Moment Alignment: Unifying Gradient and Hessian Matching for Domain Generalization

Yuen Chen[1] **Haozhe Si**[1] **Guojun Zhang**[2] **Han Zhao**[1]

[1]University of Illinois at Urbana-Champaign
[2]University of Waterloo
{yuenc2, haozhes3, hanzhao}@illinois.edu, g39zhang@uwaterloo.ca

Abstract

Domain generalization (DG) seeks to develop models that generalize well to unseen target domains, addressing distribution shifts in real-world applications. One line of research in DG focuses on aligning domain-level gradients and Hessians to enhance generalization. However, existing methods are computationally inefficient and the underlying principles of these approaches are not well understood. In this paper, we develop a theory of moment alignment for DG. Grounded in *transfer measure*, a principled framework for quantifying generalizability between domains, we prove that aligning derivatives across domains improves transfer measure. Moment alignment provides a unifying understanding of Invariant Risk Minimization, gradient matching, and Hessian matching, three previously disconnected approaches. We further establish the duality between feature moments and derivatives of the classifier head. Building upon our theory, we introduce **C**losed-Form **M**oment **A**lignment (CMA), a novel DG algorithm that aligns domain-level gradients and Hessians in closed-form. Our method overcomes the computational inefficiencies of existing gradient and Hessian-based techniques by eliminating the need for repeated backpropagation or sampling-based Hessian estimation. We validate our theory and algorithm through quantitative and qualitative experiments.

1 INTRODUCTION

Classic machine learning methods rely on the assumption that training and test data are drawn from the same distribution, typically described as being independent and identically distributed (*i.i.d.*). However, the *i.i.d.* assumption is often violated in real-world scenarios due to variations in sampling populations (Santurkar et al., 2020), temporal changes (Shankar et al., 2019), and geographic differences (Hansen et al., 2013; Christie et al., 2018). Performance degradation due to distribution shifts is particularly critical in high-stake applications. For instance, an autonomous driving system (Dai and Van Gool, 2018; Hu et al., 2021) trained on data collected in the United States may encounter different traffic conditions when deployed in other regions. Similarly, in medical imaging (Wachinger et al., 2021; AlBadawy et al., 2018; Tellez et al., 2019), models trained on data from one demographic group may face challenges when applied to a different demographic.

Domain generalization (DG) aims to tackle this issue by leveraging data from multiple source domains to learn a model that performs well on unseen but related target domains. Although various approaches have been studied to address the DG problem, including Invariant Risk Minimization (IRM) Arjovsky et al. (2020), gradient matching (Shi et al., 2021; Koyama and Yamaguchi, 2021; Parascandolo et al., 2020), Hessian matching (Rame et al., 2022; Hemati et al., 2023), and domain-invariant feature representation learning (Ben-David et al., 2010; Li et al., 2018; Tzeng et al., 2017; Hoffman et al., 2017; Muandet et al., 2013; Long et al., 2015; Zhao et al., 2019), these methods often appear disconnected and are based on different underlying principles. We discuss these related research in Appendix E.

We unify these seemingly disparate methods through the theory of moment alignment. Our theory builds upon *transfer measure*, a principled DG framework proposed by Zhang et al. (2021). We first extend the definition of transfer measure to multi-source DG, inducing a target error bound. We then prove that aligning the derivatives improves transfer measure under different assumptions: when there exists a classifier that is simultaneously optimal across all domains (referred to as the *IRM assumption*), and when there is not. We show that IRM, gradient matching, and Hessian matching approaches are special cases of moment alignment. Our theory explains the success of state-of-the-art methods like HGP and Hutchinson's algorithm (Hemati et al., 2023), which perform both gradient and Hessian matching. This

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 244:705–736.

Table 1: Comparison of our method and prior algorithms.

	ERM	IRM	Fish/IGA/AND-Mask	Fishr/CORAL	HGP/Hutchinson	**CMA**
Gradient Matching	No	Yes	Yes	No	Yes	**Yes**
Hessian Matching	No	No	No	Yes	Yes	**Yes**
Closed-Form Hessian	–	–	–	No	No	**Yes**

combined approach provides an advantage over methods that only match gradients or Hessians. Furthermore, we establish the duality between feature moments and the derivatives of the classifiers, thereby unifying these approaches.

Drawing from the theoretical results, we proposed **C**losed-Form **M**oment **A**lignment (CMA), a novel algorithm to DG that aligns the first- and second-order derivatives across domains. The loss objective in CMA is similar to those of HGP and Hutchinson's, but CMA enjoys computational efficiency by analytically computing gradients and Hessians. Our method bypasses the computational limitations of existing gradient and Hessian matching techniques that rely on repeated backpropagation or sampling-based estimation. Additionally, we provide two Hessian computation methods—direct Frobenius norm computation for faster performance at higher memory cost, and a memory-efficient method that reduces memory requirement at the expense of increased computation time. This flexibility allows users to balance memory usage and computational time.

The empirical evaluation consists of two settings designed to validate our theoretical framework and proposed algorithm. First, we conduct linear probing experiments on Waterbirds, CelebA, and MultiNLI datasets, where the IRM assumption holds. Second, we perform full fine-tuning experiments on selected datasets from the DomainBed benchmark (Gulrajani and Lopez-Paz, 2020), where the IRM assumption may not be satisfied. In the DomainBed experiment, where the IRM assumption is not guaranteed. We compare CMA with ERM, CORAL (Sun and Saenko, 2016), and Fishr (Rame et al., 2022). CMA's performance aligns with our theory and matches state-of-the-art performance.

Below we summarize our main contributions:

- *Unified Theory of Moment Alignment:* We develop a theory of moment alignment that unifies IRM, gradient matching, and Hessian matching. This unified framework enhances our understanding of the interplay between these methods and their combined effect on improving generalization across domains. We further establish the duality between feature moments and the classifier derivatives.
- *New Algorithm:* We propose **C**losed-Form **M**oment **A**lignment (CMA), a novel DG algorithm that performs both gradient and Hessian matching. CMA enjoys computational efficiency by analytically computing gradients and Hessians, avoiding the need for repeated backpropagation or sampling-based estimation. We offer two Hessian computation methods to optimize memory usage and computational speed.
- *Empirical Validation:* We validate CMA through both quantitative and qualitative analyses. CMA matches state-of-the-art performance while achieving superior worst-group accuracy and feature moment alignment, reducing first- and second-moment discrepancies more effectively than Fishr and ERM.

