Information-Geometric Perspectives on Merging Variational Foundation Models

Nour Jamoussi
EURECOM, France
nour.jamoussi@eurecom.fr

Giuseppe Serra
EURECOM, France
giuseppe.serra@eurecom.fr

Photios A. Stavrou
EURECOM, France
fotios.stavrou@eurecom.fr

Marios Kountouris
University of Granada, Spain
EURECOM, France
mariosk@ugr.es

Abstract

We propose an information-geometric framework for merging variational foundation models that preserves global robustness while integrating domain-specific knowledge in a principled manner. Assuming that the foundation models have been pretrained or fine-tuned using the Improved Variational Online Newton (IVON) optimizer, matching Adam's computational cost while providing Bayesian advantages, we formulate the merging problem between the pretrained and fine-tuned models as an information-geometric projection. Under mild assumptions, this reduces to computing a barycenter in the variational parameter space, yielding a computationally efficient and theoretically grounded merging rule. The framework naturally extends to multi-model barycentric merging, minimizing the average discrepancy among fine-tuned models.

1 Introduction

Foundation Models (FMs) have emerged as powerful general-purpose learners, capable of adapting to a wide range of downstream tasks after large-scale pretraining. However, as data distributions shift and new domains appear, keeping these models up to date without retraining from scratch remains a major challenge [2]. Approaches such as continual pretraining [9, 19], fine-tuning [6], and model merging [15, 13, 3, 5] offer promising paths forward, allowing FMs to integrate new knowledge while retaining broad generalization. Yet, updating such large systems at scale faces key obstacles: high computational cost, catastrophic forgetting [14], and potential misalignment in uncertainty quantification. Addressing these challenges requires principled and efficient update rules that incorporate domain-specific adaptations into a global FM while preserving statistical rigor and scalability.

Within the spectrum of adaptation strategies, variational approaches provide a principled framework for representing model uncertainty, making them particularly well-suited to settings where both reliability and interpretability are critical. We focus on variational FMs whose parameters encode posterior distributions, enabling updates to be expressed as operations on the statistical manifold of distributions. We assume that these models are pretrained or fine-tuned using the Improved Variational Online Newton (IVON) optimizer [18], which is grounded in the Bayesian learning rule [10] and matches Adam's computational cost while providing strong Bayesian performance at scale. In this setting, we formulate FM merging as an information-geometric projection from a global model (i.e., the pretrained model) onto a sphere centered at a specialized model (i.e, the fine-tuned

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on Continual and Compatible Foundation Model Updates (CCFM).

model). Under mild assumptions, this projection reduces to computing a barycenter in the variational parameter space. The formulation naturally extends from single specialization to multi-model aggregation via barycentric averaging, minimizing the average information-geometric discrepancy across multiple fine-tuned models. Our approach thus yields an interpretable, computationally efficient, and theoretically grounded merging mechanism that generalizes existing techniques such as Fisher-weighted averaging [13] and mixture or product-of-experts.

2 Related Work

Variational Foundation Models. Despite their strong empirical performance, deep learning methods often fall short in practical aspects such as reliability and calibrated uncertainty quantification. In contrast, Bayesian learning provides well-calibrated models and principled, risk-aware adaptation, but its computational overhead has limited its scalability to FM levels [17]. The IVON optimizer [18] brings variational learning to near-Adam efficiency and has demonstrated large-scale viability by pretraining GPT-2 on OpenWebText and ResNet on ImageNet from scratch, as well as fine-tuning large masked language models (e.g., DeBERTaV3). IVON has also proven to be effective in curvature-based (Fisher/Hessian) model merging [18, 3].

Model Merging. Originally introduced in the context of Federated Learning (FL) to reduce communication overhead and enhance privacy [15], model merging has since been adopted in domains including computer vision and large language models (LLMs) [5]. Wortsman *et al.* [20] show that averaging the weights of models fine-tuned under varied hyperparameters improves both accuracy and out-of-distribution robustness. Using only a small number of fine-tuned models, Jang *et al.* [8] achieve robust merges via layer-wise linear interpolation that explicitly operates in the Euclidean parameter space. In contrast, we focus on the manifold geometry of variational posteriors. Building on the barycentric aggregation framework proposed in Bayesian FL [7], we formulate merging between a pretrained and a fine-tuned variational model as an information-geometric projection in posterior space, and show that, under mild assumptions, it is equivalent to a barycentric merge. We further extend this formulation to multi-model barycentric aggregation, generalizing several established FM-merging techniques, while respecting the inherently non-Euclidean geometry of the variational parameter space.

