Model Merging as Probabilistic Inference in Fine-Tuning Parameter Space

Long Minh Bui[1] **Tuan Anh Le Van**[2] **Tung Phi Duc**[2] **Phi Le Nguyen**[2] **Jana Doppa**[1] **Trong Nghia Hoang**[1]

[1]Washington State University
[2]Hanoi University of Science and Technology

Abstract

Model merging aims to combine existing single-task solutions into a multi-task solution without additional data-driven fine-tuning. Most existing approaches achieve this using geometric properties of local solution spaces. However, such geometric views provide limited guidance for scoring how statistically useful each task-specific update direction is across tasks during merging. To address this, we formulate model merging from a new perspective of probabilistic inference under a product-of-experts (PoE) scenario where each single-task solution defines an energy-based expert model (EBM) over the merged parameters. We show that several existing model merging methods arise as special cases of our framework under energy designs that impose implicit Gaussian assumptions on directional residuals between merged and task-specific models. Empirically, we find that these residuals are often heavy-tailed which exposes a mismatch with the imposed light-tailed Gaussian structures. We address this with a heavy-tailed PoE design based on Cauchy experts, which better captures the observed residual behavior while admitting a provably convergent inference procedure. Experiments across multiple tasks and architectures show significant improvements over state-of-the-arts baselines. Our code is available at `https://github.com/MinhLong210/PoE-EBM-Merging.git`.

1 INTRODUCTION

Large pre-trained foundation models and their task-specific fine-tuned variants [Achiam et al., 2023, Touvron et al., 2023] have become increasingly available for a wide range of downstream tasks. The growing availability of such specialized models has motivated model merging, which seeks to combine multiple task-specific models into a single multi-task model without additional data-driven fine-tuning [Hoang et al., 2019, Yurochkin et al., 2019, Hoang et al., 2020, Lam et al., 2021, Yang et al., 2024, Li et al., 2023, Hoang and Hoang, 2024]. For example, monolingual models can be combined to obtain a single multilingual model [Ahmadian et al., 2024]. Such model merging approaches are particularly valuable in many real-world production systems [Su et al., 2018] where both local datasets and training pipelines cannot be centralized and synchronized [1]. Despite requiring no additional training data, model merging can often achieve performance competitive with full multi-task fine-tuning, which might be impractical in such settings. Model merging is also attractive when storage resources are limited, such as on edge devices [Voghoei et al., 2018, Narayanswamy et al., 2024], or access to privacy-sensitive task-specific data is restricted [Liang et al., 2025, Zhang and Metaxas, 2024, Pan et al., 2024].

Prior Work. A common paradigm in model merging is to represent each task-specific model with a fine-tuning module, such as a low-rank adaptation (LoRA) matrix [Hu et al., 2022] or a task vector capturing the difference between fine-tuned and pre-trained parameters [Ilharco et al., 2022]. Model merging then reduces to aggregating these task-specific updates into a single multi-task model. Most existing approaches perform this aggregation using geometric properties of local solution spaces.

The simplest methods, including weight averaging and task arithmetic [Wortsman et al., 2022, Ilharco et al., 2022], assume that task-specific updates lie in a common solution manifold and can be directly combined. More sophisticated approaches, such as Fisher-weighted averaging [Matena and Raffel, 2022] and Gram-based weighting [Jin et al., 2022], incorporate curvature or data-dependent geometric information to better account for differences among local solution spaces during aggregation. Another line of work explicitly seeks to align local solution spaces before merging. For ex-

[1]Federated Learning [McMahan et al., 2016] can help address data privacy but still requires synchronized local training processes.

ample, DOGE [Wei et al., 2025] constructs a shared tangent space, while KnOTS [Stoica et al., 2024] derives a common low-dimensional subspace using singular value decomposition. Task-specific models are then projected onto these aligned representations prior to aggregation.

Limitation. These approaches largely treat each task-specific fine-tuning module as a deterministic point estimate and aggregate them with some geometry-guided operations. However, this deterministic view, provides little information regarding how useful each update direction is across different tasks. As a result, update directions that are effective only for individual tasks can be aggregated with directions that are consistently useful across tasks. Such task-specific directions can then cancel or dominate useful cross-task directions, pulling the merged update away from shared directions that benefit multiple tasks. This suggests that model merging should not only aggregate update directions, but also estimate how confidently each direction should influence the merged module.

Motivation and Solution Vision. Motivated by the above intuition, we investigate model merging within a broader probabilistic framework in which each task update induces a distribution over candidate shared update directions rather than a single point estimate. The local probabilistic score assigned for each candidate direction thus reflects how it is supported by the corresponding task. In this view, merging becomes evidence aggregation. Directions supported by multiple tasks receive higher aggregate confidence while directions supported only by individual tasks receive low support from the remaining tasks and are down-weighted.

We instantiate this idea by casting model merging as MAP inference in the fine-tuning parameter space under a product of task-specific energy-based experts. This view recovers existing merging rules as special cases under particular energy designs and exposes the uncertainty assumptions they implicitly impose. In particular, quadratic energies recover classical averaging-style methods and correspond to Gaussian experts. This approach is also closely related to a recent uncertainty-aware gradient matching method of Daheim et al. [2023] which imposes an implicit Gaussian expert design via its Laplace approximation interpretation. While this method provides important empirical evidence for the effectiveness of probabilistic model merging, our analysis reveals that its implicit Gaussian structure does not fit well with the distribution over directional residuals between individual and merged updates. In particular, we show that these directional residuals exhibit substantial heavy-tailed behavior while Gaussian models are inherently light-tailed (see Fig. 1). To address this limitation, we develop a product-of-experts (PoE) formulation with heavy-tailed energy-based expert models (EBM). Our solution approach is substantiated by the following technical contributions:

1. A unified probabilistic model merging framework. We formulate model merging as MAP inference in the fine-tuning parameter space under a product of task-specific energy-based experts. In this view, each fine-tuning module induces an energy over the directional residual between a candidate merged update and the task-specific update. The merged module then corresponds to the MAP estimate under the resulting product of experts. We show that existing merging rules, including uniform averaging and Fisher-weighted averaging, arise as special cases under particular energy designs. This provides a unified lens for these methods and exposes the implicit probabilistic assumptions imposed by their aggregation rules (Section 3 and 4).

2. Heavy-tailed merging with convergence guarantees. Under the above probabilistic framework, we show that existing merging methods correspond to light-tailed distributions over the directional residual $r = (\boldsymbol{\zeta} - \boldsymbol{\theta})^\top \boldsymbol{\theta}$, which measures how far a candidate merged update $\boldsymbol{\zeta}$ drifts from task update $\boldsymbol{\theta}$ along that task's own update direction. Our empirical analysis shows that these residuals exhibit heavy-tailed behavior, revealing a mismatch with the Gaussian expert structure often implicitly assumed in prior work. To address this, we develop a novel heavy-tailed Cauchy-based expert designs and an efficient fixed-point MAP inference algorithm with convergence guarantees (Section 5).

3. Evaluation across vision and language models. We evaluate the proposed framework across diverse vision and language benchmarks spanning multiple model families. Our approach consistently improves merged model performance over state-of-the-art (SOTA) baselines (Section 6).

2 PROBLEM SETUP AND BACKGROUND

Model merging aims to aggregate N existing models which are fine-tuned from a large pre-trained model $\boldsymbol{W}_0$ on different downstream datasets. Each fine-tuned task-specific model has parameters $\boldsymbol{W}_i$ which are learned from the local dataset D_i. To extract task-specific information for task $i \in [N]$, Ilharco et al. [2022] introduced the task vector $\boldsymbol{\theta}_i = \boldsymbol{W}_i - \boldsymbol{W}_0$. The task vector encodes task-specific information and allows the analysis of individual task's characteristics. The goal of model merging is to find merged parameters $\boldsymbol{W}_m$ that performs well on all tasks by designing an aggregation algorithm A to combine the task vectors:

$$\boldsymbol{W}_m \;=\; A\big(\boldsymbol{W}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\big)\,.$$

Ilharco et al. [2022] show that we can obtain multi-task models $\boldsymbol{W}_m$ by performing simple task arithmetic operations on the task vectors $\theta_1, \theta_2, \ldots, \theta_N$:

$$\boldsymbol{W}_m \;=\; \boldsymbol{W}_0 \;+\; \lambda \sum_{i=1}^{N} \boldsymbol{\theta}_i\,, \tag{1}$$

where λ is a scaling coefficient, usually tuned on a validation set. When $\lambda = 1/N$, Eq. (1) reduces to performing

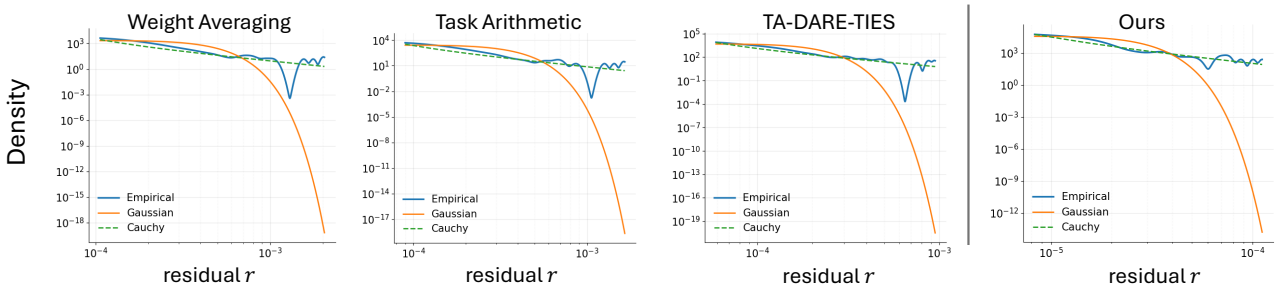

Figure 1: Empirical distributions of the directional residual $r = (\boldsymbol{\zeta} - \boldsymbol{\theta})^\top \boldsymbol{\theta}$, which measures the drift of a merged update $\boldsymbol{\zeta}$ from a task update $\boldsymbol{\theta}$ along that task's own update direction. We compare residuals produced by different merging methods when merging 7 fine-tuned ViT-L/14 models. Log-scale density plots show that empirical tails decay substantially more slowly than the fitted Gaussian and align more closely with a fitted Cauchy distribution, indicating pronounced heavy-tailed residual behavior in model merging, revealing a mismatch with the (implicit) Gaussian expert structure in prior work.

weight averaging over $\boldsymbol{W}_i$ [Wortsman et al., 2022]. Matena and Raffel [2022] further generalize this idea by scaling local parameters with their corresponding Fisher Information Matrix $\boldsymbol{F}_i$ which results in the following Fisher Averaging:

$$\boldsymbol{W}_m \quad = \quad \sum_{i=1}^{N} \alpha_i \left(\sum_{t=1}^{N} \alpha_t \boldsymbol{F}_t \right)^{-1} \boldsymbol{F}_i \boldsymbol{\theta}_i \ , \qquad (2)$$

where

$$\mathbf{F}_i \quad = \quad \mathbb{E}_{x \sim D_i} \left[\mathbb{E}_{y \sim p_{\mathbf{W}_i}(y|x)} \right] \mathbf{G}_i \mathbf{G}_i^\top \qquad (3)$$

with $\mathbf{G}_i = \nabla_{\mathbf{W}_i} \log p_{\mathbf{W}_i}(y|x)$.