Our work offers a unified perspective that enhances theoretical understanding and practical performance in addressing distribution shifts. As summarized in Table 1, our method is, to the best of our knowledge, the first to achieve exact gradient and Hessian matching.

2 PRELIMINARIES

We consider the problem of DG, where predictors are trained on data drawn from a set of source domains and are evaluated on an unseen target domain. The goal is to learn a predictor that generalizes well to the target domain. Formally, the data are drawn from K source domains $\mathcal{S} := \{\mu_1, \ldots, \mu_K\}$ and a target domain $\mathcal{T} := \mu_{\mathcal{T}}$. Each domain μ_i is a distribution over the input space $\mathcal{X}$ and the label space $\mathcal{Y}$. The loss of a predictor $h : \mathcal{X} \to \mathcal{Y}$ on domain μ is defined as $\mathcal{L}_\mu(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mu}[\ell(h(\mathbf{x}), y)]$, where ℓ is the loss function on a single example. The goal of domain generalization is to learn $h \in \mathcal{H}$ to minimize the loss on the target domain $\mathcal{T}$: $\mathcal{L}_{\mu_{\mathcal{T}}}(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mu_{\mathcal{T}}}[\ell(h(\mathbf{x}), y)]$. ERM minimizes the average loss over the source domains, $\mathcal{L}_{\text{ERM}} := \frac{1}{K}\sum_{i=1}^{K}\mathcal{L}_{\mu_i}$. However, ERM often fails under distribution shifts, especially when the data exhibits spurious correlation. To address this, Arjovsky et al. (2020) proposes the IRM principle, aiming to jointly learn a features extractor and a predictor such that there exists a predictor on the extracted features that is optimal for all domains simultaneously. Subsequent studies, such as those by Rosenfeld et al. (2021) and Wang et al. (2022, 2023), have shown that IRM alone is not sufficient for DG. Recent work by Zhang et al. (2021) on *transferability* introduces a framework to measure how much success a predictor trained on one domain can transfer to another. Below we review the original definition of transfer measures between two domains and extend it to multi-source domain settings.

2.1 TRANSFER MEASURES

We restate the definitions of *transfer measures* (Zhang et al., 2021) and its induced target error bound.

Definition 1 (transfer measures (Zhang et al., 2021)). *Given some $\Gamma \subseteq \mathcal{H}$, $\mathcal{L}^*_{\mathcal{S}} := \inf_{h\in\Gamma}\mathcal{L}_{\mathcal{S}}(h)$ and $\mathcal{L}^*_{\mathcal{T}} := \inf_{h\in\Gamma}\mathcal{L}_{\mathcal{T}}(h)$, we define the one-sided transfer measure, symmetric transfer measure, and the realizable transfer*

measure respectively as:

$$\begin{aligned}\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) &:= \sup_{h\in\Gamma} \mathcal{L}_\mathcal{T}(h) - \mathcal{L}^*_\mathcal{T} - (\mathcal{L}_\mathcal{S}(h) - \mathcal{L}^*_\mathcal{S}) \\ \mathrm{T}_\Gamma(\mathcal{S},\mathcal{T}) &:= \max\{\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}), \mathrm{T}_\Gamma(\mathcal{T}\|\mathcal{S})\} \\ &= \sup_{h\in\Gamma} |\mathcal{L}_\mathcal{S}(h) - \mathcal{L}^*_\mathcal{S} - (\mathcal{L}_\mathcal{T}(h) - \mathcal{L}^*_\mathcal{T})| \\ \mathrm{T}^{\mathrm{r}}_\Gamma(\mathcal{S},\mathcal{T}) &:= \sup_{h\in\Gamma} |\mathcal{L}_\mathcal{S}(h) - \mathcal{L}_\mathcal{T}(h)|\end{aligned} \tag{1}$$

From the definition of one-sided transfer measure, we have the following target error bound.

Proposition 1 (target error bound (Zhang et al., 2021)). *For any $h \in \Gamma \subseteq \mathcal{H}$, the target error is bounded by:*

$$\mathcal{L}_\mathcal{T}(h) \leq \mathcal{L}_\mathcal{S}(h) + \mathcal{L}^*_\mathcal{T} - \mathcal{L}^*_\mathcal{S} + \mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \tag{2}$$

The implication of Proposition 1 is that by minimizing the loss on the source domain and the one-sided transfer measure between the source and target domains, we can effectively minimize an upper bound on the target loss.

2.2 APPROXIMATE HESSIAN ALIGNMENT

Hemati et al. (2023) proves an upper bound on the transfer measure by the spectral norm of the Hessian matrices between source and target domains and is the first to propose simultaneously aligning gradients and Hessians. However, their analysis is limited to the single source domain adaptation setting and assumes the existence of an *invariant optimal predictor*.

Definition 2 (invariant optimal predictor (Arjovsky et al., 2020)). *A predictor $h \in \mathcal{H}$ is an invariant optimal predictor if $\mathcal{L}_{\mu_i}(h) = \min_{h\in\mathcal{H}} \mathcal{L}_{\mu_i}(h)$ for all $i \in [K]$.*

Assumption 1 (IRM assumption). *There exists an invariant optimal predictor $h \in \mathcal{H}$ on the source domains $\mathcal{S} = \{\mu_i\}_{i=1}^K$.*

The algorithms in Hemati et al. (2023) approximate the Hessian matrices. Both methods are computationally intensive: Hessian-Gradient Product (HGP) requires repeated backpropagation, whereas Hutchinson's method relies on estimation through sampling.

In this work, we extend the analysis of Hessian alignment to DG, addressing scenarios both with and without the IRM assumption. We also propose a more efficient algorithm that analytically computes the Hessian matrices with respect to (w.r.t.) the classifier head.

2.3 NATURE OF DISTRIBUTION SHIFT

In the DG literature, there are two main types of assumptions on the underlying data-generating process and the nature of the distribution shift.

The first type relies on causal graphs (directed graphical models) to explicitly model the ground-truth data-generating distribution, over which one can also aim for the minimax out-of-distribution generalization performance using the invariant predictor principle (Peters et al., 2015; Arjovsky et al., 2020; Wang et al., 2023; Zhang et al., 2023). However, these explicit assumptions on the causal structure of the variables are often too restrictive and hard to verify in practice, due to the unobserved confounders.