3 Variational FM Specialization through Information-Geometric Projections

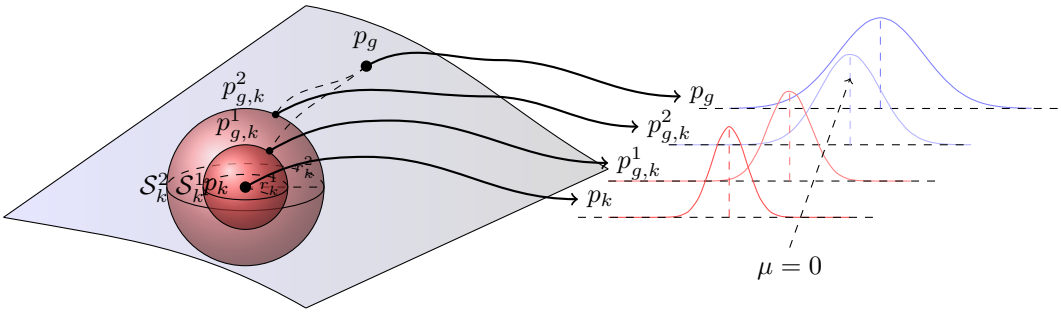

Figure 1: Specialization through information-geometric projection. The figure presents two projection scenarios illustrated with two local spheres $\mathcal{S}^1_k$ and $\mathcal{S}^2_k$ of increasing radius r^1_k and r^2_k, highlighting the impact of the radius on the closeness of the projected distribution to the global or specialized distribution.

We interpret the adaptation of a global FM to a domain-specific variant as a *projection problem* on a specialization set within the manifold of model posteriors. To formalize this, let p_g denote the variational posterior of the global FM, and let p_k denote the posterior of a specialized FM (e.g., one fine-tuned on a particular domain). We define the *specialization set* for p_k as a sphere centered at p_k with radius r_k:

Definition 1 (Specialization set $\mathcal{S}_D(p,r)$)**.** *Given a statistical manifold $\mathcal{M}$ and a divergence measure D, the specialization set $\mathcal{S}_D(p,r)$ for $p \in \mathcal{M}$ and radius $r \in [0,\infty)$ is*

$$\mathcal{S}_D(p,r) = \{q \in \mathcal{M} \,:\, D(q \,\|\, p) \leq r\} \subseteq \mathcal{M}.$$

The radius r_k encodes the degree of specialization: smaller values enforce stronger adherence to p_k, while larger values allow greater influence from p_g. We denote $\mathcal{S}_k = \mathcal{S}_D(p_k, r_k)$ for brevity.

Problem 1 (Projection for FM specialization)**.** *Given a statistical manifold $\mathcal{M}$, a divergence measure D, a global posterior $p_g \in \mathcal{M}$, and a specialization set $\mathcal{S}_k$, the projection problem is formulated as*

$$\min_{p \in \mathcal{S}_k} D(p \,\|\, p_g). \tag{1}$$

The solution, $p_{g,k} = \arg\min_{p \in \mathcal{S}_k} D(p \,\|\, p_g)$, is referred to as the specialized global posterior *for domain k.*

Remark 1. *Varying r_k between 0 and $D(p_k \,\|\, p_g)$ traces a geodesic, i.e., the shortest path, between p_k and p_g on the manifold $\mathcal{M}$, interpolating between the specialized and global posteriors.*

To support our derivation and establish the link between the projection and the barycenter formulations, we now recall the definition of a barycenter with respect to a given divergence D.

Definition 2. *(D-barycenter) Given a statistical manifold $\mathcal{M}$, a divergence measure D, and a set of distributions $\{p_k\}_{k=1}^N \subseteq \mathcal{M}$ with associated normalized weights $\{w_k\}_{k=1}^N$, the D-barycenter of the set $\{p_k\}_{k=1}^N$ is defined as*

$$p_D^*(\{p_k\}_{k=1}^N, \{w_k\}_{k=1}^N) = \arg\min_{q \in \mathcal{M}} \sum_{k=1}^N w_k D(q||p_k). \tag{2}$$

The following mild assumption is crucial for establishing the equivalence between projection and barycenter formulations, as shown in Theorem 1.