Alternatively, RegMean [Jin et al., 2022] matches model behavior by aligning activations between the merged and task-specific models at each linear layer. This leads to a linear regression problem defined by the task-specific data matrix $\boldsymbol{X}_i$ which introduces another merging rule:

$$\boldsymbol{W}_m = \left(\sum_{i=1}^{N} \frac{1}{N_i} \boldsymbol{X}_i^\top \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \frac{1}{N_i} \boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{\theta}_i \right). \qquad (4)$$

Overall, existing merging methods differ in how they align local solution spaces. Yet, they essentially view task updates as deterministic quantities to be aggregated without weighing their statistical usefulness across different tasks (Section 1). We therefore seek a probabilistic treatment of fine-tuning modules that not only account for the local update geometries, but also estimate how confidently each direction should influence the merged module.

3 A PRODUCT-OF-EXPERTS (POE) PERSPECTIVE ON MODEL MERGING

In this section, we develop a new perspective for model merging in which task-specific fine-tuning modules can be interpreted as observations of a shared latent parameter that captures information common across tasks. The merged model can be viewed as a sample drawn from a product-of-experts (PoE) model combining these local energy-based densities (Section 3.1). In this view, model merging can then be formulated as MAP inference which also provides a unified probabilistic interpretation of existing merging rules. In particular, we show that weight averaging, Fisher-weighted averaging, and RegMean arise as special cases under specific choices of task-specific energy functions, thereby making their implicit distributional assumptions explicit (Section 3.2). Section 4 then shows that these implicit assumptions are mismatched with the empirical deviations observed between candidate merged updates and task-specific modules, and introduces a redesign of the local energy functions to mitigate this mismatch.

3.1 POE FORMULATION FOR MODEL MERGING

Energy-based models (EBMs) [Teh et al., 2003, LeCun et al., 2006, Song and Kingma, 2021] define probability distributions through an unnormalized function called an *energy* $E(\boldsymbol{\zeta})$ assigning lower energy to more likely input $\boldsymbol{\zeta}$. The resulting distribution is given by

$$p(\boldsymbol{\zeta}) \quad = \quad Z^{-1} \cdot \exp\big(- E(\boldsymbol{\zeta}) \big) \ , \qquad (5)$$

where $Z \triangleq \int \exp\left(-E(\boldsymbol{\zeta})\right) d\boldsymbol{\zeta}$ is the partition function, ensuring the distribution integrates to 1. Under this view, inputs with lower energies correspond to higher probability. This formulation only needs to specify compatibility via the energy without explicit normalization.

Now, considering the problem of merging N task-specific models. We assume that there exists an unknown latent parameter $\boldsymbol{\zeta}$ that is shared across tasks. Each task-specific fine-tuning module can then be viewed as providing noisy evidence about the latent shared parameter $\boldsymbol{\zeta}$. Therefore, we associate each task with an expert distribution over $\boldsymbol{\zeta}$, where higher probability corresponds to greater compatibility between the latent parameter and the task-specific update. We model this compatibility through an energy function, result-

ing in the following energy-based expert:

$$p_i(\boldsymbol{\zeta}) \quad \propto \quad \exp\Big(-E_i(\boldsymbol{\zeta})\Big)\,, \tag{6}$$

where $E_i(\boldsymbol{\zeta})$ is an energy function determined by the task-specific parameter $\boldsymbol{\theta}_i$. The energy function is general and can be designed with specific desiderata. Local energy functions can also be combined naturally through multiplication as established in [Hinton, 2002]. Essentially, each task defines its energy $E_i(\boldsymbol{\zeta})$ and under conditional independence assumption of tasks on $\boldsymbol{\zeta}$, the PoE distribution is given as

$$p(\boldsymbol{\zeta}) \quad \propto \quad \prod_{i=1}^{N} p_i(\boldsymbol{\zeta}) \;\propto\; \exp\left(-\sum_{i=1}^{N} E_i(\boldsymbol{\zeta})\right). \tag{7}$$

This results in the global energy $E(\boldsymbol{\zeta}) \triangleq \sum_{i=1}^{N} E_i(\boldsymbol{\zeta})$, such that $p(\boldsymbol{\zeta}) \propto \exp\left(-E(\boldsymbol{\zeta})\right)$. Combining task experts then corresponds to summing up their energies. The resulting PoE distribution thus concentrates probability mass on configurations $\boldsymbol{\zeta}$ that simultaneously achieve low energy for all experts. Under this view, model merging reduces to inference with an energy-based model over the shared parameter.

A natural way to obtain a merged model is to compute the maximum-a-posteriori (MAP) estimate:

$$\begin{aligned} \boldsymbol{\zeta}^* \quad &= \quad \arg\max_{\boldsymbol{\zeta}} p(\boldsymbol{\zeta}) \;=\; \arg\min_{\boldsymbol{\zeta}} E(\boldsymbol{\zeta}) \\ &= \quad \arg\min_{\boldsymbol{\zeta}} \sum_{i=1}^{N} E_i(\boldsymbol{\zeta})\,. \end{aligned} \tag{8}$$

Interestingly, this formulation unifies existing averaging heuristics as cases of quadratic energies as shown in Section 3.2. At the same time, it also allows for new designs of more flexible and robust choices of energy function. Merging operation can thus be tightly coupled with the design of local energy functions as shown in Eq. (8). The MAP solution of the resulting PoE can then be found via minimizing Eq. (8), resulting in a merged model corresponding to the most probable shared parameter.

3.2 GAUSSIAN EXPERTS RECOVER EXISTING MERGING RULES

The key modeling aspect in the PoE-EBM framework is the design choice of the task-specific function $E_i(\boldsymbol{\zeta})$. Different choices encode different assumptions about how task-specific parameters deviate from the shared latent structure. These choices also lead to various merging designs. Several of which were rediscovered below as special cases of PoE with Gaussian experts.

A. Gaussian Expert Formulation. Assume each task-specific parameter $\boldsymbol{\theta}_i$ is generated from the shared latent parameter $\boldsymbol{\zeta}$ under Gaussian noise:

$$\boldsymbol{\theta}_i \quad = \quad \boldsymbol{\zeta} + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)\,. \tag{9}$$

This induces the likelihood over the latent $\boldsymbol{\zeta}$,

$$p_i(\boldsymbol{\zeta}) \quad \propto \quad \exp\Big(-\frac{1}{2}(\boldsymbol{\zeta}-\boldsymbol{\theta}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\zeta}-\boldsymbol{\theta}_i)\Big)\,. \tag{10}$$

Thus, the corresponding energy function is quadratic

$$E_i(\boldsymbol{\zeta}) \quad = \quad \frac{1}{2}(\boldsymbol{\zeta}-\boldsymbol{\theta}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\zeta}-\boldsymbol{\theta}_i)\,. \tag{11}$$

The MAP solution in Eq. (8) thus exhibits a closed form:

$$\boldsymbol{\zeta}^* \quad = \quad \left(\sum_{i=1}^{N} \boldsymbol{\Sigma}_i^{-1}\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\theta}_i\right). \tag{12}$$

B. Recovering Existing Merging Methods. We now show that various classical merging methods are special cases of the above PoE with Gaussian experts which exhibit quadratic local energy functions. In particular, we will show that the above MAP estimator admits several well-known model merging rules as special cases under different choices and structural assumption of the precision matrices $\boldsymbol{\Sigma}_i^{-1}$.

1. Uniform Averaging. Suppose all precision matrices are isotropic, $\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{I}$, Eq. (12) simplifies to

$$\boldsymbol{\zeta}^{\text{avg}} \quad = \quad \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\theta}_i\,, \tag{13}$$

which recovers uniform averaging [Wortsman et al., 2022].

2. Fisher-Weighted Averaging. Suppose the precision matrix of each task is chosen as its Fisher information matrix, $\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{F}_i$, Eq. (12) becomes

$$\boldsymbol{\zeta}^{\text{FA}} \quad = \quad \left(\sum_{i=1}^{N} \boldsymbol{F}_i\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{F}_i\boldsymbol{\theta}_i\right), \tag{14}$$

which recovers Fisher averaging [Matena and Raffel, 2022].

3. RegMean. When the precision matrix is taken as the Gram matrix of empirical data,

$$\boldsymbol{\Sigma}_i^{-1} \quad = \quad \frac{1}{N_i}\boldsymbol{X}_i^\top \boldsymbol{X}_i\,, \tag{15}$$

the MAP estimator reduces to

$$\boldsymbol{\zeta}^{\text{RM}} = \left(\sum_{i=1}^{N} \frac{1}{N_i}\boldsymbol{X}_i^\top \boldsymbol{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \frac{1}{N_i}\boldsymbol{X}_i^\top \boldsymbol{X}_i\boldsymbol{\theta}_i\right), \tag{16}$$

which recovers RegMean [Jin et al., 2022].

4 FROM GAUSSIAN TO HEAVY-TAILED EXPERT MODELS

Section 3 shows that several existing merging methods can be recovered as PoE models with Gaussian experts. We now show that these Gaussian formulations impose light-tailed assumptions on directional residuals, which mismatch the heavy-tailed behavior observed in practice (Section 4.1). To address this mismatch, we introduce a heavy-tailed PoE formulation based on Cauchy experts (Section 4.2).

4.1 LIMITATIONS OF GAUSSIAN EXPERTS

As shown above, the use of quadratic energy fields lead to EBMs with Gaussian shapes. Such Gaussian experts are closely related to the framework of [Daheim et al., 2023] which assumes a Gaussian prior over task-specific parameters. This is followed by a Laplace approximation to obtain the Gaussian posterior for the merged parameters. Under specific choices of the precision matrix, their formulation reduces to precision-weighted averaging and recovers standard schemes such as weight averaging and Fisher-weighted averaging. However, such Gaussian structures often do not sufficiently capture the tail behaviors of the merged models due to their fast-decaying tails. This represents a structural mismatch according to our empirical findings in Fig. 1 which shows heavy tail behavior of the merged models.

This can be seen by analyzing the **directional residual** of the merged model with respect to each task-specific module. Given a candidate merged parameter $\boldsymbol{\zeta}$ and the i-th fine-tuning module $\boldsymbol{\theta}_i$, the directional residual is defined as

$$r_i(\boldsymbol{\zeta}) \quad \triangleq \quad (\boldsymbol{\zeta} - \boldsymbol{\theta}_i)^\top \boldsymbol{\theta}_i \,. \tag{17}$$

Intuitively, this computes the difference $(\boldsymbol{\zeta} - \boldsymbol{\theta}_i)$ between the merged and individual models which is then projected onto the direction of the task update $\boldsymbol{\theta}_i$. The result measures how far the merged solution drifted from the task-specific module along that task's preferred direction. Under a Gaussian model $\boldsymbol{\zeta} - \boldsymbol{\theta}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)$, each directional residual thus follows a Gaussian $r_i(\boldsymbol{\zeta}) \sim \mathcal{N}(0, \boldsymbol{\theta}_i^\top \boldsymbol{\Sigma}_i \boldsymbol{\theta}_i)$.

As a result, existing choices of Gaussian experts, and by extension the classical merging methods they recover, implicitly impose light-tailed assumptions on the residuals. However, a closer inspection of the directional residuals of different merging methods reveals an intrinsic heavy-tailed behavior that stands in contrast to these Gaussian assumptions. In Fig. 1, we plot the empirical density of the directional residuals across all tasks and layers when merging 7 (fully) fine-tuned ViT-L-14 models and overlay Gaussian and heavy-tailed Cauchy distributions for comparison. It shows a heavy-tailed behavior: the decay in the tail region indicates a heavy-tailed distribution, deviating from Gaussian behavior and following more closely with the Cauchy distribution. In particular, large residuals occur more frequently than would be predicted under a Gaussian assumption.

