A Theoretical Framework for Federated Domain Generalization with Gradient Alignment

Mahdiyar Molahasani* **Milad Soltany*** **Farhad Pourpanah***
Michael Greenspan **Ali Etemad**
Queen's University, Canada

Abstract

Gradient alignment has shown empirical success in federated domain generalization, yet a theoretical foundation for this approach remains unexplored. To address this gap, we provide a theoretical framework linking domain shift and gradient alignment in this paper. We begin by modeling the similarity between domains through the mutual information of their data. We then show that as the domain shift between clients in a federated system increases, the covariance between their respective gradients decreases. This link is initially established for federated supervised learning and subsequently extended to federated unsupervised learning, showing the consistency of our findings even in a self-supervised setup. Our work can further aid the development of robust models by providing an understanding of how gradient alignment affects learning dynamics and domain generalization.

1 Introduction

Federated learning [1, 2] has emerged as a promising framework for training machine learning models across multiple decentralized clients while preserving data privacy. It allows clients to collaboratively train a global model without the need to exchange their sensitive and local data. Each client trains a local model using its data and a server aggregates these models at a certain frequency [3, 4, 5]. However, the inherent heterogeneity of data across different clients poses significant challenges, as variations in data distributions can cause models to perform well on local training data but fail to generalize to unseen target domains. Federated Domain Generalization (FDG) aims to address this challenge by developing models that can generalize effectively to new, unseen data distributions not represented during training [6, 7].

In recent years, various approaches have been proposed for FDG, such as learning domain-invariant representations [8, 9] and identifying common features across multiple domains [10]. One promising approach for addressing FDG is **gradient alignment** [11, 12, 13, 14, 15], which aims to align the gradients of different clients during the training process to enhance the generalizability of the aggregated model. Gradient alignment was first utilized in multi-task learning [16], where a technique known as gradient surgery was proposed. This method projects a task's gradients onto the orthogonal plane of any other task's gradient that exhibits a conflicting direction. Later, this technique was used to improve generalization in centralized (non-federated) learning through gradient alignment [17, 18, 19]. The study in [17] updates the weights if the signs of the gradient components are aligned across all domains. Fish [18] maximizes the inner product between gradients from different domains, while Fishr [19] leverages domain-level gradient variances.

Although the effectiveness of gradient alignment in federated domain generalization has been demonstrated in practice [11, 12, 13, 14, 15], no theoretical basis has been proposed for how observing the local gradients of different clients can infer information about the local data distributions without

[*] Equal contribution

Mathematics of Modern Machine Learning Workshop at NeurIPS 2024.

violating privacy constraints. To this end, in this paper, we propose a theoretical framework to establish a link between domain shift and gradient alignment for the first time. We first model the similarity between the domains using the mutual information between their data. Then, we demonstrate that when the domain shift between two clients in a federated system increases, the covariance between their respective gradients decreases. We prove this concept for a federated supervised domain generalization setup. We then extend our framework to federated unsupervised domain generalization and show that our findings hold even when the clients are trained using self-supervised learning loss. Furthermore, since using cosine similarity as a criterion for discarding unaligned gradients has been empirically shown to effectively enhance gradient alignment [20, 21], we theoretically analyze its applicability within a federated learning setup. The contributions of this work can be summarized as follows:

- For the first time, we provide a theoretical basis for using gradient alignment for federated domain generalization.
- We demonstrate the applicability of this method not only for supervised setup but also for self-supervised representation learning.

By providing a rigorous theoretical basis for using gradient alignment for domain generalization without violating privacy constraints, this work is a step toward more robust federated learning systems.

2 Approach

2.1 Problem formulation

To formalize federated domain generalization, assume K clients, C_i, in a federated setup, each with its own *unlabeled* data $\mathcal{D}_i = \{\mathbf{x}_i^{(n)}\}_{n=1}^{N_i}$. Each dataset consists of N_i data points sampled from a distinct data distribution $p(\mathbf{x}_i)$, where $\mathbf{x}_i$ is a vector of F features, i.e., $\mathbf{x}_i = [x_i^1, x_i^2, ...x_i^F]^T$. The data distributions are assumed to be different among the clients with each distribution $p(\mathbf{x}_i)$ sampled from a family of distributions $\mathcal{P}$. Privacy constraints prevent the transfer of data between clients or to the server S. The objective is to learn generalized representations from $\mathbf{x}_i$ that perform well across unseen distributions $p(\mathbf{x}_t) \sim \mathcal{P}$, where $p(\mathbf{x}_i) \neq p(\mathbf{x}_t)$. This is formulated as minimizing the expected loss over the unseen distributions:

$$\min_{\theta} \mathbb{E}_{p(\mathbf{x}_t)\sim\mathcal{P}} \left[\mathbb{E}_{p(\mathbf{x}_t)} \left[\mathcal{L}(\theta; \mathbf{x})\right]\right], \tag{1}$$

where $\mathcal{L}$ is the unsupervised loss function, and θ is the set of global model parameters. Each client contributes to this goal by computing a local objective function approximating the expected loss with respect to its own data distribution:

$$\min_{\theta_i} \mathbb{E}_{p(\mathbf{x}_i)} \left[l_i(\theta_i; \mathbf{x})\right] \approx \frac{1}{N_i} \sum_{n=1}^{N_i} l(\theta_i; \mathbf{x}_i^{(n)}), \tag{2}$$

where θ_i indicates the local parameters of client C_i, and θ is the global aggregation of all local θ_i.

2.2 Gradient Alignment and Domain Shift

Under a federated learning framework, privacy constraints prevent clients and servers from accessing each other's data, including distribution information such as data means and variances. They can, however, observe individual client gradients at the server level and the average aggregated gradient across clients at the client level. We motivate our work on the fact that alignment of gradients may infer characteristics of the client domain distributions, thus facilitating improved model generalization. While empirically the utility of gradients has been demonstrated in the area of domain generalization [17], no theoretical basis has been proposed for this approach under federated constraints. Accordingly, we aim to establish a link between gradient alignment and domain shifts. Our theoretical findings provide the basis for effective local parameter updates and global model aggregation to address federated domain generalization. To this end, we analyze this link in two different setups of **Supervised** and **Unsupervised** learning.