Assumption 1. *The divergence measure D is convex in its first argument.*

Remark 2. *Most commonly used divergences, including the family of f-divergences and the Wasserstein-p distances, are convex in both arguments.*

Theorem 1. *Under Assumption 1, the solution of the projection problem* (1)*) is equivalent to the weighted barycenter problem* (2)*, i.e.,*

$$p_{g,k} = p_D^*(\{p_g, p_k\}, \{w_g, w_k\}) \tag{3}$$

where the weights w_g and w_k are given by $w_g = \frac{1}{\lambda+1}, w_k = \frac{\lambda}{\lambda+1}$, for some $\lambda \in [0,\infty)$.

We highlight the following observations regarding the relationship between r_k and λ.

Remark 3. *As $\lambda \to 0$, we have $r_k \to \infty$ and $p_{g,k}$ coincides with p_g; conversely, as $\lambda \to \infty$, we have $r_k \to 0$ and $p_{g,k}$ coincides with p_k.* ***Therefore, the choice of λ implicitly determines the specialization radius r_k.***

The key advantage of the equivalence between projection and barycenter formulations lies in the greater tractability of the latter. For instance, under the assumption of independent marginal Gaussian distributions, analytical solutions exist for both the reverse Kullback-Leibler (KL) divergence and the Wasserstein-2 distance [11, 1]. These closed-form solutions enable a straightforward and computationally efficient merging process for FMs. To support such covariance-based aggregations and to avoid tuning IVON's effective sample size, which may be unknown at aggregation time, we implement an IVON variant that explicitly maintains the posterior covariance and samples from it directly, rather than relying on a Hessian proxy. More details are provided in Appendix B.

4 Information-Geometric Barycenters: An Interpretable Generalization of FM Merging Techniques

Averaging the weights of fine-tuned models trained under varied hyperparameters has been shown to improve both accuracy and robustness [20]. Motivated by this observation, we extend FM adaptation beyond the specialization framework to multi-model merging through information-geometric barycenters, which minimize the average discrepancy across models. This approach yields a single posterior that balances global generality and domain specificity without relying on ad-hoc parameter heuristics.

Generalization via D-Barycenters. Let $\{p_k\}_{k=1}^N$ denote the variational posteriors of the fine-tuned FMs, for example, those obtained using IVON, and let $p_k(y|x)$ represent the predictive distribution of the k-th FM.

- **Forward KL** With $D = \mathrm{KL}(p_k \,\|\, q)$, the minimizer is the *mixture* in the posterior space: $p_D^\star(\theta) = \sum_{k=1}^N w_k\, p_k(\theta)$. After marginalizing over θ, the resulting predictive distribution also mixes pointwise as $p_D^\star(y \mid x) = \sum_{k=1}^N w_k\, p_k(y \mid x)$, a construction commonly referred to as *Mixture of Experts*.
- **Reverse KL** With $D = \mathrm{KL}(q \,\|\, p_k)$, the solution is the *log-opinion pool* or *Product of Experts*: $p_D^\star(\theta) \propto \prod_{k=1}^N p_k(\theta)^{w_k}$. In exponential families, this corresponds to *natural-parameter averaging*. For Gaussians posteriors, $\Lambda^\star = \sum_{k=1}^N w_k\, \Lambda_k$ and $\mu^\star = (\Lambda^\star)^{-1} \sum_{k=1}^N w_k\, \Lambda_k \mu_k$, where Λ_k denotes the precision matrix of the k-th model; this formulation connects directly to Fisher merging [13].
- **Wasserstein-2.** With $D = \mathrm{W}_2^2(p_k \,\|\, q)$, the minimizer is the *Wasserstein-2 barycenter*. For Gaussians posteriors, the barycenter remains Gaussian with $\Sigma_{\mathrm{W}_2^2} = \left(\sum_{k=1}^N w_k \Sigma_k^{\frac{1}{2}}\right)^2$, $\mu_{\mathrm{W}_2^2} = \sum_{k=1}^N w_k \mu_k$, often yielding more robust summaries than naive parameter averaging.