4.2 HEAVY-TAILED EXPERT MODELS

To capture this inherent heavy-tailed residual distribution, we adopt a robust energy function whose logarithmic growth reduces the influence of task-specific updates with large directional residuals. The negative log-density of a Cauchy distribution [Liu and Tao, 2014] naturally exhibits this behavior, motivating the following Cauchy expert design:

$$E_i^{Cauchy}(\boldsymbol{\zeta}) \quad \triangleq \quad \log\left(1 + \frac{r_i(\boldsymbol{\zeta})^2}{\gamma^2}\right) \,, \tag{18}$$

where $r_i(\boldsymbol{\zeta})$ denotes the directional residual at task-specific parameter $\boldsymbol{\theta}_i$ and $\gamma > 0$ is a user-defined scale controlling the tail heaviness of the distribution. This energy function induces the following (unnormalized) Cauchy density:

$$p_i(\boldsymbol{\zeta}) \quad \propto \quad \exp\Big(-E_i^{Cauchy}(\boldsymbol{\zeta})\Big) \;= \frac{\gamma^2}{\gamma^2 + r_i(\boldsymbol{\zeta})^2} \,, \tag{19}$$

which is heavy-tailed. Under conditional independence assumption of tasks, the PoE-EBM posterior factorizes according to Eq. (7), leading to the following global energy:

$$E^{Cauchy}(\boldsymbol{\zeta}) \quad = \quad \sum_i \log\Big(\gamma^2 + r_i(\boldsymbol{\zeta})^2\Big) + \mathcal{C} \,, \tag{20}$$

where $\mathcal{C}$ is a constant independent of $\boldsymbol{\zeta}$. This defines the Cauchy score as negative of the global energy gradient:

$$\begin{aligned} S^{Cauchy}(\boldsymbol{\zeta}) \quad &\triangleq \quad -\nabla_{\boldsymbol{\zeta}} E^{Cauchy} \\ &= \quad -\sum_{i=1}^{N} \frac{2 r_i(\boldsymbol{\zeta})}{\gamma^2 + r_i(\boldsymbol{\zeta})^2} \nabla_{\boldsymbol{\zeta}} r_i(\boldsymbol{\zeta}) \,, \end{aligned} \tag{21}$$

where the residual gradient is $\nabla_{\boldsymbol{\zeta}} r_i(\boldsymbol{\zeta}) = \boldsymbol{\theta}_i$. We further provide insights into the Cauchy score and the connection between Cauchy and Gaussian experts in Appendix D.

5 MAP INFERENCE ALGORITHM AND CONVERGENCE GUARANTEE

We will now develop a practical MAP inference algorithm for the previously established heavy-tailed PoE model. We first derive a fixed-point characterization of the optimal merged update and then use it to obtain an iterative procedure with a convergence guarantee.

As shown in the general PoE-EBM setting in Section 3.1, merging corresponds to computing the MAP estimator in Eq. (8). For Cauchy experts with global Cauchy energy in Eq. (20), the inference task then becomes

$$\boldsymbol{\zeta}^* \quad = \quad \arg\min \sum_{i=1}^{N} \log\Big(\gamma^2 + r_i(\boldsymbol{\zeta})^2\Big) \,. \tag{22}$$

Unlike the quadratic case in Eq. (12), this optimization loss is nonconvex and does not admit a closed-form minimizer. However, we can exploit its structure to derive an explicit optimality characterization. In particular, the MAP estimator can be expressed as the closed-form solution of a nonlinear equation, revealing that robust model merging amounts to a residual-dependent weighted consensus among task-specific models. The following theorem formalizes this closed-form characterization of the MAP solution.

Theorem 5.1 (Closed-form characterization of MAP)**.** *Let $\{\boldsymbol{\theta}_i\}_{i=1}^N$ be a set of task-specific fine-tuning modules, and define the following auxiliary function:*

$$u_i(\boldsymbol{\zeta}) \;=\; \Big(\big[(\boldsymbol{\zeta}-\boldsymbol{\theta}_i)^\top \boldsymbol{\theta}_i\big]^2 + \gamma^2\Big)^{-1} \quad \textit{with} \;\; \gamma > 0\,. \tag{23}$$

It then follows that any stationary point $\boldsymbol{\zeta}^*$ of the MAP loss in Eq. (22) satisfies $\boldsymbol{\zeta}^* = F(\boldsymbol{\zeta}^*)$ with a closed-form mapping F defined below;

$$\begin{aligned} F(\cdot) \;&\triangleq\; \left(\sum_{i=1}^N u_i(\cdot)\,\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top + \eta \boldsymbol{I}\right)^{-1} \left(\sum_{i=1}^N u_i(\cdot)\,\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top \boldsymbol{\theta}_i\right) \\ &\triangleq\; \boldsymbol{H}(\cdot)^{-1}\boldsymbol{b}(\cdot)\,, \end{aligned} \tag{24}$$

where we define $\boldsymbol{H}(\cdot) \triangleq \sum_{i=1}^N u_i(\cdot)\,\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top + \eta\boldsymbol{I}$ and $\boldsymbol{b}(\cdot) \triangleq \sum_{i=1}^N u_i(\cdot)\,\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top\boldsymbol{\theta}_i$ for ease of notation. Here, $\eta > 0$ is a conditioning hyper-parameter to ensure the inversion operator in the definition of F is well-defined.

Under mild boundedness assumptions on the fine-tuning modules (see Assumption 5.2), Theorem 5.4 further shows that F is a contractive map with a unique fixed point solution. Consequently, repeated application of F converges to this fixed point, which corresponds to the MAP estimate (i.e., the optimal merged solution) of `PoE-EBM`. This forms the basis of our merging algorithm (see Algorithm 1).

Assumption 5.2 (Bounded fine-tuning modules)**.** There exists a constant $M > 0$ such that $\|\boldsymbol{\theta}_i\|_2 \le M,\ \forall i \in [N]$.

We note that in fine-tuning regimes, task-specific parameters are fine-tuned from a common pre-trained model with common practices such as ℓ_2 regularization with small learning rates or constraining the rank of the fine-tuning modules. Consequently, these modules often remain within a bounded region of the parameter space.

To validate this assumption, we compute the ℓ_2 norm of full fine-tuned task vectors across vision tasks and their ratios relative to the pretrained model. As shown in Table 6, the ratio is consistently in the range of $0.6\% - 0.8\%$ across all tasks and models. Motivated by this empirical observation, we restrict our analysis to a local neighborhood around the fixed point $\boldsymbol{\zeta}^*$ and assume that all iterates remain within a unit ball centered at $\boldsymbol{\zeta}^*$ as formally stated below.

Assumption 5.3 (Local Contraction Neighborhood)**.** Let $\boldsymbol{\zeta}^*$ be a fixed point of F. The iterates $\{\boldsymbol{\zeta}^{(k)}\}_k$ generated by F are assumed to remain in a closed unit ball around $\boldsymbol{\zeta}^*$:

$$\forall k \;:\; \boldsymbol{\zeta}^{(k)} \in B_1(\boldsymbol{\zeta}^*) \;\triangleq\; \big\{\boldsymbol{\zeta} \;:\; \|\boldsymbol{\zeta}-\boldsymbol{\zeta}^*\| \;\le\; 1\big\}\,. \tag{25}$$

Given the above assumption, we can now show that F is contractive via Theorem 5.4 below.

Theorem 5.4 (Contraction of F)**.** *Under Assumptions 5.2 and 5.3, it follows that F is a contractive mapping where $\|F(\boldsymbol{\zeta}^*) - \boldsymbol{\zeta}^{(k+1)}\| = \|F(\boldsymbol{\zeta}^*) - F(\boldsymbol{\zeta})^{(k)}\| \le L\|\boldsymbol{\zeta}^* - \boldsymbol{\zeta}^{(k)}\|$ with L being a provably small Lipschitz constant,*

$$L \;\triangleq\; 2M\sum_{i=1}^N \big\|\boldsymbol{H}^{-1}\boldsymbol{J}_i(\boldsymbol{\zeta}^*)\big\| \;\le\; 1\,, \tag{26}$$

where $\boldsymbol{H}$ is previously defined in Theorem 5.1 and $\boldsymbol{J}_i(\boldsymbol{\zeta}^*) \triangleq u_i(\boldsymbol{\zeta}^*)\Big(\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top\boldsymbol{\theta}_i - \boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top F(\boldsymbol{\zeta}^*)\Big)$.

Due to limited space, the proof of Theorem 5.4 is deferred to Appendix B. We also show empirically in Fig. 4 that $L < 1$ in all experiments, asserting that F is indeed a contracting map in a neighborhood of the true MAP $\boldsymbol{\zeta}^*$. This leads to a practical merging procedure via iterating F in Algorithm 1. Its complexity analysis is provided in Appendix C.

Algorithm 1 `PoE-EBM` merging

INPUT: Backbone model $\boldsymbol{W}_0$, fine-tuning modules $\{\boldsymbol{\theta}_i\}_{i=1}^N$, conditioning parameter $\eta > 0$, scaling coefficient λ, and number of iterations T.
OUTPUT: Merged module $\boldsymbol{\zeta}$, merged model $\boldsymbol{W}_m$.
Initialize $\boldsymbol{\zeta}^{(0)}$
for $k = 0$ to $T-1$ **do**
  for $i = 1$ to N **do**
    Compute residual $r_i^{(k)} \leftarrow (\boldsymbol{\zeta}^{(k)} - \boldsymbol{\theta}_i)^\top\boldsymbol{\theta}_i$
    Compute $u_i^{(k)} \leftarrow 1/((r_i^{(k)})^2 + \gamma\|\boldsymbol{\theta}_i\|^2)$
  end for
  Compute $\boldsymbol{H}^{(k)} = \boldsymbol{H}(\boldsymbol{\zeta}^{(k)})$, $\boldsymbol{b}^{(k)} = \boldsymbol{b}(\boldsymbol{\zeta}^{(k)})$ via 24
  Compute new iterate $\boldsymbol{\zeta}^{(k+1)} = \big(\boldsymbol{H}^{(k)}\big)^{-1}\boldsymbol{b}^{(k)}$
end for
return $\boldsymbol{\zeta} \triangleq (\boldsymbol{\zeta}^{(T)})$, $\boldsymbol{W}_m = \boldsymbol{W}_0 + \lambda\boldsymbol{\zeta}^{(T)}$

6 EXPERIMENTS AND RESULTS

In this section, we validate the effectiveness our framework on diverse empirical settings covering both vision and language tasks. We describe the setup of our experiments in Section 6.1 and provide the main results in Section 6.2. We also provide additional ablation analysis in Section 6.4.

For clarity, we use the following highlighting convention: (1) best accuracy is **bolded**; (2) second best accuracy is underlined; and (3) task-specific fine-tuning accuracy (as performance upper bound) is colored blue.

6.1 EXPERIMENTAL SETUP

Tasks. We evaluate on both vision and language benchmarks. For vision, we consider a 7-task benchmark and an extended 13-task benchmark. For language task, we use 8 datasets from GLUE benchmark [Wang et al., 2018].

Models. For vision tasks we merge CLIP ViT-B/32 and ViT-B/14 models [Radford et al., 2021] under both fully

Table 1: Multi-task performance comparison when merging ViT-B/32 (fully finetuned) across 7 vision benchmarks (absolute accuracy). Task-specific finetuning accuracy is blue - performance accuracy upper bound.