The second type of assumption explicitly models the nature of the distribution shift, such as covariate shift, label shift, concept shift (Ben-David et al., 2010; Heinze-Deml et al., 2018; Zhao et al., 2019). These assumptions make technical analysis possible but often oversimplify the true real-world shifts, which rarely adhere strictly to such constraints. Moreover, these assumptions are sufficient but not necessary for provable OOD generalizations.

Given the limitations of these two types of assumptions, our work aims to broaden its potential applicability by avoiding explicit assumptions on the underlying data-generating distributions and the nature of distribution shifts. Instead, our approach focuses on the loss landscapes of the train and test distributions, which are more fine-grained and fundamental. We would also like to point out that typical explicit distribution shift assumptions, such as the covariate shift assumption, which is closely related to the line of work on invariant risk minimization, can be used to simplify certain terms in our generalization upper bound.

3 THEORY OF MOMENT ALIGNMENT

In this section, we first extend the transfer measures to multi-source domain generalization (Section 3.1) and prove a bound on the transfer measure independent of the target distribution (Section 3.2). We then apply this bound and Proposition 1 to show that aligning derivatives across domains minimizes the target loss both under the IRM assumption (Section 3.3), and when it does not hold (Section 3.4). We defer the proof of propositions, theorems, and corollaries to Appendix A, Appendix B, and Appendix C respectively.

3.1 TRANSFER MEASURES FOR MULTI-SOURCE DOMAINS

The original definition of transfer measures is defined only for a single source domain $\mathcal{S}$ and a target domain $\mathcal{T}$. Next, we first state the generalized definition to multiple source domains $\mathcal{S} = \{\mu_i\}_{i=1}^K$.

Definition 3 (transfer measures on multiple source domains). *Given $\mathcal{S} = \{\mu_i\}_{i=1}^K$, some $\Gamma \subseteq \mathcal{H}$, $\mathcal{L}^*_{\mu_i} := \inf_{h\in\Gamma} \mathcal{L}_{\mu_i}(h)$ for all $i \in [K]$, $\mathcal{L}^*_\mathcal{T} := \inf_{h\in\Gamma} \mathcal{L}_\mathcal{T}(h)$, $\mu^* := \arg\min_\mu \max_{i\in[K]} \mathrm{T}_\Gamma(\mu_i\|\mu)$, and $\mathcal{L}_\mathcal{S}(h) := \mathcal{L}_{\mu^*}(h)$. we define the one-sided transfer measure, symmetric transfer measure, and the realizable transfer measure*

respectively as:

$$\begin{aligned}
\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) :=& \sup_{h\in\Gamma} \mathcal{L}_\mathcal{T}(h) - \mathcal{L}_\mathcal{T}^* - (\mathcal{L}_\mathcal{S}(h) - \mathcal{L}_\mathcal{S}^*) \\
=& \sup_{h\in\Gamma} \mathcal{L}_\mathcal{T}(h) - \mathcal{L}_\mathcal{T}^* - \left(\mathcal{L}_{\mu^*}(h) - \mathcal{L}_{\mu^*}^*\right) \\
=& \mathrm{T}_\Gamma(\mu^*\|\mathcal{T}) \\
\mathrm{T}_\Gamma(\mathcal{S},\mathcal{T}) :=& \max\left\{\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}), \mathrm{T}_\Gamma(\mathcal{T}\|\mathcal{S})\right\} \\
=& \mathrm{T}_\Gamma(\mu^*,\mathcal{T}) \\
\mathrm{T}_\Gamma^{\mathrm{r}}(\mathcal{S},\mathcal{T}) :=& \sup_{h\in\Gamma} |\mathcal{L}_\mathcal{S}(h) - \mathcal{L}_\mathcal{T}(h)| \\
=& \sup_{h\in\Gamma} |\mathcal{L}_{\mu^*}(h) - \mathcal{L}_\mathcal{T}(h)| = \mathrm{T}_\Gamma^{\mathrm{r}}(\mu^*,\mathcal{T})
\end{aligned}$$

In words, we define the transfer measure between a set of domains $\mathcal{S}$ and a domain $\mathcal{T}$ as the transfer measure between a distribution μ^* and $\mathcal{T}$, where μ^* is the center of mass. Note that it is not necessary to find μ^* explicitly, as shown next. For the remainder of this paper, we use transfer measure to refer to the one-sided transfer measure and leave the analogous results for the symmetric and realizable transfer measures to Appendix D.

3.2 BOUNDING TRANSFER MEASURE

Although Definition 3 defines the transfer measure in the multiple source domain setting, in DG, one can only access the source domains $\mathcal{S} = \{\mu_i\}_{i=1}^K$. In this section, we prove an upper bound on the transfer measure on the target domain $\mathcal{T}$ under the following mixture assumption:

Assumption 2 (convex combination of source domains). *The target domain $\mathcal{T}$ is a convex combination of the source domains $\mathcal{S} = \{\mu_i\}_{i=1}^K$, i.e., $\exists w_i \geq 0$ and $\sum_{i=1}^K w_i = 1$ such that $\mu_\mathcal{T} = \sum_{i=1}^K w_i \mu_i$.*

Note that although Assumption 2 seems restrictive, this assumption generally holds. In particular, the assumption is satisfied in the simple case where each source distribution is a Gaussian. Moreover, as a well-known result in the literature of mixture models shows, when the number of mixture components is large enough, any smooth continuous distribution can be well-approximated by a mixture of Gaussians. Furthermore, in the literature of domain generalization, similar assumptions have been adopted as well for the purpose of analysis and design of algorithms (Hu et al., 2018; Sagawa et al., 2020; Krueger et al., 2021).

From the definition of transfer measure and Assumption 2, we have the following proposition.

