# Communication-Efficient Deep Learning Algorithms for Distributed Research Networks: A Model Merging Approach with Pareto Fronts

Lu Li[1,2,3], Jenna Reps[3], Patrick Ryan[3], and Yong Chen[1,2,3,4]

[1]Applied Mathematics and Computational Sciences, University of Pennsylvania, PA, USA
[2]Center for Health AI and Synthesis of Evidence (CHASE), University of Pennsylvania, Philadelphia, PA, USA
[3]Observational Health Data Sciences and Informatics, New York, New York
[4]Department of Biostatistics, Epidemiology and Informatics (DBEI), the Perelman School of Medicine, University of Pennsylvania, PA, USA

## Background

Real-world evidence synthesis in healthcare often spans multiple sites characterized by data heterogeneity and distribution shifts. This complexity necessitates sophisticated methods to harness data across diverse sources effectively. Deep learning models have shown promise in addressing various healthcare tasks; however, the direct aggregation of individual-level patient data from different sites to train a unified model is often impractical due to stringent data privacy regulations and logistical challenges.

Traditional approaches like federated learning [1] offer a promising solution by enabling model training on decentralized data while preserving data privacy. However, conventional federated learning typically involves iterative rounds of exchanging gradients between local sites and a central server, which can be computationally expensive and impractical for healthcare institutions with limited resources and strict operational constraints.

An emerging method known as model merging [2, 3, 4] presents a novel approach to overcome these challenges. Unlike conventional federated learning, model merging does not require multiple rounds of gradient updates, while also not require the sites to share their data. Instead, this approach focuses on combining multiple models finetuned on local data into a unified model within the parameter space. By merging models in the parameter space, rather than combining data or gradients, this method preserves data privacy while still allowing communication-efficiency in collaborations across distributed sites.

In this context, our study explores a novel distributed algorithm, which is does not require the data to be share, and only uses two rounds of communications. The only things that are shared are the parameter residuals between the pretrained model parameters and finetuned model parameters, and the coefficients for the fitted quadratic surrogate functions.

Unlike traditional federated learning methods which produces a single final model, our method finds the Pareto set of solutions. The Pareto set of solutions allows the practitioners to visualize the trade-offs within the synthesis, thus allowing them to make more informed decisions. In addition, the Pareto front solutions can enhance fairness by maximizing the worst performing site. Our approach aims to optimize the synthesis of real-world evidence across healthcare networks, offering a scalable and privacy-preserving solution to leverage collective knowledge while respecting data constraints and regulatory requirements.

# Methods

## Model merging

A recent work by [2] introduced *task arithmetic* as a simple and effective way for performing model merging. The task vector for task $n$ is defined as $\mathbf{v}_n = \theta_{ft}^n - \theta_{pre}$, which is the element-wise difference between the pre-trained parameters and the fine-tuned parameters for the task $n$. To perform the model merging with task vectors, we can compute $\theta_{pre} + \sum_{n=1}^{N} c_n \mathbf{v}_n$, where $c_n$ is some scaling factors and has shown to be essential to the performance of the merged model [5, 4].

Denoting the metric of task $n$ as $M_n$, most of the existing approaches for model merging aim to improve an equal weight average metric $\frac{1}{N}\sum_{n=1}^{N} M_n$. This target implies the user of the algorithm has equal preferences between tasks. However, in real-world applications, users might have biased preferences for the importance of tasks, necessitating trade-offs. In such cases, the goal of model merging is no longer the equal weight average metric. Instead, a Pareto set of solution is preferable.

## Pareto fronts

**Pareto dominance** Let $X$ be a set representing the solution space, where each element $x \in X$ is a possible solution to the multi-objective optimization problem. Let there be $n$ objectives. Define an evaluation function $f_i : X \to \mathbb{R}$, where $i \in \{1, 2, \cdots, n\}$. Given two solutions $x, y \in X$, we define that $x$ *Pareto dominates* $y$, denoted by $x \succ_P y$, if and only if: $\forall i \in \{1, 2, \ldots, n\}, f_i(x) \leq f_i(y)$ and $\exists j \in \{1, 2, \ldots, n\}, f_j(x) < f_j(y)$.