Method	ViT-B/32							
	MNIST	**SVHN**	**Cars**	**DTD**	**GTSRB**	**EuroSAT**	**RESISC45**	**Average**
Finetuning	99.67	97.46	76.36	97.29	99.11	99.78	95.44	95.02
Weight averaging	85.00	64.05	60.10	51.76	55.38	63.44	68.30	64.00
Task Arithmetic (TA)	90.65	71.66	60.17	55.00	62.69	66.15	69.22	67.93
TA-DARE-TIES	95.98	82.03	59.41	60.90	70.84	70.48	67.52	72.45
Fisher Merging	89.77	82.35	69.15	54.52	57.98	77.11	74.03	72.13
DOGE-TA	98.41	87.52	**70.53**	64.31	87.76	89.93	**82.37**	82.97
Concrete-TA	96.99	80.01	57.99	58.56	71.27	76.22	71.59	73.23
PoE-EBM (Ours)	**98.88**	**91.65**	65.93	**76.06**	**88.80**	**90.04**	77.92	**84.18**

fine-tuned and LoRA fine-tuned settings. For language task, we merge LoRA fine-tuned Flan-T5-base and Flan-T5-large.

Model Merging Methods. We compare our PoE-EBM against common model merging baselines including Weight Averaging [Wortsman et al., 2022], Task Arithmetic [Ilharco et al., 2022], DARE-TIES [Yadav et al., 2023, Yu et al., 2024]. We also compare with recent subspace-based and optimization-based merging methods such as Concrete [Tang et al., 2023], KnOTS [Stoica et al., 2024] (for LoRA-fine-tuned vision models) and DOGE [Wei et al., 2025].

Metrics. We report absolute accuracy and normalized accuracy (relative to each task's fine-tuned model). Full experimental details are provided in Appendix F.

6.2 MODEL MERGING FOR VISION TASKS

We present the results for merging fully-finetuned ViT-B-32 and ViT-L-14 models on 7 vision benchmarks in Table 1 and Table 8, respectively. We also present the results for LoRA-finetuned ViT-L-14 models in Table 12.

PoE-EBM consistently outperforms both simple weight averaging and task arithmetic baselines on every individual task. More recent alignment-based methods, such as Concrete and DOGE, yield further improvements over task arithmetic; however, their average accuracies remain lower to our approach. In general, PoE-EBM consistently achieves best or second best accuracies on individual tasks and best accuracy on average.

We note that while DOGE-TA outperforms PoE-EBM on a subset of tasks, these gains come at the expense of substantial degradation on others. For instance, when merging ViT-B/32 and ViT-L/14 models, DOGE-TA achieves absolute accuracies of 64.31 and 72.23 on DTD, respectively, which are markedly lower than those obtained by PoE-EBM (76.06 and 83.56). This behavior highlights that our approach yields a more favorable trade-off across tasks, maintaining strong general performance without disproportionately sacrificing any individual task. PoE-EBM also

Table 2: Average performance comparison when merging fully finetuned ViT-L/14 models and ViT-L/32 models across 13 vision benchmarks. Task-specific fine-tuning accuracy is blue - performance accuracy upper bound.

Method	**ViT-L-14**	**ViT-B-32**
Finetuning	95.93	93.20
Weight Averaging	71.08	65.22
Task Arithmetic (TA)	80.57	66.38
TA-DARE-TIES	81.18	67.09
Fisher Merging	79.61	69.87
PoE-EBM (Ours)	**87.10**	**77.96**

demonstrates strong robustness as the number of tasks increases. As shown in Table 2, our framework achieves substantial performance gains over all baselines when merging a larger set of tasks, for both ViT-B/32 and ViT-L/14 models.

Table 12 shows that under LoRA-fine-tuned settings, our merging framework substantially outperforms task arithmetic and weight averaging variants and achieves better performance than LoRA-specific merging methods such as KnOTS-TIES and KnOTS-DARE-TIES [Stoica et al., 2024]. Taken together, these results demonstrate that PoE-EBM consistently generalizes across model scales, numbers of merged tasks, and fine-tuning regimes, without requiring architecture-specific modifications.

6.3 MODEL MERGING FOR LANGUAGE TASKS

We now present results on eight datasets from the GLUE benchmark using Flan-T5-base and Flan-T5-large models in Tables 3 and 9. As noted in [Tang et al., 2023], pre-trained text-generation models already exhibit strong inherent multitask capabilities, which can limit the extent of gains achievable through task-specific fine-tuning. Despite this, PoE-EBM achieves the best average generation performance across tasks on both model architectures, indicating that our framework remains effective even in regimes where

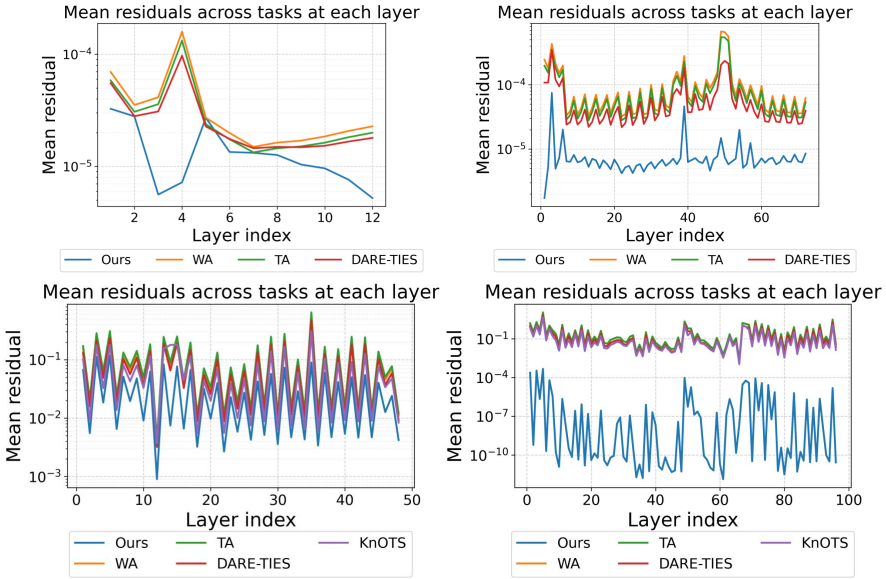

Figure 2: Plots of average directional residual over tasks (log-scale) at every layer weight in ViT-B-32 (left column) and ViT-L-14 (right column) incurred by different merging methods. Top row - merging fully fine-tuned models. Bottom row - merging LoRA fine-tuned models. It can be observed that PoE−EBM consistently achieve lower directional residual values at every layer than other merging methods.

Table 3: Multi-task performance comparison when merging Flan-T5-base models (LoRA-fine-tuned) across 8 GLUE benchmarks. Task-specific fine-tuning accuracy is colored blue which indicates the upper bound on performance accuracy for model merging. It can be observed that PoE−EBM achieves the best rank among the considered methods across tasks.

Method	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	Average	Rank
Finetuning	69.13	82.70	85.50	90.90	84.00	84.40	92.90	87.40	84.62	NA
Weight Averaging	**69.70**	59.65	78.92	90.07	83.79	80.51	91.12	71.89	78.21	2.75
Task Arithmetic	69.32	59.00	78.68	90.13	83.84	79.06	91.51	72.87	78.05	3.00
TIES-Merging	69.13	59.09	78.68	90.08	**83.91**	80.14	91.51	71.85	78.05	3.50
Concrete-TA	69.22	58.21	78.19	89.97	83.60	79.42	91.63	73.24	77.93	3.50
DOGE-TA	69.12	71.92	80.93	**90.32**	83.51	79.82	92.53	71.13	79.91	3.13
PoE−EBM (Ours)	69.22	**75.33**	**82.84**	88.87	82.91	**80.52**	**92.55**	**82.77**	**81.87**	**2.38**

performance improvements are intrinsically constrained.

Scaling to larger LLMs. We further evaluate PoE−EBM by merging three 7B-parameter models: Vicuna-7B, Llama-2-Coder, and WizardMath, all fine-tuned from the same Llama-2-7B backbone. The merged model is evaluated on GSM8K math problems. Despite operating on multi-billion-parameter models, PoE−EBM requires only **122 seconds on 2×A100 GPUs**, demonstrating its practical computational cost. In addition, PoE−EBM achieves the highest accuracy among all compared methods, outperforming both individual expert models and strong merging baselines. As shown in Table 4, the merged model surpasses the strongest expert (WizardMath) by more than **4 percentage points**, indicating that PoE−EBM can effectively combine complementary capabilities from specialized models.

6.4 ANALYSIS AND ABLATIONS

We provide ablation studies on the averaged directional residuals incurred by PoE−EBM and other merging baselines, the empirical convergence of the fixed point map and runtime analysis of PoE−EBM. Additional ablation on the performance sensitivity with respect to scaling parameter γ is provided in Appendix G.

Residual values across layers. To compare directional alignment of different merging methods, we plot the task-average squared residual value

$$Mean\Big(||(\boldsymbol{\zeta}^* - \boldsymbol{\theta}_i)^\top \boldsymbol{\theta}_i||^2\Big)\,,$$

where $\boldsymbol{\zeta}^*$ is the merged model at each layer of both fully and LoRA fine-tuned ViT-B/32 and ViT-L/14 models, and $Mean()$ represents the average taken across tasks. The re-

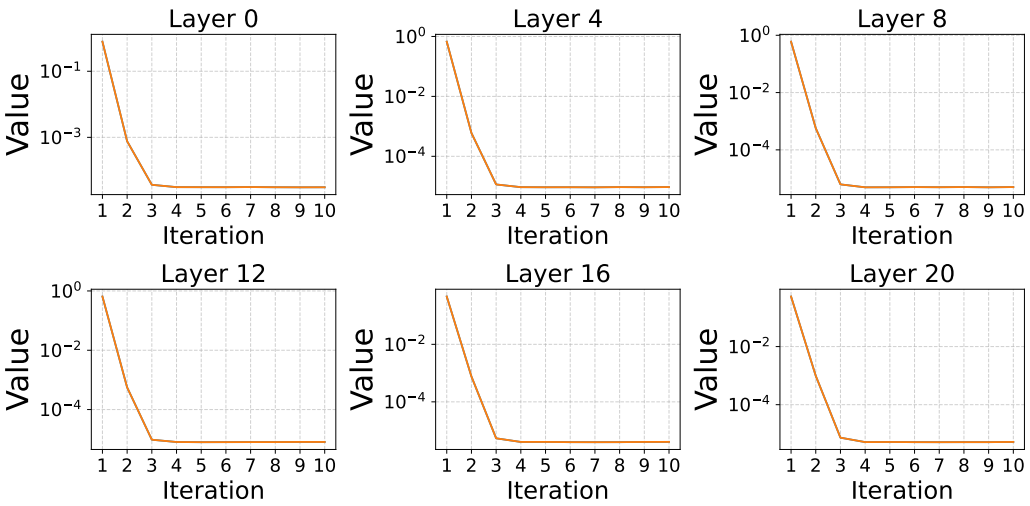

Figure 3: Convergence of the fixed point iteration method (see Algorithm 1) across layers of the merged ViT-L-14 model on 7 vision benchmarks using our PoE−EBM framework with heavy-tailed experts. We visualize $||\zeta^{(k+1)} - \zeta^{(k)}||_2$ across 10 iterations. Our results show rapid convergence to the MAP solution.

Table 4: Accuracy (%) on GSM8K dataset after merging three 7B-parameter models (Vicuna-7B, WizardMath, and Llama-2-Coder) fine-tuned from a shared Llama-2-7B backbone. PoE−EBM achieves the highest accuracy, outperforming both individual expert models and standard merging baselines (Task Arithmetic, Weight Averaging).