Practical implications. Barycentric merging provides a single interpretable control parameter (the weights $\{w_k\}$) that balances global and domain-specific knowledge. It recovers popular FM-merging schemes as special cases and admits closed-form solutions for common variational families (e.g., diagonal Gaussians) under widely used divergences. When combined with the IVON training regime, these weights can be derived from curvature estimates, assigning greater importance to higher-curvature models [3, 18]. Consequently, models with higher uncertainty (i.e., lower curvature) are naturally down-weighted in the aggregation.

5 Preliminary Experimental Evaluation

As a preliminary study, we evaluate our approach on non-FM image tasks within a Bayesian FL setup with 10 heterogeneous clients. The analysis focuses on how the Lagrangian multiplier λ governs the trade-off between generalization and specialization (interpreted as personalization in the FL context).

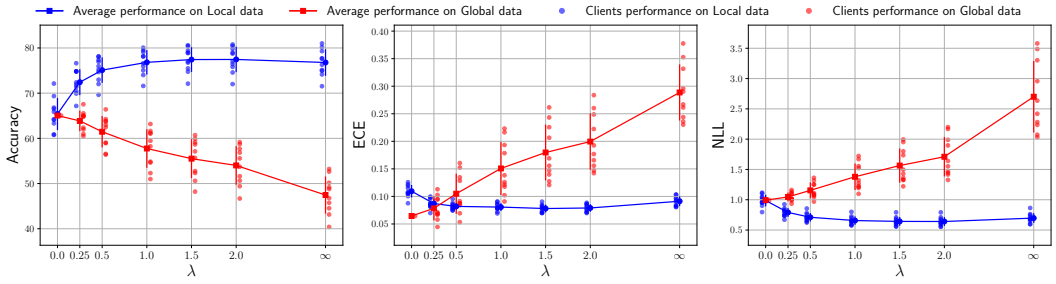

Figure 2: Effect of the Lagrangian multiplier λ on model performance across local and global data distributions. Results are reported for the CIFAR-10 dataset. Note that $\lambda = 0$ corresponds to the global model, while $\lambda \to \infty$ corresponds to the local model.

We analyze the specialization parameter λ in three image datasets with Dirichlet-simulated heterogeneity. By design, λ=0 corresponds to the global model, while $\lambda \to \infty$ yields the fully local one. In CIFAR-10 (Fig. 2), and consistently across datasets, increasing λ improves local performance (Accuracy↑, ECE↓, NLL↓) up to a plateau, while degrading performance on the global distribution (the union of client test sets, approximately class-balanced). Excessively large values of λ lead to over-personalization (worsening global ECE/NLL), whereas small λ underfits client-specific patterns. These results underscore the critical role of λ in controlling the generalization-specialization trade-off, enabling effective adaptation to heterogeneous, non-i.i.d. data distributions in federated settings.

6 Challenges and Future Directions

Like most merging methods in the distribution space (Bayesian) or parameter space (deterministic), our approach assumes architecturally aligned models, i.e., compatible layers and widths, to enable layer-wise aggregation. This constraint is particularly limiting for foundation models, where specialized adapters or domain-specific variants are often smaller than the pretrained backbone. As the next step, our aim is to relax this assumption using Gromov–Wasserstein Optimal Transport maps (e.g., see [4, 12]), which enable mappings between spaces of different dimensionalities. Furthermore, we plan to conduct foundation model-scale experiments to assess the method's efficiency at large scale.

References

[1] Pedro C Álvarez-Esteban, E Del Barrio, JA Cuesta-Albertos, and C Matrán. A Fixed-point Approach to Barycenters in Wasserstein Space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

[2] Rishi Bommasani and et al. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*, 2021.

[3] 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.

[4] Julie Delon, Agnes Desolneux, and Antoine Salmona. Gromov–Wasserstein Distances Between Gaussian Distributions. *Journal of Applied Probability*, 59(4):1178–1198, 2022.

[5] Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's Mergekit: A Toolkit for Merging Large Language Models. *arXiv preprint arXiv:2403.13257*, 2024.

[6] 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.

[7] Nour Jamoussi, Giuseppe Serra, Photios A. Stavrou, and Marios Kountouris. Information-Geometric Barycenters for Bayesian Federated Learning, 2025.

[8] Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. Model Stock: All We Need Is Just a Few Fine-tuned Models. In *European Conference on Computer Vision*, pages 207–223. Springer, 2024.

[9] Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual Pre-training of Language Models. *arXiv preprint arXiv:2302.03241*, 2023.