**Pareto optimal solutions** The Pareto front is the set of solutions in the solution space $X$ that are not Pareto dominated by any other solutions in $X$. PF $= \{x \in X \mid \nexists y \in X \text{s.t.} y \succ_P x\}$

## Quadratic approximation

In many cases, approximating the Pareto front can be computationally expensive and data inefficient. We introduce our method, MAP [6], a computationally efficient method to find the Pareto front for model merging.

Given the task vectors $\{\mathbf{v}_n\}_{n\in[N]}$ and the initialization $\theta_{pre} \in \mathbb{R}^d$, we denote the merged model parameters as $\theta_{\text{merge}}(\mathbf{c}) = \theta_{pre} + \mathbf{V}\mathbf{c} = \theta_{pre} + \sum_{n=1}^{N} c_n \mathbf{v}_n$, where $\mathbf{V} = \text{concat}(\mathbf{v}_1, ..., \mathbf{v}_N) \in \mathbb{R}^{d\times N}$ is the task matrix and $\mathbf{c} = \text{concat}(c_1, ..., c_N) \in \mathbb{R}^N$ is the scaling coefficients for the task vectors.

The main idea is to use the second-order Taylor expansion to approximate $M_n$:

$$
\begin{aligned}
M_n(\mathbf{c}) \equiv M_n(\theta_{\text{merge}}(\mathbf{c})) &= M_n(\theta_{pre}) + \nabla M_n(\theta_{pre})^\top (\theta_{\text{merge}}(\mathbf{c}) - \theta_{pre}) \\
&+ \frac{1}{2}(\theta_{\text{merge}}(\mathbf{c}) - \theta_{pre})^\top \mathbf{H}_n(\theta_{pre})(\theta_{\text{merge}}(\mathbf{c}) - \theta_{pre}) + R_n(\theta_{\text{merge}}(\mathbf{c}) - \theta_{pre}) \\
&\approx M_n(\theta_{pre}) + \nabla M_n(\theta_{pre})^\top \mathbf{V}\mathbf{c} + \frac{1}{2}(\mathbf{V}\mathbf{c})^\top \mathbf{H}_t(\theta_{pre})\mathbf{V}\mathbf{c}
\end{aligned}
$$

where $\mathbf{H}_n(\theta_{pre}) = \nabla^2 M_n(\theta_{pre}) \in \mathbb{R}^{d\times d}$ is the Hessian matrix and $R_n(\theta_{\text{merge}}(\mathbf{c}) - \theta_{pre}) = R_n(\mathbf{V}\mathbf{c})$ is the third-order remainder, which is negligible when $||\mathbf{V}\mathbf{c}||^3 = ||\theta_{\text{merge}}(\mathbf{c}) - \theta_{pre}||^3$ is small. Leveraging this quadratic approximation, we can define surrogate models for each task $n$, $\tilde{M}_n(\mathbf{c}; \mathbf{A}_n, \mathbf{b}_n, e_n) \equiv e_n + \mathbf{b}_n^\top \mathbf{c} + \frac{1}{2}\mathbf{c}^\top \mathbf{A}_n \mathbf{c}$ where

$$
\mathbf{A}_n = \mathbf{V}^\top \mathbf{H}_n(\theta_{pre})\mathbf{V} \in \mathbb{R}^{T\times T}, \mathbf{b}_n = \mathbf{V}^\top \nabla M_n(\theta_{pre}) \in \mathbb{R}^T, e_n = M_n(\theta_{pre}) + R_n \tag{1}
$$

We can further leverage existing methods to learn the coefficients by minimizing the empirical risk over multiple $\mathbf{c}$: for instance,

$$\mathbf{A}_n^*, \mathbf{b}_n^*, e_n^* = \arg\min_{\mathbf{A}_n, \mathbf{b}_n, e_n} \sum_{\mathbf{c} \in \Omega} |M_n(\theta_{\text{merge}}(\mathbf{c})) - \tilde{M}_n(\mathbf{c}; \mathbf{A}_n, \mathbf{b}_n, e_n)|^2 \tag{2}$$

where $\Omega = \{\mathbf{c}^{(1)}, ..., \mathbf{c}^{(N_c)}\}$ is the set of $\mathbf{c}$ and $M_n(\theta_{\text{merge}}(\mathbf{c}))$ is the corresponding evaluation metric.