Method	**Accuracy (%)**
Vicuna-7B	14
WizardMath	58
Llama-2-Coder	10
Task Arithmetic	46
Weight Averaging	40
PoE-EBM	**62**

sults are shown in Figure 2, respectively. Across all settings, PoE−EBM consistently achieves lower residual magnitude, indicating better directional alignment than other approaches. Notably, other methods exhibit pronounced outlier residuals at certain layers. For example, it can be observed that the residual magnitude is abnormally large at layer 5 of the LoRA-fine-tuned ViT-L/14 in Figure 2. In contrast, PoE−EBM maintains low residual values across all layers, reflecting more stable and robust alignment behavior.

Empirical convergence of the fixed point map. We empirically validate the convergence of the fixed point iteration algorithm (see Algorithm 1) to compute the MAP point of our PoE−EBM. We report the merging results with ViT-L-14 models on 7 vision tasks. As shown in Figure 3 in Appendix G, the sequence $\{\zeta^{(k)}\}$ converges rapidly across all examined layers, with the update magnitude $||\zeta^{(k+1)} - \zeta^{(k)}||_2$ decreasing sharply within the first few iterations and stabilizing thereafter. This convergence pattern indicates that the contracting map admits a solution and the resulting estimate corresponds to the MAP solution under our probabilistic formulation. Importantly, the convergence behavior is uniform across layers, implying that the optimization landscape induced by our model is well-conditioned in practice and that the algorithm is reliable for large-scale model merging. The convergence behavior is observed across layers, implying that the optimization landscape induced by PoE−EBM is well-conditioned in practice.

Runtime analysis. We also evaluate the efficiency of PoE−EBM by measuring the wall-clock runtime of the merging procedure across different tasks and model scales. As reported in Table 7, the proposed merging procedure completes in under one minute for all models considered. For the largest model, Flan-T5-large with 0.7B parameters, the full-layer merging process takes only 59.19s. The detailed complexity analysis is deferred to Appendix C.

7 CONCLUSION

We revisit model merging through the lens of MAP inference in the parameter space under a product-of-experts (PoE) energy-based model. We showed that many existing methods can be rediscovered from this new perspective. It enriches the solution space for model merging and at the same time exposes critical limitations of existing work. In this new view, existing merging methods can be interpreted as imposing a Gaussian structure with light tails on the directional residual between the merged and individual models. To address this limitation, we introduce Cauchy-based experts that better capture the observed heavy-tailed behavior, resulting in improved performance. We developed practical algorithms to perform MAP inference on the new PoE design and proved convergence guarantees.

ACKNOWLEDGEMENT

This work utilized GPU compute resources at SDSC and ACES through allocation CIS230391 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services and Support (ACCESS) program [Boerner et al., 2023], which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. Trong Nghia Hoang is supported by the National Science Foundation CAREER Award IIS-2544071.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Arash Ahmadian, Seraphina Goldfarb-Tarrant, Beyza Ermis, Marzieh Fadaee, Sara Hooker, et al. Mix data or merge models? optimizing for diverse multi-task learning. *arXiv preprint arXiv:2410.10801*, 2024.

Timothy J. Boerner, Stephen Deems, Thomas R. Furlani, Shelley L. Knuth, and John Towns. Access: Advancing innovation: Nsf's advanced cyberinfrastructure coordination ecosystem: Services & support. In *Practice and Experience in Advanced Research Computing 2023: Computing for the Common Good*, PEARC '23, page 173–176, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399852. doi: 10.1145/3569951.3597559. URL https://doi.org/10.1145/3569951.3597559.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

Nico Daheim, Thomas Möllenhoff, Edoardo Maria Ponti, Iryna Gurevych, and Mohammad Emtiyaz Khan. Model merging by uncertainty-based gradient matching. *arXiv preprint arXiv:2310.12808*, 2023.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodola. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18695–18705, 2025.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14 (8):1771–1800, 2002.

Minh Hoang and Trong Nghia Hoang. Few-shot learning via repurposing ensemble of black-box models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 2024.

Nghia Hoang, Thanh Lam, Bryan Kian Hsiang Low, and Patrick Jaillet. Learning task-agnostic embedding of multiple black-box experts for multi-task model fusion. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4282–4292. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hoang20b.html.

Quang Minh Hoang, Trong Nghia Hoang, Kian Hsiang Low, and Carleton Kingsford. Collective model fusion for multiple black-box experts. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. *arXiv preprint arXiv:2212.09849*, 2022.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

Erwin Kreyszig. *Introductory functional analysis with applications*. John Wiley & Sons, 1991.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

Thanh Chi Lam, Nghia Hoang, Bryan Kian Hsiang Low, and Patrick Jaillet. Model fusion for personalized learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5948–5958. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/lam21a.html.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Lu Li, Tianyu Zhang, Zhiqi Bu, Suyuchen Wang, Huan He, Jie Fu, Yonghui Wu, Jiang Bian, Yong Chen, and Yoshua Bengio. Map: Low-compute model merging with amortized pareto fronts via quadratic approximation. *arXiv preprint arXiv:2406.07529*, 2024.

Weishi Li, Yong Peng, Miao Zhang, Liang Ding, Han Hu, and Li Shen. Deep model fusion: A survey. *arXiv preprint arXiv:2309.15698*, 2023.

Pengchen Liang, Bin Pu, Haishan Huang, Yiwei Li, Hualiang Wang, Weibo Ma, and Qing Chang. Vision foundation models in medical image analysis: Advances and challenges. In *International Conference on Blockchain and Trustworthy Systems*, pages 170–181. Springer, 2025.

Tongliang Liu and Dacheng Tao. On the robustness and generalization of cauchy regression. In *2014 4th IEEE International Conference on Information Science and Technology*, pages 100–105. IEEE, 2014.

Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

H. B. McMahan, Eider Moore, Daniel Ramage, S.C.D. Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2016. URL https://api.semanticscholar.org/CorpusID:14955348.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, et al. Scaling wearable foundation models. *arXiv preprint arXiv:2410.13638*, 2024.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 7. Granada, 2011.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. *Advances in Neural Information Processing Systems*, 36:66727–66754, 2023.

Bikang Pan, Wei Huang, and Ye Shi. Federated learning from vision-language foundation models: Theoretical analysis and method. *Advances in Neural Information Processing Systems*, 37:30590–30623, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with svd to tie the knots. *arXiv preprint arXiv:2410.19735*, 2024.

Chengwei Su, Rahul Gupta, Shankar Ananthakrishnan, and Spyridon Matsoukas. A re-ranker scheme for integrating large scale nlu models. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 670–676, 2018. URL https://api.semanticscholar.org/CorpusID:52823816.

Anke Tang, Li Shen, Yong Luo, Liang Ding, Han Hu, Bo Du, and Dacheng Tao. Concrete subspace learning based interference elimination for multi-task model fusion. *arXiv preprint arXiv:2312.06173*, 2023.

Yee Whye Teh, Max Welling, Simon Osindero, and Geoffrey E Hinton. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4(Dec):1235–1260, 2003.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Sahar Voghoei, Navid Hashemi Tonekaboni, Jason G Wallace, and Hamid R Arabnia. Deep learning at the edge. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 895–901. IEEE, 2018.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355, 2018.

Yongxian Wei, Anke Tang, Li Shen, Zixuan Hu, Chun Yuan, and Xiaochun Cao. Modeling multi-task model merging as adaptive projective gradient descent. *arXiv preprint arXiv:2501.01230*, 2025.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115, 2023.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in llms, mllms, and beyond: Methods, theories, applications, and opportunities. *ACM Computing Surveys*, 2024.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan H. Greenewald, and Trong Nghia Hoang. Statistical model aggregation via parameter matching. In *Neural Information Processing Systems*, 2019. URL `https://api.semanticscholar.org/CorpusID:202784500`.

Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical image analysis*, 91:102996, 2024.

A PROOF FOR THEOREM 5.1

Consider the MAP loss under PoE-EBM (22):

$$\boldsymbol{\zeta}^* \;=\; \arg\min_{\boldsymbol{\zeta}} \sum_{i=1}^{N} \log\Big(r_i(\boldsymbol{\zeta})^2 + \eta\Big) \;=\; \arg\min_{\boldsymbol{\zeta}} \sum_{i=1}^{N} \log\Big([(\boldsymbol{\zeta}-\boldsymbol{\theta}_i)^T\,\boldsymbol{\theta}_i]^2 + \eta\Big). \tag{27}$$

We write the squared residual term as

$$r_i(\boldsymbol{\zeta})^2 \;=\; [(\boldsymbol{\zeta}-\boldsymbol{\theta}_i)^T\,\boldsymbol{\theta}_i]^2 \;=\; (\boldsymbol{\zeta}-\boldsymbol{\theta}_i)^T\,\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top(\boldsymbol{\zeta}-\boldsymbol{\theta}_i) \tag{28}$$

A necessary condition for optimality is that the gradient of the global energy vanishes at $\boldsymbol{\zeta}^*$. This implies the Cauchy score $S^{Cauchy}(\boldsymbol{\zeta}^*) = 0$, resulting in the following stationary equation:

$$\sum_{i=1}^{N} \frac{2r_i(\boldsymbol{\zeta}^*)}{\gamma^2 + r_i(\boldsymbol{\zeta}^*)^2}\boldsymbol{\theta}_i \;=\; 0. \tag{29}$$

$$S^{Cauchy}(\boldsymbol{\zeta}^*) \;=\; \sum_{i=1}^{N} \frac{2\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top(\boldsymbol{\zeta}^*-\boldsymbol{\theta}_i)}{(\boldsymbol{\zeta}^*-\boldsymbol{\theta}_i)^\top\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top(\boldsymbol{\zeta}^*-\boldsymbol{\theta}_i)+\gamma} \;=\; 2\sum_{i=1}^{N} u_i(\boldsymbol{\zeta}^*)\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top(\boldsymbol{\zeta}^*-\boldsymbol{\theta}_i) \;=\; 0. \tag{30}$$

Solving for $\boldsymbol{\zeta}^*$ thus reveals the fixed-point equation as desired,

$$\boldsymbol{\zeta}^* \;=\; \left(\sum_{i=1}^{N} u_i(\boldsymbol{\zeta}^*)\,\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top\right)^{\dagger}\left(\sum_{i=1}^{N} u_i(\boldsymbol{\zeta}^*)\,\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top\boldsymbol{\theta}_i\right). \tag{31}$$

B PROOF FOR THEOREM 5.4

We prove that the fixed-point mapping

$$F(\boldsymbol{\zeta}^*) \;=\; \left(\sum_{i=1}^{N} u_i(\boldsymbol{\zeta}^*)\,\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top + \eta\boldsymbol{I}\right)^{-1}\left(\sum_{i=1}^{N} u_i(\boldsymbol{\zeta}^*)\,\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top\boldsymbol{\theta}_i\right) \;\text{ where} \tag{32}$$

$$u_i(\boldsymbol{\zeta}^*) \;=\; \frac{1}{\left[(\boldsymbol{\zeta}^*-\boldsymbol{\theta}_i)^\top\boldsymbol{\theta}_i\right]^2+\gamma} \;\triangleq\; \frac{1}{r_i(\boldsymbol{\zeta}^*)^2+\gamma}, \qquad \eta>0,\ \gamma>0,$$

is a contractive map under Assumptions 5.2 and 5.3, i.e.

$$\|F(\boldsymbol{\zeta}^*) - F(\boldsymbol{\nu})\| \;\le\; L\|\boldsymbol{\zeta}^*-\boldsymbol{\nu}\| \;\text{ where }\; 0 \;<\; L \;<\; 1. \tag{33}$$

Choosing $\boldsymbol{\nu} = \boldsymbol{\zeta}^{(k)}$ and factoring in that $\boldsymbol{\zeta}^{(k+1)} = F(\boldsymbol{\zeta}^{(k)})$ then leads to the desired result. For notational ease, we denote $\boldsymbol{H}(\boldsymbol{\zeta}^*) = \sum_{i=1}^{N} u_i(\boldsymbol{\zeta}^*)\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top + \eta\boldsymbol{I}$ and $\boldsymbol{b}(\boldsymbol{\zeta}^*) = \sum_{i=1}^{N} u_i(\boldsymbol{\zeta}^*)\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top\boldsymbol{\theta}_i$. This results in

$$F(\boldsymbol{\zeta}^*) \;=\; \Big(\boldsymbol{H}(\boldsymbol{\zeta}^*)\Big)^{-1}\boldsymbol{b}(\boldsymbol{\zeta}^*)\,. \tag{34}$$

We first need the following lemma for computing the directional gradient of F.