Proposition 2 (upper bound on transfer measure). *Given $\mathcal{S} = \{\mu_i\}_{i=1}^K$ and some $\Gamma \subseteq \mathcal{H}$. Define $\mathcal{L}_{\mu_i}^* := \inf_{h\in\Gamma} \mathcal{L}_{\mu_i}(h)$ for all $i \in [K]$, $\mathcal{L}_\mathcal{T}{}^* := \inf_{h\in\Gamma} \mathcal{L}_\mathcal{T}(h)$, $\mu^* := \arg\min_\mu \max_{i\in[K]} \mathrm{T}_\Gamma(\mu_i\|\mu)$, and $\mathcal{L}_\mathcal{S}(h) := \mathcal{L}_{\mu^*}(h)$. Under Assumption 2, we have:*

$$\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq \frac{1}{2} \max_{i\neq j} \mathrm{T}_\Gamma(\mu_j\|\mu_i) \tag{3}$$

A direct consequence of Proposition 2 is a target error bound analogous to Proposition 1, but does not require the knowledge of the target domain $\mathcal{T}$ except that it is a mixture of the source distributions.

Proposition 3 (target error bound – multiple source domains). *Given $\Gamma \subseteq \mathcal{H}$, for any $h \in \Gamma$, the target error is bounded by:*

$$\mathcal{L}_\mathcal{T}(h) \leq \mathcal{L}_\mathcal{S}(h) + \mathcal{L}_\mathcal{T}^* - \mathcal{L}_\mathcal{S}^* + \frac{1}{2} \max_{i\neq j} \mathrm{T}_\Gamma(\mu_j\|\mu_i) \tag{4}$$

Having established the transfer measure provides an upper bound on the target error, we now focus on bounding the transfer measure.

3.3 MOMENT ALIGNMENT UNDER IRM ASSUMPTION

For ease of notation, we assume that the classifier $h \in \mathcal{H}$ is parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$ so that $\mathcal{L}_\mu(h) = \mathcal{L}_\mu(\boldsymbol{\theta})$.

Theorem 1 (moment alignment under IRM). *Given K source domains $\mathcal{S} = \{\mu_i\}_{i=1}^K$ and a target domain $\mathcal{T} \in conv(\mu_1,\ldots,\mu_k)$, assume the losses $\mathcal{L}_{\mu_i}\ \forall i \in [K]$ are ν-strongly convex w.r.t. the classifier head and M-times differentiable. Under the IRM assumption (Assumption 1), let $\boldsymbol{\theta}^*$ be the optimal invariant predictor, $\Gamma = \arg\min(\mathcal{L}_\mathcal{S}, \delta_\mathcal{S})_\mathcal{H} := \left\{\boldsymbol{\theta} \mid h_{\boldsymbol{\theta}\in\mathcal{H}} : \max_{i\in[K]} (\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)) \leq \delta_\mathcal{S}\right\}$, and $\delta = \frac{2\delta_\mathcal{S}}{\nu}$, we have:*

$$\begin{aligned}
\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq \max_{i\neq j} \Bigg(& \sum_{n=2}^{N} \frac{1}{n!} \delta^{\frac{n}{2}} \|\nabla_{\boldsymbol{\theta}}^n \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) \\
& - \nabla_{\boldsymbol{\theta}}^n \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)\|_F^n \Bigg) + o(\delta^{\frac{N}{2}})
\end{aligned} \tag{5}$$

for any integer $2 \leq N \leq M$. $\nabla_{\boldsymbol{\theta}}^{\mathrm{n}} \mathcal{L}(\boldsymbol{\theta})$ is an n^{th} order tensor with dimension $d \times \cdots \times d$ (n times) where $\nabla_{\boldsymbol{\theta}}^{\mathrm{n}} \mathcal{L}(\boldsymbol{\theta})_{(k_1,\cdots,k_n)} = \frac{\partial^n \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{k_1} \ldots \partial \boldsymbol{\theta}_{k_n}}$.

The implication of Theorem 1 is that the transfer measure is upper-bounded by the sum of the differences in higher-order derivatives across domains. Specifically, this suggests that aligning higher-order moments of the loss function promotes domain generalization.

Consider the special case when $N = 2$, we recover an upper bound similar to the Hessian alignment theorem in Hemati et al. (2023), which we state next as a corollary.

Corollary 2 (hessian alignment under IRM). *Under the same setup as in Theorem 1, we have:*

$$\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq \frac{1}{2}\delta \max_{i\neq j} \|\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*) - \mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\|_F + o(\delta) \tag{6}$$

where $\mathbf{H}(\boldsymbol{\theta})$ denotes the Hessian matrix of $\mathcal{L}(\boldsymbol{\theta})$.

Corollary 2 implies that the transfer measure is upper-bounded by the Frobenius norm of the difference of the Hessian matrices across domains. This result aligns with the findings of Hemati et al. (2023). However, unlike their result, Corollary 2 does not require knowledge of the target domain; instead, it relies on the assumption of $\mathcal{T}$ being a convex combination of the source domains.

3.4 MOMENT ALIGNMENT WITHOUT IRM ASSUMPTION

In practice, unless the features are explicitly trained or post-processed to satisfy the IRM assumption—such as through IRM training or Invariant Feature Subspace Recovery (Wang et al., 2023)—invariant optimal predictors generally do not exist. In this section, we derive a bound on the transfer measure under this setting.

Assumption 3 (bounded gradients, approximate IRM). *There exists a constant $g > 0$ such that $\min_{\boldsymbol{\theta}\in\mathcal{H}} \max_{i\in[K]} \|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta})\|_2 \leq g$.*

Theorem 3 (moment alignment). *Given K source domains $\mathcal{S} = \{\mu_i\}_{i=1}^K$ and target domain $\mathcal{T} \in conv(\mu_1, \ldots, \mu_k)$. Assume loss $\mathcal{L}_{\mu_i}$, $\forall i \in [K]$ are ν-strongly convex and M-times differentiable w.r.t. the classifier head. Let $\mathcal{P}(\{\mathcal{L}_{\mu_i}\}_{i=1}^K) := \{\boldsymbol{\theta} : \max_{i\in[K]}(\boldsymbol{\theta'} - \boldsymbol{\theta})^\top \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) \geq 0, \ \forall \boldsymbol{\theta'} \in \Gamma\}$ (a set of weakly Pareto optimal points for the objectives $\{\mathcal{L}_{\mu_i}\}_{i=1}^K$), and let $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\mathcal{P}} \max_{i\in[K]} \|\nabla\mathcal{L}_{\mu_i}(\boldsymbol{\theta})\|_2$, $\Gamma := \left\{\boldsymbol{\theta} \in \mathcal{H} : \max_{i\in[K]} \left(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)\right) \leq \delta_{\mathcal{S}}\right\}$, and $\delta = \frac{2\delta_{\mathcal{S}}}{\nu}$, we have:*