Here we formulate MAP as a federated learning algorithm which only requires two rounds of communications:

**Algorithm 1** MAP

**Input**: pretrained model with weights $\theta_{pre}$, $K$ site each with data $D_k$
**Output**: a pareto front of different scaling coefficients that can be used to combine the models.
Prepare models $\{\theta_{ft}^n\}$ and compute task vectors $\{\mathbf{v}_n = \theta_{ft}^n - \theta_{pre}\}$.
**for** $k = 1 \dots K$ **do**
    Site $k$ to finetune the pretrained model $\theta_{pre}$ on their local data $D_k$ and obtain $\theta_{ft}^k$
    Site $k$ to upload the task vector $\{\mathbf{v}_k = \theta_{ft}^k - \theta_{pre}\}$ to the server
    Site $k$ to download the other $K - 1$ task vectors from the server
**end for**
The lead site to share pre-sampled $M$ scaling coefficients $C = [c_1, \dots, c_M]^T$ to the $K$ sites
**for** $k = 1 \dots K$ **do**
    **for** $m = 1 \dots M$ **do**
        Site $k$ to compute the combined model $\theta_m = \theta_{pre} + \sum_k c_{mk} \theta_{ft}^k$
        Site $k$ to evaluate the combined model $\theta_m$ on their private data $D_k$ and obtain $\tilde{M}_k$.
        Site $k$ to fit a quadratic function on $\tilde{M}_k$ by learning $\mathbf{A}_n^*, \mathbf{b}_n^*, e_n^*$ in (2) and share back $\mathbf{A}_n^*, \mathbf{b}_n^*, e_n^*$.
    **end for**
**end for**
Fit the quadratic approximation surrogate model $\tilde{M}_n$.
Apply MOOP algorithm (e.g. NSGA-III) to $\{\tilde{M}_n\}$ and get the Pareto front

# Results

## Performance of the quadratic approximation

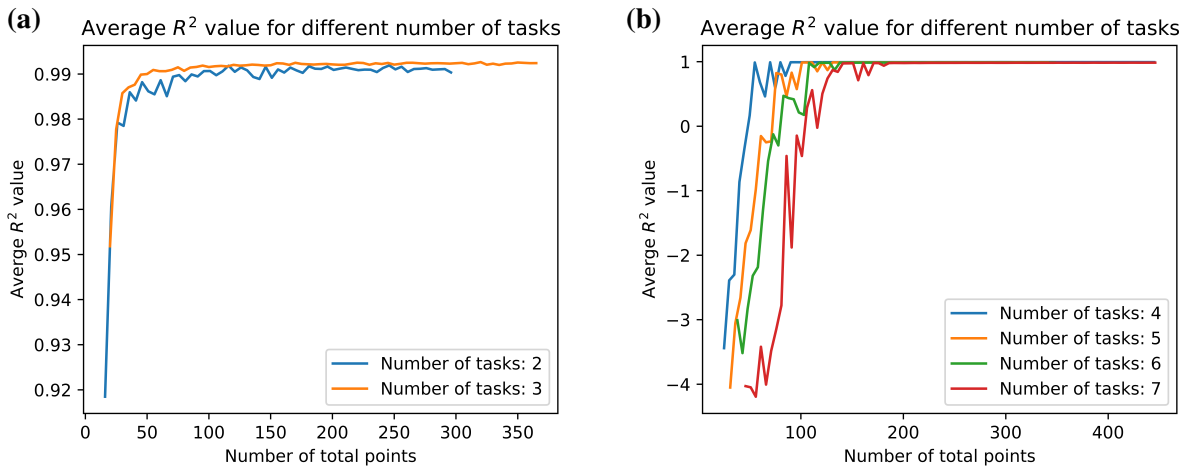


Figure 1: (a) An example of uncertainty in 2D polar coordinate system. The more uncertain within a bin, the more $(\mathbf{c}, \{M_n(\theta_{\text{merge}}(\mathbf{c}))\}_{n=1}^{N})$ pairs information we should collect in the next round. (b) An illustration of discretization of a 3D spherical coordinate system. When the dimension is higher, the discretization would be in a hyper-spherical coordinate system along with the angular dimensions.