Lemma B.1. *Given the mapping F defined in* (34)*. Pick an arbitrary weight $\boldsymbol{\zeta}^*$ and direction $\mathbf{h}$. Then the directional gradient of F at $\boldsymbol{\zeta}^*$ in the direction $\mathbf{h}$ is given by*

$$DF(\boldsymbol{\zeta}^*)_{\boldsymbol{h}} \;=\; -2\boldsymbol{H}^{-1}\sum_{i=1}^{N} \frac{r_i(\boldsymbol{\zeta}^*)}{(r_i(\boldsymbol{\zeta}^*)^2+\gamma)^2}(\boldsymbol{\theta}_i^\top\boldsymbol{h})\Big(\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top\boldsymbol{\theta}_i - \boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top F(\boldsymbol{\zeta}^*)\Big). \tag{35}$$

Proof. Applying the definition of directional derivative:

$$DF(\boldsymbol{\zeta}^*)_{\boldsymbol{h}} \quad = \quad \frac{d}{dt}F(\boldsymbol{\zeta}^* + t\boldsymbol{h})\Big|_{t=0}.$$

Denote $\boldsymbol{H}_t \triangleq \boldsymbol{H}(\boldsymbol{\zeta}^* + t\boldsymbol{h})$ and $\boldsymbol{b}_t \triangleq \boldsymbol{b}(\boldsymbol{\zeta}^* + t\boldsymbol{h})$. Then $F_t \triangleq F(\boldsymbol{\zeta}^* + t\boldsymbol{h}) = \boldsymbol{H}_t^{-1}\boldsymbol{b}_t$. Applying chain rule, we have

$$\frac{dF_t}{dt} \quad = \quad \frac{d}{dt}\boldsymbol{H}_t^{-1}\boldsymbol{b}_t \; + \; \boldsymbol{H}_t^{-1}\frac{d\boldsymbol{b}_t}{dt} \; = \; -\boldsymbol{H}_t^{-1}\frac{d\boldsymbol{H}_t}{dt}\underbrace{\boldsymbol{H}_t^{-1}\boldsymbol{b}_t}_{F_t} + \boldsymbol{H}_t^{-1}\frac{d\boldsymbol{b}_t}{dt}\,, \tag{36}$$

where we apply the matrix inverse derivative identity. At $t = 0$, (36) becomes

$$DF(\boldsymbol{\zeta}^*)_{\boldsymbol{h}} \quad = \quad -\boldsymbol{H}(\boldsymbol{\zeta}^*)^{-1}[D\boldsymbol{b}(\boldsymbol{\zeta}^*)_{\boldsymbol{h}} - D\boldsymbol{H}(\boldsymbol{\zeta}^*)_{\boldsymbol{h}}F(\boldsymbol{\zeta}^*)], \tag{37}$$

where

$$\begin{aligned} D\boldsymbol{b}(\boldsymbol{\zeta}^*)_{\boldsymbol{h}} \quad &= \quad \sum_{i=1}^{N}(\nabla u_i(\boldsymbol{\zeta}^*)^\top \boldsymbol{h})\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top\boldsymbol{\theta}_i \; = \; \sum_{i=1}^{N}\frac{2r_i(\boldsymbol{\zeta}^*)}{(r_i(\boldsymbol{\zeta}^*)^2+\gamma)^2}(\boldsymbol{\theta}_i^\top\boldsymbol{h})\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top\boldsymbol{\theta}_i, \\ D\boldsymbol{H}(\boldsymbol{\zeta}^*)_{\boldsymbol{h}} \quad &= \quad \sum_{i=1}^{N}(\nabla u_i(\boldsymbol{\zeta}^*)^\top \boldsymbol{h})\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top \; = \; \sum_{i=1}^{N}\frac{2r_i(\boldsymbol{\zeta}^*)}{(r_i(\boldsymbol{\zeta}^*)^2+\gamma)^2}(\boldsymbol{\theta}_i^\top\boldsymbol{h})\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top. \end{aligned} \tag{38}$$

Substituting (38) to (37) thus completes the proof. ■

It is now sufficient to show that the directional gradient $\|DF(\boldsymbol{\zeta}^*)\|$ of F in an arbitrary direction $\boldsymbol{h}$ in the neighborhood of $\boldsymbol{\zeta}^*$ is less than 1 to show local contraction of F [Kreyszig, 1991].

First, we compute the gradient of $u_i(\boldsymbol{\zeta}^*)$ as:

$$\nabla u_i(\boldsymbol{\zeta}^*) \quad = \quad \frac{2r_i(\boldsymbol{\zeta}^*)}{(r_i(\boldsymbol{\zeta}^*)^2+\gamma)^2}\cdot\boldsymbol{\theta}_i \tag{39}$$

Next, we define $\boldsymbol{h} = \boldsymbol{\nu} - \boldsymbol{\zeta}^*$ for arbitrary weights $\boldsymbol{\zeta}^*$ and $\boldsymbol{\nu}$ such that $\|\boldsymbol{h}\| = 1$ (as per Assumption 5.3). We bound the norm of the directional gradients $DF(\boldsymbol{\zeta}^*)_{\boldsymbol{h}}$ in direction $\mathbf{h}$. Using Lemma B.1, we can compute $\nabla F(\boldsymbol{\zeta}^*)_{\boldsymbol{h}}$ as

$$DF(\boldsymbol{\zeta}^*)_{\boldsymbol{h}} \quad = \quad -2\boldsymbol{H}^{-1}\sum_{i=1}^{N}\underbrace{\frac{r_i(\boldsymbol{\zeta}^*)}{(r_i(\boldsymbol{\zeta}^*)^2+\gamma)^2}\Big(\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top\boldsymbol{\theta}_i - \boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top F(\boldsymbol{\zeta}^*)\Big)}_{\triangleq\boldsymbol{J}_i}(\boldsymbol{\theta}_i^\top\boldsymbol{h}). \tag{40}$$

Therefore, $DF(\boldsymbol{\zeta}^*)_{\boldsymbol{h}}$ is bounded by

$$\|DF(\boldsymbol{\zeta}^*)_{\boldsymbol{h}}\| \quad = \quad 2\left\|\sum_{i=1}^{N}\boldsymbol{H}^{-1}\boldsymbol{J}_i(\boldsymbol{\theta}_i^\top\boldsymbol{h})\right\| \quad \leq \quad 2\sum_{i=1}^{N}\|\boldsymbol{H}^{-1}\boldsymbol{J}_i\boldsymbol{\theta}_i^\top\| \quad \leq \quad 2M\sum_{i=1}^{N}\|\boldsymbol{H}^{-1}\boldsymbol{J}_i\|\,, \tag{41}$$

where the first inequality follows from triangle inequality and $\|\boldsymbol{h}\| = 1$, the second inequality follows from Assumption 5.2. We empirically show in Figure 4 that this upper bound is less than 1 in our experiment settings, making F a contracting map around the neighborhood of $\boldsymbol{\zeta}^*$.

C COMPLEXITY ANALYSIS OF ALGORITHM 1.

Let $\boldsymbol{\theta}_i \in \mathbb{R}^{d\times d}$ be a fine-tuning module. Outer loop over maximum iterations (line 4-12) is executed T times. For each outer iteration, the inner loop (line 5-8) over tasks is executed N times.

Inner loop complexity:

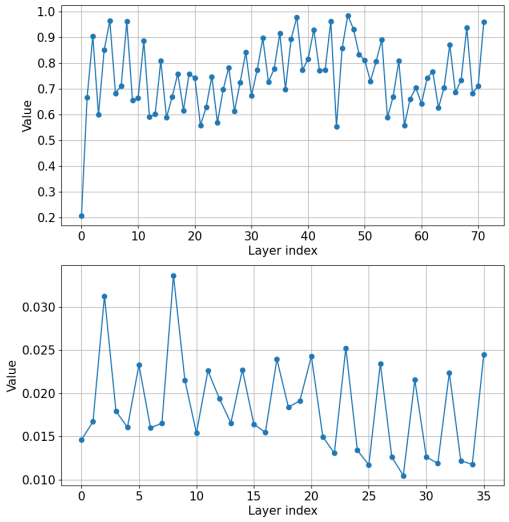

Figure 4: Lipschitz constant values of F across multiple layers of our PoE-EBM when merging 7 fullfinetuned ViT models. Top: ViT-L-14. Bottom: ViT-B-32. The Lipschitz constant is consistently less than 1, indicating that F is a contracting map, ensuring convergence of our algorithm 1 (see Figure 3).

1. Residual computation (line 6): $\mathcal{O}(d^2)$.
2. Task weight computation (line 7): $\mathcal{O}(1)$.

After the inner loop

1. Compute $\boldsymbol{H}$ and $\boldsymbol{b}$ (line 9): $\mathcal{O}(Nd^2)$
2. Update merged fine-tuning module (line 11): $\mathcal{O}(d^3)$.

Therefore, total complexity is $\mathcal{O}\left(T\left(N(d^2+1)+Nd^2+d^3\right)\right) \;=\; \mathcal{O}(TNd^3)$

D INSIGHTS INTO CAUCHY SCORE 21

D.1 GAUSSIAN ENERGY AND BASE SCORE

For Gaussian experts (11) with the precision matrix chosen as $\boldsymbol{\Sigma}_i^{-1} = \frac{1}{\gamma^2}\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top$, the corresponding energy function, which we refer to as the *base energy*, for the i-th task has the following form:

$$E_i^{base}(\boldsymbol{\zeta}) \;\triangleq\; \frac{1}{2\gamma^2}(\boldsymbol{\zeta}-\boldsymbol{\theta}_i)^\top\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top(\boldsymbol{\zeta}-\boldsymbol{\theta}_i) \;=\; \frac{1}{2\gamma^2}r_i(\boldsymbol{\zeta})^2\,. \tag{42}$$

The global base energy is simply the sum over task energies $E^{base}(\boldsymbol{\zeta}) = \sum_{i=1}^N \frac{1}{2\gamma^2}r_i(\boldsymbol{\zeta})^2$, which admits the following:

$$S^{base}(\boldsymbol{\zeta}) \;\triangleq\; -\nabla_{\boldsymbol{\zeta}}E^{base} \;=\; -\sum_{i=1}^N \frac{r_i(\boldsymbol{\zeta})}{\gamma^2}\boldsymbol{\theta}_i\,. \tag{43}$$

D.2 CAUCHY EXPERTS AS ADDITIVE ROBUST GUIDANCE

To understand of the effect of heavy-tail experts, we compare the Cauchy score (21) with the base score (43). The Cauchy score (21) can be expressed as

$$S^{Cauchy} \;=\; S^{base} \,+\, G(\boldsymbol{\zeta})\,,$$

Table 5: Hyperparameter settings (λ, γ, η) for PoE−EBM when used to merge fine-tuned models in vision and NLP benchmarks. The numerical algorithm (Section 5) used to compute the MAP of PoE−EBM model is configured with convergence tolerance parameter 10^{-5} in all scenarios.