2.2.1 Federated Supervised Domain Generalization

Assumption 1. *Let each $\mathbf{x}_i^f$ be a random variable drawn from a Normal distribution $p(\mathbf{x}_i^f) \sim \mathcal{P}$ [22]. Within a single domain, following [23, 24], features are assumed to be independent ($Cov(x_i^{f_1}, x_i^{f_2}) = 0$). Across different domains, corresponding features of $\mathbf{x}_i^f$ and $\mathbf{x}_j^f$ are bivariate with a covariance of $\sigma_{x_i^f, x_j^f}$. In line with contemporary practices in deep learning, the features are normalized with $\mu = 0$ and $\sigma^2 = 1$ [25, 26, 27]. We also assume that each client is a logistic regression classifier trained in a supervised federated setup using cross-entropy loss. After each epoch, the local models are aggregated in the server and sent back to all clients. The gradient $\mathbf{g}_i$ of the model is assumed to be differentiable.*

Theorem 1 (Gradient Misalignment in Federated Supervised Learning Dependent upon Domain Shift)**.** *Given Assumption 1, under the problem of Federated **Supervised** Domain Generalization, for two distinct domains characterized by random variables $\mathbf{x}_i$ and $\mathbf{x}_j$ belonging to two different clients C_i and C_j, an increase in domain shift across the clients results in a decrease in the covariance Cov$(\mathbf{g}_i, \mathbf{g}_j)$ of the corresponding gradients $\mathbf{g}_i$, $\mathbf{g}_j$ across C_i and C_j's respective local models.*

Proof Sketch. The proof proceeds by first modeling the similarity between two domains C_i and C_j using mutual information denoted $I(\mathbf{x}_i; \mathbf{x}_j)$. We then introduce Lemma A1, demonstrating that the mutual information can be expressed as a function of the covariance $\sigma_{x_i^f, x_j^f}$ between the features of the two domains (Equation 3). As domain shift increases, $\sigma_{x_i^f, x_j^f}$ decreases, leading to a reduction in mutual information. Next, Lemma A2 is introduced establishing that the covariance between each dimension of gradients $\mathrm{Cov}(\mathbf{g}_i, \mathbf{g}_j)$ depends on the corresponding feature covariance $\sigma x_i^f, x_j^f$ through Equation 4. By introducing Claim A1, we show that the covariance of gradients is positively and monotonically related to the covariance of the features under the federated supervised setup where local models are logistic regression classifiers. Thus, as the feature covariance decreases due to domain shift, the gradient covariance also decreases. This establishes the positive correlation between mutual information and gradient covariance. The full proof of this theorem is presented in Appendix A.1. □

2.2.2 Federated Unsupervised Domain Generalization

We extend our theoretical analysis to federated unsupervised domain generalization setup under the following assumptions, which differ slightly from those previously presented in Assumption 1:

Assumption 2. *Based on Assumption 1, we extend our framework to an unsupervised learning context. The assumptions regarding the differentiability of the gradients and data distribution remain the same except the clients now do not have access to the labels. Regarding the model, we assume each client is a one-layer encoder with the sigmoid activation function. Following [28], local training is performed using contrastive loss where a random augmentation of a data sample is used as the positive pair and all other samples are utilized as the negative pairs. The random augmentation is performed with the same random Affine transformation ($Ax_i + B$) [29] broadcast to all clients.*

Theorem 2 (Gradient Misalignment in Federated Unsupervised Learning Dependent upon Domain Shift)**.** *The insights from Theorem 1 also apply to the federated **unsupervised** learning framework described in Assumption 2, demonstrating that an increase in domain shift leads to a monotonic decrease in the covariance of the gradients of local models within the context of federated unsupervised domain generalization.*

Proof Sketch. The proof of Theorem 2 follows a similar structure to that of Theorem 1, with the key difference being the transition to a federated unsupervised domain generalization setup using self-supervised learning. The assumptions regarding the data distribution remain the same, so the relationships between mutual information and feature covariance established in Theorem 1 continue to hold. However, Claim A2 is introduced to extend the result to the self-supervised learning setting, showing that for both positive and negative contrastive pairs, the gradient covariance remains positively correlated with feature covariance. This leads to the same conclusion as in Theorem 1: as domain shift increases, feature and gradient covariances decrease, establishing a positive correlation between mutual information and gradient covariance. The full proof is provided in Appendix A.5. □

For a more detailed analysis of the relationship between the domain distributions and their corresponding gradients, refer to Corollary A1, which is presented and proved in Appendix A.7.

2.3 Cosine Similarity for Gradient Alignment

More recently, several works employed the cosine similarity of the gradients as a proxy for the generalizability of the models through different domains or tasks [30, 31, 32, 20, 33, 21, 34]. To this end, we demonstrate that using cosine similarity as the criterion for discarding unaligned gradients can indeed promote generalization through the following proposition.

Proposition 1. *Given two sets of gradient vectors* $\mathbf{g}_i$ *and* $\mathbf{g}_j$*, by removing the* K^{th} *vector in* $\mathbf{g}_j$ *where* $cos(\mathbf{g}_{j,K}, \mathbf{g}_{est}) < 0$*, the covariance of two sets increases.*

Proof Sketch. The proposition is proved by first partitioning the gradient set $\mathbf{g}_j$ into two parts: the unaligned vector $\mathbf{g}_{j,K}$ and the remaining aligned vectors. We then express the covariance between the sets $\mathbf{g}_i$ and $\mathbf{g}_j$ as a sum of two terms: the covariance between $\mathbf{g}_i$ and the aligned set, and the contribution of the unaligned vector $\mathbf{g}_{j,K}$. By analyzing the second term, we show that the contribution of $\mathbf{g}_{j,K}$ introduces mostly negative values due to its opposite direction relative to the mean vector of the two gradient sets, as captured by the cosine similarity. Therefore, removing $\mathbf{g}_{j,K}$, which is negatively aligned, increases the overall covariance between the sets. For more details on this Proposition and its proof, see Appendix A.9. □

3 Discussions

Here we empirically verify our theoretical finding by training different models with the same architecture and initialization on different domains of the PACS dataset [35]. We then measure the covariance between the gradients of each pair of these models and plot them against the amount of domain shift between their training domains, as illustrated in Figure 1. Each point in the plot is denoted by two letters representing the corresponding pair of domains. We observe that, except for a single outlier, the trend follows our prediction based on the insight introduced in Theorems 1 and 2. For more details on this experiment see Appendix B.