$$\begin{aligned}\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq & \frac{1}{2}\max_{i\neq j}\Bigg(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}_j^*) \\ & - \left(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*)\right) \\ & + \sum_{n=1}^{N}\frac{1}{n!}\delta^{\frac{n}{2}}\left\|\nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_\mathtt{j}}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_\mathtt{i}}(\boldsymbol{\theta}^*)\right\|_F\Bigg) \\ & + o(\delta^{\frac{N}{2}})\end{aligned} \tag{7}$$

where $\boldsymbol{\theta}_i^*$ is the minimizer of $\mathcal{L}_{\mu_i}(\boldsymbol{\theta})$. Furthermore, suppose Assumption 3 holds with $g > 0$:

$$\begin{aligned}\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq & \delta^{\frac{1}{2}}g + \frac{1}{2}\max_{i\neq j}\Bigg(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}_j^*) \\ & - \left(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*)\right) \\ & + \sum_{n=2}^{N}\frac{1}{n!}\delta^{\frac{n}{2}}\left\|\nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_\mathtt{j}}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_\mathtt{i}}(\boldsymbol{\theta}^*)\right\|_F\Bigg) \\ & + o(\delta^{\frac{N}{2}})\end{aligned} \tag{8}$$

As a special case, when $N = 2$, we have the following guarantee on gradient and Hessian alignment.

Corollary 4 (hessian alignment). *Under the same setup as in Theorem 3, we have:*

$$\begin{aligned}\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq & \frac{1}{2}\max_{i\neq j}\Bigg(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}_j^*) \\ & - \left(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*)\right) \\ & + \delta^{\frac{1}{2}}\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*)\right\|_2 \\ & + \frac{1}{2}\delta\left\|\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*) - \mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\right\|_F\Bigg) + o(\delta)\end{aligned} \tag{9}$$

where $\boldsymbol{\theta}_i^*$ is the minimizer of $\mathcal{L}_{\mu_i}(\boldsymbol{\theta})$.

Furthermore, suppose Assumption 3 holds with $g > 0$:

$$\begin{aligned}\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq & \delta^{\frac{1}{2}}g + \frac{1}{2}\max_{i\neq j}\Bigg(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}_j^*) \\ & - \left(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*)\right) \\ & + \frac{1}{2}\delta\left\|\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*) - \mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\right\|_F\Bigg) + o(\delta)\end{aligned} \tag{10}$$

To summarize, under the IRM assumption, the transfer measure is bounded by the differences in higher-order derivatives (second order and above) across domains. Conversely, when the IRM assumption does not hold, the transfer measure is bounded by the maximum optimality gaps and the gradient norms, in addition to the differences in higher-order derivatives across domains.

The implication of Theorem 3 is the necessity of minimizing the gradient norm, as the upper bound depends on it regardless of the differences in higher-order derivatives. Fortunately, this bound can be reduced by incorporating gradient norm minimization—a strategy already embedded in many existing methods, as we will see later.

The results above rely on the assumption that the loss is strongly convex w.r.t. the classifier head, which is satisfied by widely used losses with L2 regularization, such as cross-entropy loss or mean-square error.

4 MOMENT ALIGNMENT: A UNIFYING FRAMEWORK

While various approaches to DG exist, they appear largely disconnected, and, to the best of our knowledge, no prior work has explicitly drawn connections between them. In this section, we unify IRM, gradient matching, and Hessian matching under the CMA framework. We further establish a duality between feature learning space and classifier fitting.

4.1 IRM AS MOMENT ALIGNMENT

When the features are fixed and satisfy the IRM assumption, minimizing the IRMv1 objective (Arjovsky et al., 2020)

$$\mathcal{L}_{\text{IRM}} := \mathcal{L}_{\text{ERM}} + \lambda\frac{1}{K}\sum_{i=1}^{K}\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta})\|_2^2, \tag{IRMv1}$$

recovers such invariant optimal predictor, and Theorem 1 provides an upper bound on the target error. On the other hand, when the fixed features do not satisfy the IRM assumption, The IRMv1 penalty seeks a parameter $\boldsymbol{\theta}$ whose average gradient norm is small, thereby minimizing g in the upper bound in Theorem 3.

4.2 GRADIENT AND HESSIAN MATCHING AS MOMENT ALIGNMENT

Their general gradient and Hessian matching objectives are either the following or their variants:

$$\mathcal{L}_{\text{GM}} := \mathcal{L}_{\text{ERM}} + \lambda \frac{1}{K} \sum_{i=1}^{K} \left\| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \overline{\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})} \right\|_2^2 \quad \text{(GM)}$$

$$\mathcal{L}_{\text{HM}} := \mathcal{L}_{\text{ERM}} + \lambda \frac{1}{K} \sum_{i=1}^{K} \| \mathbf{H}_{\mu_i}(\boldsymbol{\theta}) - \overline{\mathbf{H}(\boldsymbol{\theta})} \|_F^2 \quad \text{(HM)}$$

By their definitions, gradient matching and Hessian matching are special cases of moment alignment, reducing the first-order and second-order terms, respectively, in the upper bound of the transfer measure. Notably, when the IRM assumption holds, the penalty in Eq. (GM) will favor an invariant optimal predictor.

From the results in Section 3, aligning both gradients and Hessians improves DG over aligning only one of them. This explains the success of HGP and Hutchinson (Hemati et al., 2023) over methods that focus on gradient matching (Shi et al., 2021; Parascandolo et al., 2020; Koyama and Yamaguchi, 2020) or Hessian matching (Rame et al., 2022; Sun and Saenko, 2016).

4.3 FEATURE MATCHING AS MOMENT ALIGNMENT

So far, we have discussed moment alignment under fixed features. Next, we establish a connection between the derivatives of the classifier and moments of features, where the classifier is assumed to be the last layer of an NN, i.e., linear predictor over the learned features.