Domain	**Training**	**Model**	**Scale** λ	γ	η
Vision	FFT	ViT-L/14	1.0	0.01	10^{-3}
Vision	FFT	ViT-B/32	1.0	0.3	10^{-3}
Vision	LoRA	ViT-L/14	0.25	0.01	10^{-3}
NLP	LoRA	Flan-T5-base	1.0	10^{-3}	10^{-3}
NLP	LoRA	Flan-T5-large	1.0	10^{-2}	10^{-3}

where we define the guidance term as:

$$G(\boldsymbol{\zeta}) \triangleq \sum_{i=1}^{N} \left(\frac{r_i(\boldsymbol{\zeta})}{\gamma^2} - \frac{2r_i(\boldsymbol{\zeta})}{\gamma^2 + r_i(\boldsymbol{\zeta})^2} \right) \boldsymbol{\theta}_i \, .$$

Simplifying the expression yields:

$$G(\boldsymbol{\zeta}) = \sum_{i=1}^{N} \frac{r_i(\boldsymbol{\zeta}) \Big(r_i(\boldsymbol{\zeta})^2 - \gamma^2 \Big)}{\gamma^2 \Big(\gamma^2 + r_i(\boldsymbol{\zeta})^2 \Big)} \boldsymbol{\theta}_i \, . \tag{44}$$

Intuitively, $G(\boldsymbol{\zeta})$ acts as a residual-dependent adjustment that pushes $\boldsymbol{\zeta}$ away from directions with large $|r_i(\boldsymbol{\zeta})|$. In contrast, the base Gaussian score (43) amplifies large residuals linearly, pulling the solution toward conflicts regime. The Cauchy guidance thus provides automatic robustness where it tempers the influence of misaligned tasks while still encouraging alignment in well-aligned directions.

E LIMITATIONS

PoE−EBM currently assumes offline model merging with simultaneous access to all task models and a shared model architecture. Moreover, our formulation focuses on Cauchy experts as the underlying heavy-tailed distribution. Extending the framework to continual merging, alternative heavy-tailed experts, heterogeneous architectures, and multi-modal settings remains potential scopes for future work.

F ADDITIONAL EXPERIMENT DETAILS

Tasks. We conduct our experiments on vision and natural language processing (NLP) tasks. The downstream vision tasks contain the 7-task benchmark: MNIST [LeCun, 1998], SVHN [Netzer et al., 2011], Stanford Cars [Krause et al., 2013], DTD [Cimpoi et al., 2014], GTRSB [Stallkamp et al., 2011], EuroSAT [Helber et al., 2019], Resisc45 [Cheng et al., 2017]. We also include an additional 6 datasets to create a more challenging 13-task benchmark: CIFAR10, CIFAR100 [Krizhevsky et al., 2009], FashionMNIST [Xiao et al., 2017], Flowers102 [Nilsback and Zisserman, 2008], Food [Bossard et al., 2014] and Oxford-IIIT Pet. For NLP task, we use 8 datasets from the GLUE benchmark [Wang et al., 2018], icluding CoLA, MNLI, MRPC, QNLI, QQP, RTE, SST2 and STSB.

Models. For vision experiments, we leverage pretrained ViT-B/32 and ViT-B/14 models from CLIP Radford et al. [2021]. We consider merging models in both fully-finetuned and LoRA fine-tuned [Hu et al., 2022] settings. We use the checkpoints provided by Ilharco et al. [2022] for fully-finetuned models. The LoRA-finetuned version of these models are provided by Stoica et al. [2024]. For NLP tasks, we merge LoRA-finetuned Flan-T5-base models and Flan-T5-large models whose checkpoints are provided by Wei et al. [2025].

Metrics. We report absolute accuracy of the merging methods as well as those of individual fine-tuned models. Following [Ilharco et al., 2022], we also report the "normalized accuracy", i.e. the ratio between absolute accuracy of the merged model on task i-th and the finetuned model accuracy on the same task. Normalized accuracy shows how close the merged model performs in relative to the finetuned model for each task. Additional experiment details are provided in Appendix F.

Table 6: Average ℓ_2 norms of full fine-tuning task vectors $\boldsymbol{\theta}_i$ across seven vision tasks, together with their magnitudes relative to the pretrained model $\boldsymbol{W}_0$. All values are averaged across layers. The consistently small ratios (<1%) indicate that task-specific updates remain localized perturbations of the pretrained model, supporting the fine-tuning parameterization adopted throughout this work.

Backbone	ℓ_2-norm & ratio of ℓ_2-norms	MNIST	SVHN	Cars	DTD	GTSRB	EuroSAT	RESISC45
ViT-B-32	$\|\boldsymbol{\theta_i}\|_2$	0.124	0.136	0.140	0.154	0.117	0.116	0.128
	$\|\boldsymbol{\theta}_i\|_2/\|\boldsymbol{W}_0\|_2$	0.65%	0.71%	0.73%	0.81%	0.61%	0.61%	0.67%
ViT-L-14	$\|\boldsymbol{\theta_i}\|_2$	0.128	0.149	0.156	0.170	0.117	0.134	0.152
	$\|\boldsymbol{\theta}_i\|_2/\|\boldsymbol{W}_0\|_2$	0.60%	0.70%	0.73%	0.80%	0.55%	0.63%	0.80%

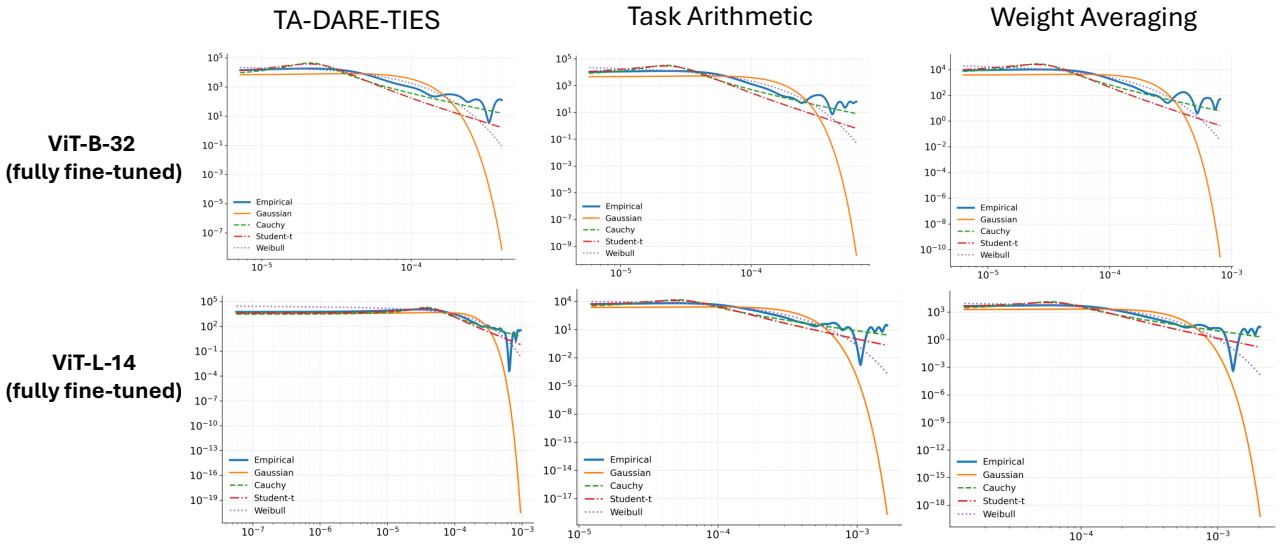

Figure 5: Empirical distributions of directional residuals r (see Eq. 17) produced by different merging methods when merging seven fully fine-tuned ViT models. Across all methods, the observed tails are substantially heavier than those predicted by light-tailed models and are most accurately captured by a Cauchy distribution among the distributions considered.

Hyperparameters. All hyperparameters are selected via extensive grid search on validation performance for each architecture and training regime. We use the same convergence criterion across all experiments and report the best-performing configuration for each setting. Convergence criteria: For all experiments with `PoE-EBM`, fixed-point iterations are terminated when the norm of the difference between two iterates is smaller or equals to 10^{-5}. The conditioning parameter η is fixed to 10^{-3}. Other hyperparameter configurations for using `PoE-EBM` on vision and NLP tasks are reported in Table 5.

G ADDITIONAL RESULTS

We present additional experimental results and ablation studies in this section. The normalized accuracy for fully fine-tuned ViT-B/32 and ViT-B/14 models is reported in Tables 10 and 11, respectively. Results for merging LoRA-finetuned and fully fine-tuned ViT-L/14 models on the 7-task vision benchmark are shown in Tables 12 and 8. We report the merging results on the 13 vision benchmarks in Table 2. Table 9 summarizes performance on 8 GLUE datasets when merging LoRA-finetuned Flan-T5-large models. Across all settings, `PoE-EBM` consistently outperforms the compared baselines, including KnOTS [Stoica et al., 2024], which is specifically designed for merging LoRA-finetuned models. We also provide additional empirical results demonstrating the residuals' heavy-tail behavior when merging Flan-T5 models and ViTs which closely aligns with the Cauchy distribution in Figs. 5 and 6.

Ablation on the covariance scale parameter. We study the effect of the scale parameter γ in (19) on the averaged absolute accuracy of our merging algorithm on both ViT-B-32 and ViT-L-14. We vary the value of γ, which controls the computation of the task-wise weights $u_i(\cdot)$, over the range $[10^{-3}, 3]$ and plot the corresponding accuracies in Figure 7. We observe that the performance is relatively stable across a wide range of γ values for both ViT models. For ViT-B-32, accuracy peaks at

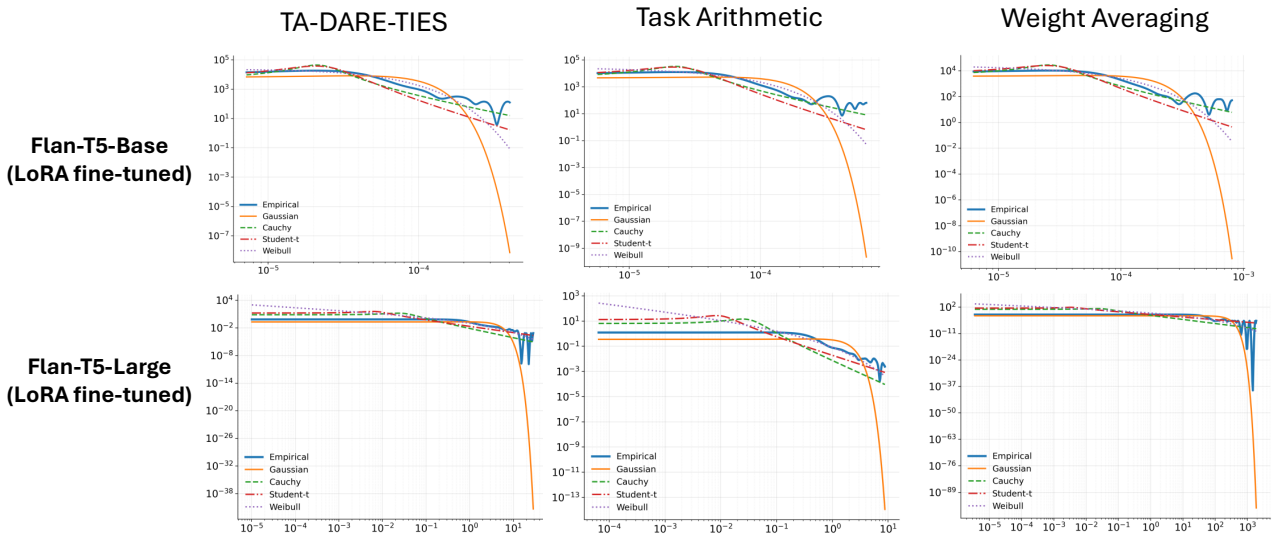

Figure 6: Empirical distributions of directional residuals r (see Eq. 17) produced by different merging methods when merging 8 LoRA fine-tuned Flan-T5 models.