## Zero-shot Medical Image Classification

We used the NIH dataset consisting of over 112,000 chest X-rays and 30,000 unique patients [7]. It originally contained 15 classes (14 diseases and 1 class for no finding). We split the dataset into two groups, where medical task 1 specifically tries to classify Atelectasis, Consolidation, Infiltration, Pneumothorax, and medical task 2 tries to classify Nodule, Mass and Hernia. An example image taken from the dataset is shown in Figure 2 (a).

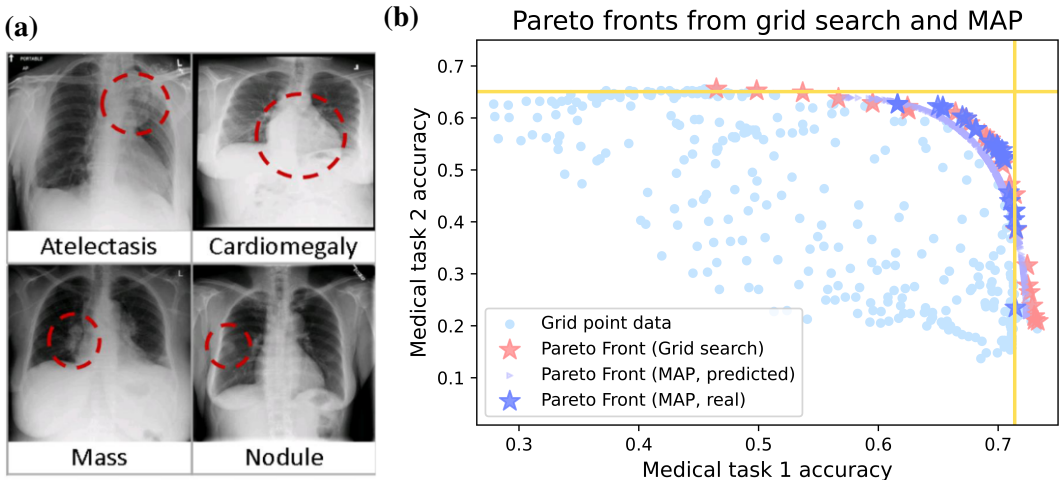


Figure 2: (a) Example figure from the NIH [7] dataset. (b) Pareto fronts found by brute-force direct search using 400 points and by MAP using 30 points. We randomly sampled 25 points from the predicted Pareto front by MAP and evaluated its performance.

Our next step is to apply MAP to electronic health records data using the 17 benchmark tasks established in OHDSI PatientLevelPrediction package.

## Conclusion

Evidence synthesis plays a central role in distributed research networks such as OHDSI. Communication-efficient and privacy-preserving distributed learning algorithms have been developed for regression-based machine learning methods [8]. This work introduces an innovative approach for deep learning-based models. Our research aligns with the mission of OHDSI by advancing the frontier of methodology in clinical evidence generation and evidence synthesis.

# References


[1] McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. PMLR; 2017. p. 1273-82.

[2] Ilharco G, Ribeiro MT, Wortsman M, Gururangan S, Schmidt L, Hajishirzi H, et al. Editing models with task arithmetic. arXiv preprint arXiv:221204089. 2022.

[3] Matena MS, Raffel CA. Merging models with fisher-weighted averaging. Advances in Neural Information Processing Systems. 2022;35:17703-16.

[4] Yang E, Wang Z, Shen L, Liu S, Guo G, Wang X, et al. Adamerging: Adaptive model merging for multi-task learning. arXiv preprint arXiv:231002575. 2023.

[5] Yadav P, Tam D, Choshen L, Raffel CA, Bansal M. Ties-merging: Resolving interference when merging models. Advances in Neural Information Processing Systems. 2024;36.

[6] Li L, Zhang T, Bu Z, Wang S, He H, Fu J, et al. MAP: Low-compute Model Merging with Amortized Pareto Fronts via Quadratic Approximation. arXiv preprint arXiv:240607529. 2024.

[7] of Health NI, et al.. NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community; 2017.

[8] Duan R, Boland MR, Moore JH, Chen Y. ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. In: BIOCOMPUTING 2019: Proceedings of the Pacific Symposium. World Scientific; 2018. p. 30-41.