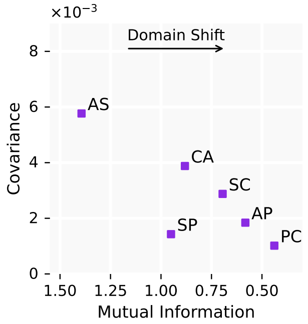

Figure 1: Empirical verification of the proposed framework.

Our framework introduces mutual information as an effective metric for quantifying domain similarities, setting a foundation for future generalization techniques that can leverage this measure for design and performance evaluation. Furthermore, establishing a link between gradient alignment and domain shift suggests a rigorous way of retrieving domain shift information without violating privacy constraints which can be incorporated in designing generlizable pipelines. Given the inherent heterogeneity of decentralized data, our framework is broadly applicable to a wide range of federated learning problems, encompassing both supervised and unsupervised training paradigms.

4 Conclusion

In this work, for the first time, we establish a link between gradient alignment and domain shift in a federated setup, filling a gap in the theoretical basis for recent federated domain generalization methods that rely on gradient alignment. In both supervised and unsupervised setups, we demonstrate that an increase in domain shift across the client's local data leads to a decrease in their respective gradient alignment. These findings provide insights into the learning dynamics of federated systems, which can be leveraged to develop models that facilitate better generalization across diverse and decentralized datasets without violating the privacy constraints.

Acknowledgments

We would like to thank Geotab Inc., the City of Kingston, and NSERC for their support of this work.

References

[1] 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*, 2017, pp. 1273–1282.

[2] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, pp. 106775, 2021.

[3] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran, "An efficient framework for clustered federated learning," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 19586–19597.

[4] Zachary Charles, Zachary Garrett, Zhouyuan Huo, Sergei Shmulyian, and Virginia Smith, "On large-cohort training for federated learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 20461–20475, 2021.

[5] Qinbin Li, Bingsheng He, and Dawn Song, "Model-contrastive federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10713–10722.

[6] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng, "FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1013–1023.

[7] Ying Li, Xingwei Wang, Rongfei Zeng, Praveen Kumar Donta, Ilir Murturi, Min Huang, and Schahram Dustdar, "Federated domain generalization: A survey," *arXiv:2306.01334*, 2023.

[8] A Tuan Nguyen, Philip Torr, and Ser Nam Lim, "FedSR: A simple and effective domain generalization method for federated learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38831–38843, 2022.

[9] Guile Wu and Shaogang Gong, "Collaborative optimization and aggregation for decentralized domain generalization and adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6484–6493.

[10] Liling Zhang, Xinyu Lei, Yichun Shi, Hongyu Huang, and Chao Chen, "Federated learning with domain generalization," *arXiv:2111.10487*, 2021.

[11] Yatin Dandi, Luis Barba, and Martin Jaggi, "Implicit gradient alignment in distributed and federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 6454–6462.

[12] Yikang Wei and Yahong Han, "Multi-source collaborative gradient discrepancy minimization for federated domain generalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 15805–15813.

[13] Meilu Zhu, Zhen Chen, and Yixuan Yuan, "Feddm: Federated weakly supervised segmentation via annotation calibration and gradient de-conflicting," *IEEE Transactions on Medical Imaging*, vol. 42, pp. 1632–1643, 2023.

[14] Chenglu Pan, Jiarong Xu, Yue Yu, Ziqi Yang, Qingbiao Wu, Chunping Wang, Lei Chen, and Yang Yang, "Towards fair graph federated learning via incentive mechanisms," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 14499–14507.

[15] Chris Xing Tian, Haoliang Li, Yufei Wang, and Shiqi Wang, "Privacy-preserving constrained domain generalization via gradient alignment," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[16] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn, "Gradient surgery for multi-task learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5824–5836, 2020.

[17] Lucas Mansilla, Rodrigo Echeveste, Diego H Milone, and Enzo Ferrante, "Domain generalization via gradient surgery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6630–6638.

[18] Yuge Shi, Jeffrey Seely, Philip Torr, Siddharth N, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve, "Gradient matching for domain generalization," in *International Conference on Learning Representations*, 2022.

[19] Alexandre Rame, Corentin Dancette, and Matthieu Cord, "Fishr: Invariant gradient variances for out-of-distribution generalization," in *International Conference on Learning Representations*, 2022.

[20] Joseph Oliver Pemberton and Rui Ponte Costa, "Bp (λ): Online learning via synthetic gradients," *Transactions on Machine Learning Research*, 2024.

[21] Qizhang Li, Yiwen Guo, Wangmeng Zuo, and Hao Chen, "Improving adversarial transferability via intermediate-level perturbation decay," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[22] Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall, "Learning distance functions using equivalence relations," in *Proceedings of the International Conference on Machine Learning*, 2003, pp. 11–18.

[23] Erik Štrumbelj and Igor Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and information systems*, vol. 41, pp. 647–665, 2014.

[24] Erik Štrumbelj and Igor Kononenko, "A general method for visualizing and explaining black-box regression models," in *International Conference on Adaptive and Natural Computing Algorithms*. Springer, 2011, pp. 21–30.

[25] Jiahui Yu and Konstantinos Spiliopoulos, "Normalization effects on deep neural networks," *arXiv preprint arXiv:2209.01018*, 2022.

[26] Haobo Qi, Jing Zhou, and Hansheng Wang, "A note on factor normalization for deep neural network models," *Scientific Reports*, vol. 12, no. 1, pp. 5909, 2022.

[27] Lei Huang, Jie Qin, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao, "Normalization techniques in training dnns: Methodology, analysis and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[28] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, 2020, pp. 1597–1607.

[29] Zihao Wang, Chunxu Wu, Yifei Yang, and Zhen Li, "Learning transformation-predictive representations for detection and description of local features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11464–11473.

[30] Stanislav Fort, Paweł Krzysztof Nowak, Stanislaw Jastrzebski, and Srini Narayanan, "Stiffness: A new perspective on generalization in neural networks," *arXiv preprint arXiv:1901.09491*, 2019.