For a softmax classifier, the prediction is a function of $\mathbf{x}^\top \boldsymbol{\theta}$, where $\mathbf{x}$ is a feature vector and $\boldsymbol{\theta}$ is the classifier. Therefore, $\nabla_{\boldsymbol{\theta}}^{\text{n}} \ell(\boldsymbol{\theta})$ involves the n^{th} moment of $\mathbf{x}$, and by matching the n^{th} order derivatives w.r.t. the classifier head, we are matching the n^{th} moment of $\mathbf{x}$ across domains. Another view of this duality is that by the symmetry between $\mathbf{x}$ and $\boldsymbol{\theta}$, we can derive analogously results in Section 3 with optimization target $\mathbf{x}$.

IRM (Ahuja et al., 2020) and CORAL (Sun and Saenko, 2016) are two concrete examples of this feature-parameter duality. Going from the feature space to the parameter space, CORAL (Sun and Saenko, 2016) matches the feature covariance, namely the second moment of $\mathbf{x}$. Thus, CORAL is approximately Hessian matching in the parameter space. We refer interested readers to Proposition 4 in Hemati et al. (2023) for discussion on the attributes aligned by CORAL. Conversely, starting from the parameter space and moving to the feature space, the penalty term in Eq. (IRMv1) regularizes the gradient w.r.t. the classifier, corresponding to the first-moment alignment in the feature space, i.e., aligning the features themselves.

5 CLOSED-FORM MOMENT ALIGNMENT

Motivated by the theory of moment alignment, we introduce **C**losed-Form **M**oment **A**lignment (CMA), a DG algorithm that minimizes the following objective:

$$\begin{aligned} \mathcal{L}_{\text{CMA}} = & \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}_{\mu_i} + \alpha \| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i} - \overline{\nabla_{\boldsymbol{\theta}} \mathcal{L}} \|_2^2 \\ & + \beta \| \mathbf{H}_{\mu_i} - \overline{\mathbf{H}} \|_F^2, \end{aligned} \quad \text{(CMA)}$$

where $\overline{\nabla_{\boldsymbol{\theta}} \mathcal{L}} = \frac{1}{K} \sum_{i=1}^{K} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i}$ and $\overline{\mathbf{H}} = \frac{1}{K} \sum_{i=1}^{K} \mathbf{H}_{\mu_i}$ are the average gradient and Hessian. Similar to HGP and Hutchinson (Hemati et al., 2023), CMA aligns the gradients and Hessians across domains, but we compute the derivatives in closed form. In Appendix F, we connect CMA and other DG algorithms.

Gradient and Hessian matching (Koyama and Yamaguchi, 2020; Shi et al., 2021; Hemati et al., 2023; Rame et al., 2022), despite their theoretical and empirical success, often incur significant computations due to multiple backpropagations for a single update. CMA bypasses this limitation by analytically computing gradient and Hessian penalty.

5.1 CLOSED-FORM GRADIENT AND HESSIAN

CMA computes the gradient and Hessian penalty without requiring additional backpropagations. Leveraging closed-form solutions for the gradients and Hessians of the cross-entropy loss w.r.t. a linear classifier, CMA reduces computational overhead. The derivations are provided in Appendix G.

5.2 MEMORY-EFFICIENT HESSIAN MATCHING

The Hessian of the cross-entropy loss for a single feature vector $\mathbf{x}$ w.r.t. a softmax classifier is:

$$\mathbf{H} = \left(\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top \right) \otimes \left(\mathbf{x}\mathbf{x}^\top \right),$$

where $\mathbf{p} \in \mathbb{R}^C$ is a vector of predicted probabilities, $\text{diag}(\mathbf{p}) \in \mathbb{R}^{C \times C}$ is the diagonal matrix with elements of $\mathbf{p}$, $\mathbf{x}\mathbf{x}^\top \in \mathbb{R}^{d \times d}$, and $\otimes$ is the Kronecker product.

The dimension of $\mathbf{H}$ is quadratic in the number of classes C and feature dimension d, which could be memory-prohibitive under many features or classes. To mitigate this issue, we use properties of the Frobenius norm to avoid storing the full $dC \times dC$ matrix. Instead, we compute:

$$\| \mathbf{H} \|_F^2 = \text{tr}\left(\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top \right) \text{tr}\left(\mathbf{x}\mathbf{x}^\top \right) \quad (11)$$

which only requires storing a $C \times C$ and a $d \times d$ matrix.

5.3 TRADE-OFFS IN HESSIAN COMPUTATION

Our code offers two versions of Hessian computation, each with its trade-offs. The first version directly computes the Frobenius norm of the Kronecker product, which is faster but requires more memory. The second version avoids storing the full Kronecker product matrix, reducing memory usage, but requires computing traces for all pairs of Hessians. We lay out the derivation in Appendix G.1.4.

6 EXPERIMENTS

We validate CMA through both quantitative and qualitative analyses. First, we describe the experimental setup, including dataset details and model training procedures. We then present quantitative results, evaluating CMA's performance under the IRM assumption and scenarios where it does not hold. Finally, we conduct qualitative analyses to better understand CMA's effect on worst-group performance and feature moment alignment.

6.1 IMPLEMENTATION

Linear Probing (IRM) We evaluate liner probing performance on Waterbirds (Sagawa et al., 2020), CelebA (Liu et al., 2015), and MultiNLI (Williams et al., 2018). To enforce the IRM assumption, we apply the Invariant-feature Subspace Recovery (ISR) algorithm (Wang et al., 2022, 2023), which provably yields features that induce an optimal invariant predictor. For Waterbirds and CelebA, we extract features from a CLIP-pretrained Vision Transformer (ViT-B/32). For MultiNLI, we fine-tune a BERT model using the code and hyperparameters in Sagawa et al. (2020), then extract features from the fine-tuned model. These features are transformed using the ISR-mean algorithm (Wang et al., 2022, 2023). Finally, we train a linear classifier using ERM, Fishr (Rame et al., 2022), and CMA objectives.

Full Fine-Tuning (Non-IRM) We run end-to-end fine-tuning on a subset of DomainBed (Gulrajani and Lopez-Paz, 2020), applying gradient and Hessian regularization from Eq. (CMA) to the classifier head while back-propagating the loss through both the linear classifier and the encoder. Specifically, penalizing large gradient variance aligns gradients across domains, while the ERM loss drives gradients toward zero. The two mechanisms promote a small gradient norm for each domain, aligning with the theory in Section 3.4 and Section 4.3. Given recent empirical evidence supporting strong DG capabilities of Vision Transformers (Ghosal et al., 2022; Zheng et al., 2022; Sultana et al., 2022), we have selected ViT-S as the backbone for DomainBed experiments. Using the DomainBed codebase (Gulrajani and Lopez-Paz, 2020), we compare ERM (Vapnik, 1999), CORAL (Sun and Saenko, 2016), Fishr (Rame et al., 2022), and CMA by fine-tuning small Vision Transformers (Steiner et al., 2022; Dosovitskiy et al., 2021; Wightman, 2019). For more implementation details, please refer to Appendix H.1 and Appendix H.2.