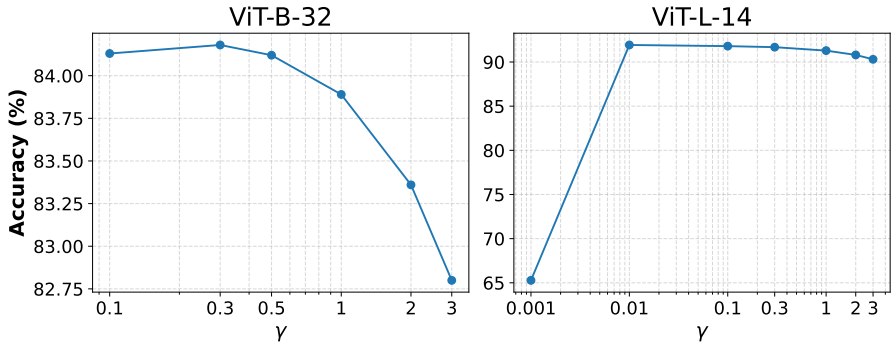

Figure 7: Effect of γ values on the merging accuracies of PoE−EBM on the performance accuracy when merging 7 full fine-tuned ViT models.

$\gamma = 0.3$ and gradually decreases as γ increases to 3. Even at $\gamma = 2$, PoE−EBM still outperforms all the baselines compared with accuracy 83.36%. For ViT-L-14, the performance remains relatively stable as the accuracy gradually increases with smaller values of γ. However, at extremely small values (e.g., $\gamma = 10^{-3}$), performance collapses, indicating numerical instability and over-sensitivity in the weight u_i's computation. These results suggest that moderate γ values result in stable performance of our merging algorithm, while overly small scales can be detrimental, particularly for larger models.

H ADDITIONAL RELATED WORKS

To resolve task interference in task arithmetic [Ilharco et al., 2022], TIES [Yadav et al., 2023] improves upon task arithmetic by reducing interference between parameters using their signs and magnitudes before merging. DARE [Yu et al., 2024] randomly removes fine-tuned weights and rescales the existing ones to create sparse tasks vectors, improving generalization of task arithmetic. Other line of work aims to resolve interference by merging within subspaces. Alternatively, Ortiz-Jimenez et al. [2023] proposes finetuning models in the tangent space, disentangling finetuned models and thus improving their mergeability. Task Singular Vectors [Gargiulo et al., 2025] combines task vectors using low-rank approximation and reduces interference through means of whitening. KnOTS [Stoica et al., 2024] leverages the Singular Value Decomposition (SVD) of the concatenated task vectors to extract the shared information across all tasks and merge only the task-specific components using task arithmetic. Concrete [Tang et al., 2023] uses meta learning to find a common low-dimensional subspace and perform merging with reduced interference.

Table 7: Reported processing time for our model merging algorithm (see Algorithm 1) across NLP and vision benchmarks.

Domain	Benchmark	Model	Time (s)
NLP	GLUE (8 tasks)	Flan-T5-base	29.26
NLP	GLUE (8 tasks)	Flan-T5-large	59.19
Vision	7 tasks	ViT-B/32	12.10
Vision	7 tasks	ViT-L/14	42.93
Vision	13 tasks	ViT-B/32	16.39
Vision	13 tasks	ViT-L/14	50.79

Table 8: Multi-task performance comparison when merging ViT-L/14 (full fine-tuned) across 7 vision benchmarks (absolute accuracy). Task-specific fine-tuning accuracy is colored blue to highlight performance upper-bound.

	ViT-L/14							
Method	**MNIST**	**SVHN**	**Cars**	**DTD**	**GTSRB**	**EuroSAT**	**RESISC45**	**Average**
Finetuning	99.762	97.881	90.113	97.766	99.129	99.852	96.762	97.320
Weight averaging	79.86	59.14	75.61	54.89	57.32	54.52	68.14	64.21
Task Arithmetic (TA)	92.90	70.65	77.80	60.27	66.38	68.93	74.33	73.04
TA-DARE-TIES	98.63	88.13	80.63	72.71	83.79	87.78	83.51	85.03
Fisher Merging	83.74	62.23	77.71	57.34	58.52	**99.04**	69.19	72.54
DOGE-TA	98.88	**94.45**	**87.85**	72.23	93.97	96.41	**91.94**	90.82
Concrete-TA	98.99	88.47	82.74	66.54	87.13	93.89	89.39	86.74
PoE-EBM (Ours)	**99.44**	94.41	85.53	**83.56**	**94.35**	97.00	89.22	**91.93**

More recent works explore model merging through the lens of optimization. DOGE [Wei et al., 2025] view model merging as a single constrained optimization problem where the objective is aligning the test performance of the merged model with the task-specific models on their respective tasks. MAP [Li et al., 2024] model the merging problem as a multi-objective optimization problem and aim to identify the Pareto front of the merging coefficients using proxy data. Nevertheless, these works leverage the geometry of the parameter space and employ non-probabilistic merging scheme and thus do not take into account the uncertainty. On the other hand, our PoE-EBM framework explicitly views the merging problem as probabilistic inference in the parameter space and takes uncertainty into account.

Table 9: Multi-task performance comparison when merging Flan-T5-large models (LoRA-fine-tuned) across 8 GLUE benchmarks. Task-specific fine-tuning accuracy is colored blue which highlights performance upper bound. Our PoE-EBM achieves the best rank across tasks.

Method	**CoLA**	**MNLI**	**MRPC**	**QNLI**	**QQP**	**RTE**	**SST-2**	**STS-B**	**Average**	**Rank**
Finetuning	80.20	88.51	89.23	94.40	87.18	91.74	95.19	90.91	89.67	NA
Weight Averaging	74.59	84.28	84.07	92.79	**86.28**	87.36	94.84	87.98	86.52	4.00
Task Arithmetic	76.89	85.44	85.29	93.92	85.84	88.09	95.18	87.75	87.30	3.38
TIES-Merging	75.55	84.69	84.31	**93.94**	86.18	**88.45**	95.07	87.82	86.93	3.13
Concrete-TA	76.89	86.16	**88.54**	93.92	85.84	88.09	95.18	87.91	87.44	2.50
DOGE-TA	**78.12**	88.08	86.52	93.80	85.82	86.72	95.00	87.71	87.72	3.63
PoE-EBM (Ours)	77.28	**88.33**	86.76	93.06	85.98	87.36	**95.30**	**88.98**	**87.88**	**2.25**

Table 10: Multi-task performance comparison when merging ViT-B/32 (full fine-tuned) across 7 vision benchmarks. The performance is reported in terms of normalized accuracy.

Method	ViT-B/32							
	MNIST	**SVHN**	**Cars**	**DTD**	**GTSRB**	**EuroSAT**	**RESISC45**	**Average**
Finetuning	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Weight averaging	85.28	65.71	78.71	53.20	55.89	63.59	71.56	67.70
Task Arithmetic (TA)	90.95	73.52	78.81	56.53	63.25	66.30	72.53	71.70
TA-DARE-TIES	96.29	84.16	77.81	62.60	71.48	70.64	70.75	76.25
Fisher Merging	90.06	84.49	90.56	56.04	58.50	77.28	77.57	75.91
DOGE-TA	98.73	89.80	**92.37**	66.10	88.55	90.13	**86.30**	87.42
Concrete-TA	97.31	82.09	75.95	60.20	71.92	76.39	75.01	77.08
PoE-EBM (Ours)	**99.20**	**94.03**	86.34	**78.19**	**89.60**	**90.24**	81.64	**88.46**

Table 11: Multi-task performance comparison when merging ViT-L/14 (full fine-tuned) across 7 vision benchmarks. The performance is reported in terms of normalized accuracy.

Method	ViT-B/32							
	MNIST	**SVHN**	**Cars**	**DTD**	**GTSRB**	**EuroSAT**	**RESISC45**	**Average**
Finetuning	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Weight averaging	80.05	60.42	83.91	56.15	57.83	54.60	70.42	66.20
Task Arithmetic (TA)	93.12	72.18	86.34	61.64	66.96	69.03	76.82	75.16
TA-DARE-TIES	98.86	90.04	89.48	74.37	84.43	87.91	86.30	87.36
Fisher Merging	83.94	63.58	86.24	58.65	59.03	**99.18**	71.51	74.89
DOGE-TA	99.12	**96.49**	**97.49**	73.88	94.80	96.55	**95.01**	93.34
Concrete-TA	99.23	90.38	91.82	68.06	87.90	94.03	92.38	89.12
PoE-EBM (Ours)	**99.68**	96.45	94.91	**85.47**	**95.18**	97.14	92.21	**94.43**

Table 12: Multi-task performance comparison when merging ViT-L/14 (LoRA-fine-tuned) across 7 vision benchmarks.

Method	Accuracy Type	ViT-L/14							
		MNIST	**SVHN**	**Cars**	**DTD**	**GTSRB**	**EuroSAT**	**RESISC45**	**Average**
Finetuning		99.53	97.72	99.77	70.05	97.20	98.59	95.70	94.08
Weight averaging	Absolute	78.55	60.19	78.22	55.80	52.19	**62.74**	72.40	65.73
	Normalized	78.93	61.59	78.40	79.65	53.69	**63.64**	75.65	70.22
Task Arithmetic (TA)	Absolute	77.93	59.70	78.14	55.69	51.85	62.70	72.06	65.44
	Normalized	78.30	61.09	78.33	79.50	53.35	63.60	75.30	69.62
TA-DARE-TIES	Absolute	70.86	71.48	81.86	57.45	60.38	57.22	79.13	68.30
	Normalized	71.20	73.14	81.75	82.01	62.12	58.04	82.68	72.99
KnOTS-TIES	Absolute	81.71	75.38	**83.83**	**58.30**	**68.66**	61.74	79.79	72.77
	Normalized	82.10	77.14	**84.03**	**83.22**	**70.63**	62.62	83.38	77.59
KnOTS-DARE-TIES	Absolute	67.43	67.06	82.68	58.19	64.13	59.67	80.00	68.45
	Normalized	67.75	68.62	82.88	83.07	65.98	60.52	83.60	73.20
PoE-EBM (Ours)	Absolute	**89.56**	**75.56**	83.37	56.97	66.40	60.15	**80.37**	**73.20**
	Normalized	**89.99**	**77.32**	83.56	81.32	68.31	61.00	**83.98**	**77.93**