[31] Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras, "The break-even point on optimization trajectories of deep neural networks," in *International Conference on Learning Representations*, 2020.

[32] Stanislav Fort and Surya Ganguli, "Emergent properties of the local geometry of neural loss landscapes," *arXiv preprint arXiv:1910.05929*, 2019.

[33] Lucas Page-Caccia, Edoardo Maria Ponti, Zhan Su, Matheus Pereira, Nicolas Le Roux, and Alessandro Sordoni, "Multi-head adapter routing for cross-task generalization," *Advances in Neural Information Processing Systems*, vol. 36, pp. 56916–56931, 2023.

[34] Gunshi Gupta, Karmesh Yadav, and Liam Paull, "Look-ahead meta learning for continual learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11588–11598, 2020.

[35] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5542–5550.

[36] Jian Gao, Yang Hua, Guosheng Hu, Chi Wang, and Neil M Robertson, "Reducing distributional uncertainty by mutual information maximisation and transferable feature learning," in *European Conference on Computer Vision*, 2020, pp. 587–605.

[37] Willi Menapace, Stéphane Lathuilière, and Elisa Ricci, "Learning to cluster under domain shift," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 736–752.

[38] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh, "Learning to diversify for single domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 834–843.

[39] Ziv Goldfeld and Kristjan Greenewald, "Sliced mutual information: A scalable measure of statistical dependence," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17567–17578, 2021.

[40] Xingxuan Zhang, Linjun Zhou, Renzhe Xu, Peng Cui, Zheyan Shen, and Haoxin Liu, "Towards unsupervised domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4910–4920.

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[42] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

A Proofs

A.1 Proof of Theorem 1

Proof. Motivated by previous works where the similarity between the representations of different domains under domain shift is measured by Mutual Information [36, 37, 38], we use this concept for modeling the similarity between different domains drawn from a family of distributions. To calculate the mutual information we introduce the following Lemma.

Lemma A1. *Given Assumption 1, the mutual information between two random variables* $\mathbf{x}_i$ *and* $\mathbf{x}_j$ *can be calculated as:*

$$I(\mathbf{x}_i; \mathbf{x}_j) = -\frac{1}{2} \sum_{f=1}^{F} \log(1 - \sigma^2_{x_i^f, x_j^f}). \tag{3}$$

The full proof of this lemma is presented in Appendix A.2. Since $\sigma_{x_i^f, x_i^f} = \sigma^2_{x_i^f}$, given identical and standardized domains we have $\sigma_{x_i^f, x_j^f} = 1$. Therefore, we observe from Eq. (3) that as the shift between the domain approaches zero, the Mutual Information approaches infinity. On the other hand, as the two domains shift apart, $\sigma_{x_i^f, x_j^f}$ approaches zero, and consequently the Mutual information monotonically decreases toward zero. Accordingly, $I(\mathbf{x}_i; \mathbf{x}_j)$ and $\sigma_{x_i^f, x_i^f}$ are positively and monotonically correlated. To establish the link between the covariance of the features and the variance of the gradient and demonstrate their relationship, the following lemma and claim are introduced.

Lemma A2. *Given Assumption 1, the covariance between the differentiable function* $\mathbf{g}$ *with inputs* $\mathbf{x}_i$ *and* $\boldsymbol{x}_j$ *can be estimated as:*

$$\mathrm{Cov}(\mathbf{g}_i, \mathbf{g}_j)_{mn} \approx \sum_{f=1}^{F} \sigma_{x_i^f, x_j^f} \left(\frac{\partial g_{i(m)}}{\partial x_i^f} \Big|_{\mu_i^f} \right) \left(\frac{\partial g_{j(n)}}{\partial x_j^f} \Big|_{\mu_j^f} \right), \tag{4}$$

where $\sigma_{x_i^f, x_j^f}$ *is the covariance between the* f^{th} *feature of* x_i *and* x_j *and* $g_{i(m)}$ *is the* m^{th} *dimension of* g_i.

Claim A1. *For all clients with a logistic regression classifier described in Assumption 2, for any* i *and* j*, the sign of* $(\frac{\partial \mathbf{g}_i}{\partial x_i^f}\Big|_{x_i^f = \mu_i^f})(\frac{\partial \mathbf{g}_j}{\partial x_j^f}\Big|_{x_j^f = \mu_j^f})$ *is always positive.*

The proof for this Lemma and Claim are provided in Appendix A.3 and A.4, respectively. From Eq. (4) and Claim A1 it can be concluded $\mathrm{Cov}(\mathbf{g}_i, \mathbf{g}_j)$ and $I(\mathbf{x}_i; \mathbf{x}_j)$ are positively and monotonically correlated, which completes the proof. □

A.2 Proof of Lemma A1

Proof. By definition, the Mutual Information between two random variables, $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$I(\mathbf{x}_i; \mathbf{x}_j) = \int \int p(\mathbf{x}_i, \mathbf{x}_j) \log \left(\frac{p(\mathbf{x}_i, \mathbf{x}_j)}{p(\mathbf{x}_i) p(\mathbf{x}_j)} \right) d\mathbf{x}_i d\mathbf{x}_j. \tag{5}$$

Since the domains from which the corresponding random variables are drawn are sets of independent features:

$$p(\mathbf{x}_i) = \prod_{f=1}^{F} p(x_i^f), \qquad p(\mathbf{x}_j) = \prod_{f=1}^{F} p(x_j^f). \tag{6}$$

Moreover, since each pair of corresponding features (x_i^f, x_j^f) across domains forms a bivariate Normal distribution, we can derive:

$$p(\mathbf{x}_i, \mathbf{x}_j) = \prod_{f=1}^{F} p(x_i^f, x_j^f). \tag{7}$$

By substituting Eqs. (6) and (7) into Eq. (5), using the multiplicative property of logarithms and the definition of Mutual Information, we obtain:

$$I(\mathbf{x}_i; \mathbf{x}_j) = \sum_{f=1}^{F} \int \int p(x_i^f, x_j^f) \log \left(\frac{p(x_i^f, x_j^f)}{p(x_i^f)p(x_j^f)} \right) dx_i^f \, dx_j^f = \sum_{f=1}^{F} I(x_i^f; x_j^f). \tag{8}$$