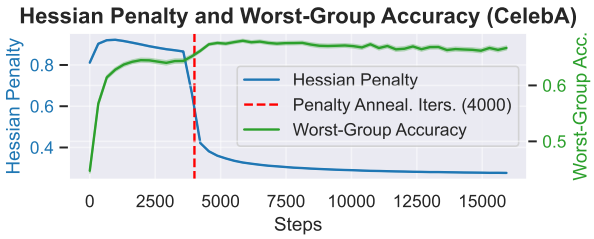

Figure 1: Hessian Penalty and worst-case accuracy on CelebA. Both curves represent the mean values, with shaded areas indicating $\pm$ one standard deviation over five runs.

6.2 QUANTITATIVE RESULTS

Our goal is not to claim that CMA surpassed existing algorithms but to demonstrate that our framework encompasses gradient matching (e.g., Koyama and Yamaguchi (2020)) and Hessian matching (e.g., Sun and Saenko (2016); Rame et al. (2022); Hemati et al. (2023)). To this end, our experimental results confirm that CMA achieves performance comparable to state-of-the-art moment matching methods.

Linear Probing (IRM) From Table 2, we observe that CMA consistently outperforms ERM on worst-group accuracy while maintaining comparable average accuracy across all datasets. Compared to Fishr, CMA achieves higher worst-group and average accuracy on two out of three datasets. In contrast, Fishr's performance varies, underperforming ERM on CelebA. Compared to CORAL, CMA achieves better worst-group performance across all datasets, while maintaining similar average accuracy.

Full Fine-Tuning (Non-IRM) We follow Rame et al. (2022) to employ the test-domain model selection method, where the validation set is a holdout set from the test domain. As shown in Table 3, CMA achieves comparable performance to Fishr, with both methods consistently outperforming ERM. This result supports the performance guarantee in Corollary 4 and validates our unified framework. Please refer to Appendix H.3 for per-dataset and training-domain validation performance.

6.3 QUALITATIVE RESULTS

Effect of Hessian Matching We analyze CMA's training progression and its impact on worst-group performance by plotting the Hessian loss:

$$\frac{\beta}{K}\sum_{i=1}^{K}\|\mathbf{H}_{\mu_i}(\boldsymbol{\theta}) - \overline{\mathbf{H}(\boldsymbol{\theta})}\|_F^2$$

for linear probing on the CelebA dataset, with the same hyperparameters as those reported for accuracy in Table 2 ($\alpha = 5000$, $\beta = 100$, penalty annealing iterations $= 4000$). As shown in Figure 1, near step 4000, when the gradient and Hessian matching terms take effect, there is a sharp drop in Hessian penalty, accompanied by a noticeable increase in worst-case accuracy, aligning with our theory that aligning Hessians across domains improves worst-case performance.

Table 2: Test accuracy (%) with standard error over three datasets. Each experiment is repeated over 5 seeds.

Method	Waterbirds (CLIP ViT-B/32)		CelebA (CLIP ViT-B/32)		MultiNLI (BERT)	
	Average	Worst-Group	Average	Worst-Group	Average	Worst-Group
ERM	89.52 ± 0.10	84.58 ± 0.20	78.76 ± 0.03	72.22 ± 0.39	81.15 ± 0.30	68.82 ± 0.64
CORAL	89.67 ± 0.14	84.85 ± 0.22	**78.81 ± 0.03**	73.00 ± 0.22	81.22 ± 0.21	68.71 ± 0.52
Fishr	89.79 ± 0.10	86.08 ± 0.10	73.95 ± 0.86	69.63 ± 1.20	**81.35 ± 0.16**	**71.55 ± 1.20**
CMA	**90.11 ± 0.17**	**86.16 ± 0.29**	77.87 ± 0.04	**74.16 ± 0.10**	81.30 ± 0.25	69.72 ± 0.66

Table 3: DomainBed results with test-domain validation model selection.

Algorithm	ColoredMNIST	RotatedMNIST	VLCS	PACS	TerraIncognita	Avg
ERM	54.5 ± 0.2	97.8 ± 0.1	76.9 ± 0.3	80.2 ± 0.5	36.5 ± 0.5	69.2
CORAL	55.7 ± 0.5	**98.0 ± 0.0**	75.9 ± 0.2	80.2 ± 0.2	33.6 ± 0.5	68.7
Fishr	62.0 ± 1.7	97.9 ± 0.0	**77.5 ± 0.5**	81.5 ± 0.2	37.3 ± 1.1	71.2
CMA	**62.5 ± 0.9**	97.9 ± 0.1	77.4 ± 0.8	**81.6 ± 0.3**	**38.4 ± 1.2**	**71.5**

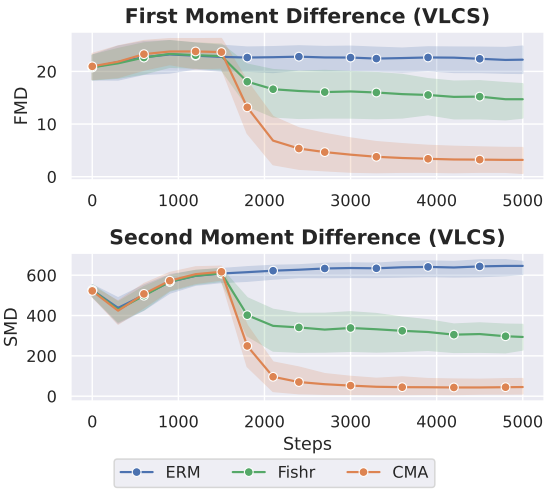

Figure 2: Comparison of first and second-moment differences across test domains for the VLCS dataset. The plots show the progression of moment differences over training steps for ERM, Fishr, and CMA. ERM fails to align the feature moments while CMA achieves the most effective alignment. The shaded regions represent one standard deviation above and below the mean across test domains.