[10] Mohammad Emtiyaz Khan and Håvard Rue. The Bayesian Learning Rule. *arXiv preprint arXiv:2107.04562*, 2021.

[11] Günther Koliander, Yousef El-Laham, Petar M Djurić, and Franz Hlawatsch. Fusion of Probability Density Functions. *Proceedings of the IEEE*, 110(4):404–453, 2022.

[12] Khang Le, Dung Q Le, Huy Nguyen, Dat Do, Tung Pham, and Nhat Ho. Entropic Gromov-Wasserstein between Gaussian Distributions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12164–12203. PMLR, 17–23 Jul 2022.

[13] Michael S Matena and Colin A Raffel. Merging Models with Fisher-weighted Averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

[14] Michael McCloskey and Neal J. Cohen. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In *Psychology of Learning and Motivation*, volume 24 of *Psy. of Le. and Mot.*, pages 109–165. Academic Press, 1989.

[15] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient Learning of Deep Networks From Decentralized Data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[16] Shivam Pal, Aishwarya Gupta, Saqib Sarwar, and Piyush Rai. Simple and Scalable Federated Learning with Uncertainty via Improved Variational Online Newton. In *OPT 2024: Optimization for Machine Learning*, 2024.

[17] Theodore Papamarkou, Maria Skoularidou, Konstantina Palla, Laurence Aitchison, Julyan Arbel, David Dunson, Maurizio Filippone, Vincent Fortuin, Philipp Hennig, José Miguel Hernández-Lobato, et al. Position: Bayesian Deep Learning Is Needed in the Age of Large-scale ai. *arXiv preprint arXiv:2402.00809*, 2024.

[18] Yuesong Shen, Nico Daheim, Bai Cong, Peter Nickl, Gian Maria Marconi, Clement Bazan, Rio Yokota, Iryna Gurevych, Daniel Cremers, Mohammad Emtiyaz Khan, et al. Variational Learning Is Effective for Large Deep Networks. *arXiv preprint arXiv:2402.17641*, 2024.

[19] Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual Learning of Large Language Models: A Comprehensive Survey. *ACM Computing Surveys*, 2024.

[20] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model Soups: Averaging Weights of Multiple Fine-tuned Models Improves Accuracy Without Increasing Inference Time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022.

A Proof of Theorem 1

Proof. Under Assumption 1, the optimization problem in (1) is convex. Therefore, Lagrangian duality applies to the projection formulation. For a fixed Lagrangian multiplier λ, the minimization problem becomes

$$\min_{p \in \mathcal{M}} \mathcal{L}_\lambda(p), \quad \text{where} \quad \mathcal{L}_\lambda(p) = D(p||p_g) + \lambda D(p||p_k).$$

We can rewrite $\mathcal{L}_\lambda(p)$ as

$$\mathcal{L}_\lambda(p) = (\lambda + 1) \left(\frac{1}{\lambda + 1} D(p||p_g) + \frac{\lambda}{\lambda + 1} D(p||p_k) \right).$$

Let w_g and w_k be defined as in (1). Then the minimization problem simplifies to

$$\min_{p \in \mathcal{M}} w_g D(p||p_g) + w_k D(p||p_k). \tag{4}$$

By Definition 2, the probability distribution p minimizing the weighted sum in (4) coincides with the D-Barycenter. This concludes the proof. □

B Covariance-Based Aggregation of IVON-Trained Models

To support aggregation strategies that operate directly on covariance matrices, and to avoid manually specifying the effective sample size parameter in IVON, which ideally corresponds to the full dataset size but is often unknown at merge time, we consider a subclass of IVON that explicitly stores the covariance matrix and samples directly from it, rather than relying solely on the Hessian approximation. This follows from the relation [18]

$$\sigma^2 = \frac{1}{N(h + \delta)}, \tag{5}$$

which expresses the variance σ^2 as a function of the dataset size N, the Hessian approximation h, and the weight decay term δ. This enables computing the covariance matrix directly from the Hessian. In practice, each model first estimates its Hessian locally and then converts it to a covariance matrix, which is subsequently aggregated across models. Unlike prior work [16] that applied IVON but was restricted to Reverse KL Barycenter (RKLB) aggregation due to its Hessian-based formulation, our approach supports a broader class of covariance-based aggregation methods, including Wasserstein barycenters (WB).