For bivariate Normal distributions x_i and x_j, Mutual Information can be measured by $I(x_i; x_j) = -\frac{1}{2}\log(1-\rho^2)$, where $\rho = \frac{\sigma_{x_i,x_j}}{\sigma_{x_i}\sigma_{x_j}}$ is the correlation coefficient [39]. Accordingly, Eq. (8) yields:

$$I(\mathbf{x}_i; \mathbf{x}_j) = -\frac{1}{2} \sum_{f=1}^{F} \log(1 - (\frac{\sigma_{x_i^f, x_j^f}}{\sigma_{x_i^f}\sigma_{x_j^f}})^2). \tag{9}$$

According to Assumption 1, we have $\sigma_{x_i^f} = \sigma_{x_j^f} = 1$. Hence, Eq. (9) yields:

$$I(\mathbf{x}_i; \mathbf{x}_j) = -\frac{1}{2} \sum_{f=1}^{F} \log(1 - \sigma^2_{x_i^f, x_j^f}), \tag{10}$$

which completes the proof. □

A.3 Proof of Lemma A2

Proof. The covariance of $\mathbf{g}_i$ and $\mathbf{g}_j$, by definition, is:

$$\text{Cov}(\mathbf{g}_i, \mathbf{g}_j) = \mathbb{E}[(\mathbf{g}_i - \mathbb{E}[\mathbf{g}_i])(\mathbf{g}_j - \mathbb{E}[\mathbf{g}_j])^T]. \tag{11}$$

First, we calculate $\mathbb{E}[\mathbf{g}_i]$. Given $\mathbf{g}$ is differentiable, as per the Assumption 1, we can estimate it using the first-degree Taylor expansion theorem around μ:

$$g(x) \approx g(\mu) + J_g(\mu).(x - \mu), \tag{12}$$

where $J_g(.)$ is the Jacobian matrix g. We can use this equation to estimate $\mathbb{E}[\mathbf{g}]$ as follows:

$$\mathbb{E}[\mathbf{g}] = \mathbb{E}\left[g(\mu) + J_g(\mu).(x - \mu))\right]. \tag{13}$$

Note that $g(\mu)$ is constant, therefore $\mathbb{E}[g(\mu)] = g(\mu)$. Moreover, since we assume that the distribution of the features is normal, we can conclude:

$$\mathbb{E}\left[J_g(\mu).(x - \mu))\right] = \mathbf{0}_F. \tag{14}$$

Since $\mathbb{E}[x - \mu] = \mathbf{0}_F$ for $x \sim \mathcal{N}$, $\mathbb{E}[g] = g(\mu)$. By replacing $\mathbf{g}_i$ and $\mathbf{g}_j$ with their Taylor expansion we derive:

$$\text{Cov}(\mathbf{g}_i, \mathbf{g}_j) = \mathbb{E}\left[(J_{g_i}(\mu_i).(x - \mu_i)) \ . \left(J_{g_j}(\mu_j).(x - \mu_j))\right)^T\right]. \tag{15}$$

Hence:

$$\text{Cov}(\mathbf{g}_i, \mathbf{g}_j) = J_{g_i}(\mu_i).\mathbb{E}\left[(x - \mu_i).(x - \mu_j)^T\right].J_{g_j}(\mu_j)^T. \tag{16}$$

Given the covariance matrix $\Sigma = \mathbb{E}[(x - \mu_i)(x - \mu_j)]$, we derive:

$$\text{Cov}(\mathbf{g}_i, \mathbf{g}_j) = J_{g_i}(\mu_i).\Sigma.J_{g_j}(\mu_j)^T. \tag{17}$$

As the features are assumed to be independent, Σ is a diagonal matrix, the general term of each entry of $\text{Cov}(\mathbf{g}_i, \mathbf{g}_j)$ can be derived as:

$$\text{Cov}(\mathbf{g}_i, \mathbf{g}_j)_{mn} = \sum_{f=1}^{F} \sigma_{x_i^f, x_j^f} \left(\frac{\partial g_{i(m)}}{\partial x_i^f}\bigg|_{\mu_i^f} \right) \left(\frac{\partial g_{j(n)}}{\partial x_j^f}\bigg|_{\mu_j^f} \right), \tag{18}$$

where $g_{i(m)}$ and $g_{j(n)}$ represented the m^{th} and the n^{th} dimension of g_i and g_j, respectively. □

A.4 Proof of Claim A1

Proof. Following the assumptions, let's describe the model's loss by:

$$l(y,\hat{y}) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}), \tag{19}$$

where y represents the label and $\hat{y}$ signifies the model output defined as $\sigma(Wx+b) = \sigma(x) = 1/(1+e^{-Wx+b})$. The gradient of the loss with respect to W is therefore given by:

$$g = \frac{\partial L}{\partial w} = (\hat{y}-y)x, \tag{20}$$

Subsequently, the derivative of g with respect to x at $x=\mu$ is expressed as:

$$\left.\frac{\partial g}{\partial x}\right|_{x=\mu} = \sigma(W\mu+b) + \mu\left.\frac{\partial\sigma(Wx+b)}{\partial x}\right|_{x=\mu} - y. \tag{21}$$

Recalling the proof of theorem 1 and noting that data is standardized, we find $\mu_i^f = \mu_j^f = 0$. This leads to

$$\frac{\partial g}{\partial x}\Big|_{x=\mu} = \sigma(b) - y. \tag{22}$$

Given that $\sigma(x) = 1/(1+e^{-x})$, its outcome always falls within the range $(0,1)$. Considering y as a data label that can be either 1 or 0, it follows that when $y=1$, $\sigma(b)-y$ is invariably negative, and when $y=0$, $\sigma(b)-y$ is invariably positive. Consequently, the sign of $\frac{\partial \mathbf{g}}{\partial x}\big|_{x=\mu}$ is determined solely by the value of y, ensuring $\text{sign}(\sigma(b_i)-y) = \text{sign}(\sigma(b_j)-y)$. Thus,

$$\text{sign}\left(\frac{\partial g_i}{\partial x_i^f}\Big|_{x_i^f=\mu_i^f}\right) = \text{sign}\left(\frac{\partial g_j}{\partial x_j^f}\Big|_{x_j^f=\mu_j^f}\right). \tag{23}$$

□

A.5 Proof of Theorem 2

Proof of Theorem 2. Since all the assumptions regarding the data distribution are the same, Eq. (5) to Eq. (4) still hold. However, since the training paradigm has changed, we introduce the following claim, which extends Claim A1 under the conditions of Assumption 2.

Claim A2. *For all clients trained under the federated unsupervised domain generalization setup using self-supervised learning described in Assumption 1, for two clients C_i and C_j, for both positive and negative contrastive data pairs, the sign of* $\left(\left.\frac{\partial \mathbf{g}_i}{\partial x_i^f}\right|_{x_i^f=\mu_i^f}\right)\left(\left.\frac{\partial \mathbf{g}_j}{\partial x_j^f}\right|_{x_j^f=\mu_j^f}\right)$ *is always positive in all dimensions. See proof in Appendix A.6.*

From Eq. (4) and Claim A2 it can be concluded $\text{Cov}(\mathbf{g}_i, \mathbf{g}_j)$ and $I(\mathbf{x}_i;\mathbf{x}_j)$ are positively and monotonically correlated, which completes the proof. □

A.6 Proof of Claim A2

Proof. Following Assumption 1, the model is defined as $\hat{y} = \sigma(Wx_1 - Wx_2)$ and the loss is formulated as:

$$l(y,\hat{y}) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}), \tag{24}$$

where $y=1$ for positive pairs and $y=0$ for the negative pairs. Consequently, the gradient of the loss w.r.t. W is

$$g = \begin{cases} (\sigma(Wx_1 - Wx_2) - 1)(x_1 - x_2) & \text{if } y = 1 \text{ (Positive pairs)} \\ \sigma(Wx_1 - Wx_2)(x_1 - x_2) & \text{if } y = 0 \text{ (Negative pairs)} \end{cases} \tag{25}$$

We first focus on the positive pairs. In the contrastive loss, an augmentation of the same sample is usually used as the positive sample. Following Assumption 1, the augmented sample is formulated as $x_2 = Ax_1 + B$. Hence, the gradient of the positive pair is derived as:

$$g^+ = (\sigma(Wx_1 - WAx_1 - WB) - 1)(x_1 - Ax_1 - B). \tag{26}$$

Subsequently, the derivative of g^+ with respect to x_1 at $x_1 = \mu_1$ is expressed as:

$$\begin{aligned}\left.\frac{\partial g^+}{\partial x_1}\right|_{x_1=\mu_1} =& \sigma(W\mu_1 - WA\mu - WB)(1 - \sigma(W\mu_1 - WA\mu_1 - WB)) \\ & (\mu_1 - A\mu - B) + \sigma(W\mu_1 - WA\mu_1 - WB)(1 - A). \end{aligned} \tag{27}$$

Recall Assumption 1 stating that the data is normalized, hence, $\mu_1 = 0$. Therefore:

$$\left.\frac{\partial g^+}{\partial x_1}\right|_{x_1=\mu_1} = \sigma(-WB)(1 - \sigma(-WB))(-B) + \sigma(-WB)(1 - A). \tag{28}$$

As a result, we can conclude that for all positive pairs, $\left.\frac{\partial g^+}{\partial x_i}\right|_{x_i=\mu_i}$ only depends on the augmentation and the network weights. According to Assumption 1, the weights of clients are the same and equal to the global model at each communication round. Moreover, since the same random augmentations are broadcasted to all the clients, A and B will be the same at each step across all the local models. Hence, for all x_i, $\text{sign}(\frac{\partial g_i}{\partial x_i^f}|_{x_i^f=\mu_i^f}) = \text{sign}(\frac{\partial g_j}{\partial x_j^f}|_{x_j^f=\mu_j^f})$.

Considering the negative pairs, from Eq. (25), we have:

$$g^- = \sigma(Wx_1 - Wx_2)(x_1 - x_2). \tag{29}$$

The derivative of g^- w.r.t. x_1 at $x_1 = \mu_1$ is derived as:

$$\left.\frac{\partial g^-}{\partial x_1}\right|_{x_1=\mu_1} = \sigma(W\mu_1 - Wx_2)(1 - \sigma(W\mu_1 - Wx_2))(\mu_1 - Ax_2) \tag{30}$$

$$+ \sigma(W\mu_1 - Wx_2)(1 - x_2). \tag{31}$$

Based on the stated assumption $\mu_1 = 0$, therefore:

$$\left.\frac{\partial g^-}{\partial x_1}\right|_{x_1=\mu_1} = \sigma(-Wx_2)(1 - \sigma(-Wx_2))(-Ax_2) + \sigma(-Wx_2)(1 - x_2). \tag{32}$$

The σ function is estimated using Taylor expansion around zero (since the data is normalized to $\mu = 0$), as:

$$\sigma(x) \simeq \frac{1}{2} + \sum_{n=0}^{\infty} \frac{(-1)^n(2^{2n+1} - 1)B_{2n+2}}{(2n+2)!} x^{2n+1}, \tag{33}$$

where B_n are Bernoulli numbers. By replacing Eq. (33) in Eq. (32):

$$\begin{aligned}\left.\frac{\partial g^-}{\partial x_1}\right|_{x_1=\mu_1} \simeq & \\ & \left(\frac{1}{2} + \sum_{n=0}^{\infty} \frac{(-1)^{n+1}(2^{2n+1} - 1)B_{2n+2}}{(2n+2)!}(Wx_2)^{2n+1}\right) \quad \times \\ & \left(\frac{1}{2} - \sum_{n=0}^{\infty} \frac{(-1)^{n+1}(2^{2n+1} - 1)B_{2n+2}}{(2n+2)!}(Wx_2)^{2n+1}\right)(-Ax_2) \\ & + \left(\frac{1}{2} + \sum_{n=0}^{\infty} \frac{(-1)^{n+1}(2^{2n+1} - 1)B_{2n+2}}{(2n+2)!}(Wx_2)^{2n+1}\right)(1 - x_2). \end{aligned} \tag{34}$$

Since the coeffiencts are only consist of n, B_n and W, we can simplify as:

$$\left.\frac{\partial g^-}{\partial x_1}\right|_{x_1=\mu_1} \approx \sum_{n=0}^{\infty} \zeta(W, n) x_2^n \tag{35}$$

Since the gradient is a linear operation, the gradient of the sum equals the sum of gradients. According to Assumption 2, we update the model with the loss of the entire dataset as the negative samples. Hence:

$$\left.\frac{\partial g^-}{\partial x_1}\right|_{x_1=\mu_1} = \sum_{i=2}^{N} \sum_{n=0}^{\infty} \zeta(W, n) x_i^n. \tag{36}$$

By substituting the sums, we derive:

$$\frac{\partial g^-}{\partial x_1}\bigg|_{x_1=\mu_1} = \sum_{n=0}^{\infty}\sum_{i=2}^{N}\zeta(W,n)x_i^n. \tag{37}$$

Since the distribution of data is assumed to be normal, these sums can be estimated using the expected value of $\mathbb{E}(\sum_{i=1}^{N} x_i^n) = N\mathbb{E}(x^n)$. Since x belongs to a normal distribution with $\mu = 0$ and $\sigma = 1$, the sum can be estimated as:

$$\sum_{i=1}^{N} x^n \simeq \begin{cases} (n-1)!! & n = 2k \quad \text{(Even powers)} \\ 0 & n = 2k+1 \quad \text{(Odd powers)} \end{cases} \tag{38}$$

By applying this estimate to Eq. (36), we derive:

$$\frac{\partial g^-}{\partial x_1}\bigg|_{x_1=\mu_1} = \sum_{k=0}^{\infty}\sum_{i=2}^{N}\zeta(W,n)(n-1)!! \tag{39}$$

Consequently, we can conclude that for all negative pairs, the gradient derivative $\frac{\partial g^-}{\partial x_i}\Big|_{x_i=\mu_i}$ only depends on the network weights. Hence, due to the fact that both matrices are symmetric, $\text{sign}(\frac{\partial g_i}{\partial x_i^f}\big|_{x_i^f=\mu_i^f}) = \text{sign}(\frac{\partial g_j}{\partial x_j^f}\big|_{x_j^f=\mu_j^f})$.

□

A.7 Corollary A1

Corollary A1. *Given the assumptions stated (under either Assumption 1 or Assumption 2), for two distinct domains characterized by random variables $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ from the distribution family $\mathcal{P}$, as $I(\boldsymbol{x}_i, \boldsymbol{x}_j)$ decreases, then the variance of the difference of the corresponding gradients, Var($\mathbf{g}_i - \mathbf{g}_j$), increases.*

This conclusion highlights our finding in Theorems 1 and 2 regarding the relationship between the distribution of different domains and their corresponding gradients.

Proof. The variance of $\mathbf{g}_i$ and $\mathbf{g}_j$ is:

$$\text{Var}(\mathbf{g}_i - \mathbf{g}_j) = \text{Var}(\mathbf{g}_i) + \text{Var}(\mathbf{g}_j) - 2\text{Cov}(\mathbf{g}_i, \mathbf{g}_j). \tag{40}$$

For deriving Var($\mathbf{g}_i$) and Var($\mathbf{g}_j$), we first introduce Lemma A3.

Lemma A3. *Given the assumptions stated, the variance of a differentiable function $\mathbf{g}$ can be estimated as:*

$$Var(\mathbf{g}) = \sum_{f=1}^{F}(\sigma^f)^2(\frac{\partial g}{\partial x^f}\bigg|_{x^f=\mu^f})^2, \tag{41}$$

where x^f is the f^{th} feature of $\mathbf{x}$, σ^f is its standard deviation, and μ^f is its mean. The full proof of this lemma is presented in Appendix A.8.

By derive Var($\mathbf{g}_i$) and Var($\mathbf{g}_j$) using Lemma 2 and Cov($\mathbf{g}_i, \mathbf{g}_j$) from Lemma 1, we derive:

$$\text{Var}(\mathbf{g}_i - \mathbf{g}_j) = \sum_{f=1}^{F}(\sigma_i^f)^2(\frac{\partial g_i}{\partial x_i^f}\bigg|_{x_i^f=\mu_i^f})^2 + (\sigma_j^f)^2(\frac{\partial g_j}{\partial x_j^f}\bigg|_{x_j^f=\mu_j^f})^2 - 2\sigma_{x_i^f,x_j^f}(\frac{\partial g_i}{\partial x_i^f}\bigg|_{x_i^f=\mu_i^f})(\frac{\partial g_j}{\partial x_j^f}\bigg|_{x_j^f=\mu_j^f}). \tag{42}$$

If we assume two domains that have the same distribution but different covariances with $\mathbf{g}_i$ denoted by $\mathbf{g}_{j1}$ and $\mathbf{g}_{j2}$, then:

$$\text{Var}(\mathbf{g}_i - \mathbf{g}_{j1}) - \text{Var}(\mathbf{g}_i - \mathbf{g}_{j2}) = \sum_{f=1}^{F}\alpha_f(\sigma_{ij1}^f - \sigma_{ij2}^f), \tag{43}$$

where $\alpha_f = 2(\frac{\partial g_i}{\partial x_i^f}\Big|_{x_i^f=\mu_i^f})(\frac{\partial g_j}{\partial x_j^f}\Big|_{x_j^f=\mu_j^f})$ which is always positive (According to Claims A1 and A2). □

This conclusion highlights our finding in Theorems 1 and 2 regarding the relationship between the distribution of different domains and their corresponding gradients.

A.8 Proof of Lemma A3

Proof. The variance of **g**, by definition, is:

$$\mathrm{Var}(\mathbf{g}) = \mathbb{E}(\mathbf{g} - \mathbb{E}\mathbf{g})^2. \tag{44}$$

From the proof of Lemma 2, we know that $\mathbb{E}g = g(\mu)$. By expanding g around μ we derive:

$$\mathrm{Var}(\mathbf{g}) = \mathbb{E}\left[\left(\sum_{f=1}^{F}(x^f - \mu^f)\left(\frac{\partial g}{\partial x^f}\Big|_{x^f=\mu^f}\right)\right)^2\right]. \tag{45}$$

The squared term can be expanded as:

$$\mathrm{Var}(\mathbf{g}) = \mathbb{E}[\sum_{f=1}^{F}\sum_{e=1}^{F}(x^f - \mu^f)(x^e - \mu^e)\left(\frac{\partial g}{\partial x^f}\Big|_{x^f=\mu^f}\right)\left(\frac{\partial g}{\partial x^e}\Big|_{x^e=\mu^e}\right)]. \tag{46}$$

As the features are assumed to be independent, when $f \neq e$, $\mathbb{E}(x^f - \mu^f)(x^e - \mu^e)$ represents $\mathrm{Cov}(x^f, x^e)$ and equals zero. Therefore:

$$\mathrm{Var}(\mathbf{g}) = \mathbb{E}\left[\sum_{f=1}^{F}(x^f - \mu^f)^2\left(\frac{\partial g}{\partial x^f}\Big|_{x^f=\mu^f}\right)^2\right]. \tag{47}$$

Since $\mathbb{E}(x^f - \mu^f)^2 = (\sigma^f)^2$, we derive:

$$\mathrm{Var}(\mathbf{g}) = \sum_{f=1}^{F}(\sigma^f)^2\left(\frac{\partial g}{\partial x^f}\Big|_{x^f=\mu^f}\right)^2, \tag{48}$$

which completes the proof for the variance of **g**. □

A.9 Proof of Proposition 1

Assumption 3. *Let $\mathbf{g}_i$ and $\mathbf{g}_j$ be two sets of gradients. $\mathbf{g}_{est}$ is defined as the average of both sets. We assume $cos(\mathbf{g}_{j,k}, \mathbf{g}_{est}) > 0$ for all $k \neq K$, $cos(\mathbf{g}_{i,k}, \mathbf{g}_{est}) > 0$ for all k, and $cos(\mathbf{g}_{j,K}, \mathbf{g}_{est}) < 0$. Without the loss of generality, we assume the last vector in the set is the unaligned vector.*

Proof. We first define $\hat{\mathbf{g}_j}$ as:

$$\mathbf{g}_j = \hat{\mathbf{g}_j} \cup \{\mathbf{g}_{j,K}\}. \tag{49}$$

Hence, the mean of $\mathbf{g}_j$ can be derived as:

$$\mu_{\mathbf{g}_j} = \frac{1}{K}((K-1)\mu_{\hat{\mathbf{g}_j}} + \mathbf{g}_{j,K}), \quad \mu_{\hat{\mathbf{g}_j}} = \frac{1}{K-1}\sum_{k=1}^{K-1}\mathbf{g}_{j,K}. \tag{50}$$

Now we shift to the covariance of the two sets:

$$\mathrm{Cov}(\mathbf{g}_i, \mathbf{g}_j) = \frac{1}{K-1}(\mathbf{g}_i - \mu_{\mathbf{g}_i})(\mathbf{g}_j - \mu_{\mathbf{g}_i})^T. \tag{51}$$

Substituting $\mathrm{Cov}(\mathbf{g}_i, \hat{\mathbf{g}_j})$ in Eq. 51, we derive:

$$\mathrm{Cov}(\mathbf{g}_i, \mathbf{g}_j) = \frac{K-2}{K-1}\mathrm{Cov}(\mathbf{g}_i, \hat{\mathbf{g}_j}) + \frac{1}{K-1}(\mathbf{g}_i - \mu_{\mathbf{g}_i})(\mathbf{g}_{j,k} - \mu_{\mathbf{g}_i})^T. \tag{52}$$

By substituting $\mu_{\mathbf{g}_j}$ in this equation we have:

$$\text{Cov}(\mathbf{g}_i, \mathbf{g}_j) = \frac{K-2}{K-1}\text{Cov}(\mathbf{g}_i, \hat{\mathbf{g}_j}) + \frac{1}{K-1}(\mathbf{g}_i - \mu_{\mathbf{g}_i})(\mathbf{g}_{j,k} - \frac{1}{K}((K-1)\mu_{\hat{\mathbf{g}_j}} + \mathbf{g}_{j,K}))^T. \quad (53)$$

Simplifying this equation we derive:

$$\text{Cov}(\mathbf{g}_i, \mathbf{g}_j) = \frac{K-2}{K-1}\text{Cov}(\mathbf{g}_i, \hat{\mathbf{g}_j}) + \frac{1}{K}(\mathbf{g}_i - \mu_{\mathbf{g}_i})(\mathbf{g}_{j,K} - \mu_{\hat{\mathbf{g}_j}})^T. \quad (54)$$

The term $(\mathbf{g}_i - \mu_{\mathbf{g}_i})(\mathbf{g}_{j,K} - \mu_{\hat{\mathbf{g}_j}})^T$ is a matrix with mostly negative components because of the following reasons. $(\mathbf{g}_i - \mu_{\mathbf{g}_i})$ represents the deviations of $\mathbf{g}_i$ vectors from their mean which points around $\mathbf{g}_{est}$ as $cos(\mathbf{g}_{i,k}, \mathbf{g}_{est}) > 0$ for all k. On the other hand, since $cos(\mathbf{g}_{j,K}, \mathbf{g}_{est}) < 0$, $(\mathbf{g}_{j,K} - \mu_{\hat{\mathbf{g}_j}})^T$ representing the deviation of $\mathbf{g}_{j,K}$ from the mean vector is in the opposite direction of $\mathbf{g}_{est}$. Hence, the dot product of two sets of vectors pointing in two opposite directions results in a negative covariance matrix which completes the proof. □

B Experiment Setup

We conduct our experiments using the **PACS** dataset [35], which comprises 9,991 images across four domains: ‘Photo’, ‘Art-painting’, ‘Cartoon’, and ‘Sketch’ denoted by the initial letter of their names. The images are divided into seven distinct classes. Following the methodology described in [40], we employ ResNet-18 [41] as the encoder network architecture, which is trained from scratch. SimCLR is utilized as the self-supervised learning technique to train this encoder due to its effectiveness in domain generalization. Data augmentations are implemented as per the guidelines in [28]. Specifically, a random patch of each image is selected and resized to 32×32. Two random transformations—horizontal flipping and color distortion—are subsequently applied. We also use a batch size of 128 and the optimization is performed using the Adam optimizer [42] with a learning rate of 3×10^{-3}.