Feature Moment Matching As discussed in Section 4.3, we illustrate the effect of CMA in matching the moments of features across domains. Figure 2 presents the moment differences between domains on VLCS dataset, where we average over all test domains. While ERM shows significant discrepancies in feature moments between domains, both Fishr and CMA successfully reduce these differences. Notably, CMA is more effective in reducing both first and second-moment disparities.

6.4 RUNTIME AND MEMORY COMPARISON

We report the average time per step (in seconds) and memory usage (in GB) for each (algorithm, dataset) pair in Table 4 and Table 5. It is important to note that, in addition to the algorithms' efficiency, the wall-clock time also depends on the hardware status at the time of training. We include additional comparisons of two versions of CMA, HGP, and Hutchinson algorithms, where "CMA (Speed)" uses the time-efficient Hessian computation, while "CMA (Memory)" uses the memory-efficient Hessian computation.

Among the methods compared, only CMA and Hutchinson compute full Hessian matrices. While CMA is inherently slower than Fishr, CORAL, and HGP, which rely on diagonal approximations of the Hessian, it remains more time-efficient than Hutchinson's.

To highlight the scalability of "CMA (Memory)", we run small-scale experiments on OfficeHome, a dataset with 65 classes. In this setting, "CMA (Speed)" requires more than 75 GB of memory and could not run on a single GPU, whereas "CMA (Memory)" completed successfully with peak usage under 13.7 GB.

7 LIMITATIONS AND FUTURE DIRECTIONS

Our analysis assumes that the target distribution is in the convex hull of the source distributions, which may not always hold or be verifiable in practice. It might be of future interest to relax this convexity assumption to accommodate a broader range of target distributions.

While closed-form Hessians eliminate the need for sampling-based approximations or multiple backpropagations, they introduce scalability challenges. Specifically, the Hessian matrix of the cross-entropy loss w.r.t. the classifier head scales quadratically with the number of classes and feature dimensions. To remedy this challenge, we provide a memory-efficient alternative for computing the Hessian Frobenius norm, albeit at the cost of longer computation time. Future work could explore Hessian approximations to further balance efficiency and accuracy.

The primary focus of this work is on a unifying theory for DG using gradient and Hessian matching. Our theory suggests that aligning higher-order derivatives improves

Table 4: Wall-clock time across datasets (in seconds). Algorithms are grouped by the type of moment matching. For each dataset, we bold the most time-efficient algorithm within each category.

Algorithm	ColoredMNIST (2)	RotatedMNIST (10)	VLCS (5)	PACS (7)	TerraIncognita (10)	OfficeHome (65)
No Moment Matching						
ERM	0.0278	0.0403	0.4019	0.3620	0.4216	0.4064
Approximate Second-Order						
CORAL	**0.0457**	**0.1003**	0.6241	**0.5244**	0.7697	0.5279
Fishr	0.0925	0.1331	0.7472	0.6757	0.6057	0.6600
HGP	0.0657	0.1292	**0.6048**	0.6729	**0.6045**	**0.4977**
Exact Second-Order						
Hutchinson	4.1663	9.7935	7.7604	7.3284	7.7270	7.8446
CMA (Speed)	**0.0676**	**0.1326**	**0.7354**	**0.7266**	**0.7421**	–
CMA (Memory)	0.1226	0.4723	0.8874	1.1699	1.0685	**0.8495**

Table 5: Memory usage across datasets (in GB). For each dataset, we bold the most memory-efficient algorithm within each category.

Algorithm	ColoredMNIST (2)	RotatedMNIST (10)	VLCS (5)	PACS (7)	TerraIncognita (10)	OfficeHome (65)
No Moment Matching						
ERM	0.1550	0.3728	6.8865	6.8865	6.8865	6.8868
Approximate Second-Order						
CORAL	**0.1391**	0.3190	6.7093	6.7093	6.7093	6.7097
Fishr	0.3192	0.7936	14.1433	14.1436	14.1441	14.1522
HGP	0.1477	**0.3099**	**5.6835**	**5.6835**	**5.6836**	**5.6843**
Exact Second-Order						
Hutchinson	**0.1502**	**0.3496**	**5.7047**	**5.7125**	**5.7284**	**6.0029**
CMA (Speed)	0.2867	1.4537	13.9511	14.8391	16.7272	~75
CMA (Memory)	0.3914	0.7776	13.6447	13.6448	13.6448	13.6474

generalization, but in practice, even second-order alignment is computationally demanding. The feasibility and potential benefits of higher-order alignments remain open questions, presenting intriguing directions for future research.

8 CONCLUSIONS

We introduced a unified theory of moment alignment for DG, providing upper bounds by examining differences in derivatives. The moment alignment framework reinterprets IRM, gradient matching, and Hessian matching, explaining the success of algorithms that match both gradients and Hessians across domains. Additionally, we established the duality between moments of features and derivatives of classifier heads, a novel perspective that we believe will open new research avenues.

Inspired by our theory, we proposed Closed-Form Moment Alignment (CMA), a DG algorithm that aligns gradients and Hessians analytically, avoiding the computational inefficiencies of previous methods that relied on repeated backpropagation or sampling-based Hessian estimations. We validated the efficacy of CMA through both quantitative and qualitative experiments. The results demonstrated that CMA achieves performance on par with state-of-the-art methods (e.g., Fishr). These findings not only confirm our theoretical predictions but also underscore the practical benefits of our moment alignment framework.

Acknowledgements

This work is partially supported by an NSF IIS grant with No. 2416897. HZ would like to thank the support from a Google Research Scholar Award. This research used the Delta advanced computing and data resource which is supported by the National Science Foundation (award OAC 2005572) and the State of Illinois. We thank Haoxiang Wang for suggestions of our experiments. The views and conclusions expressed in this paper are solely those of the authors and do not necessarily reflect the official policies or positions of the supporting companies and government agencies.

References

Kartik Ahuja, Karthikeyan Shanmugam, Kush R. Varshney, and Amit Dhurandhar. Invariant Risk Minimization Games, March 2020. URL http://arxiv.org/abs/2002.04692. arXiv:2002.04692 [cs, stat].

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization, November 2022a. URL http://arxiv.org/abs/2106.06607. arXiv:2106.06607 [cs, stat].

Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R. Varshney. Empirical or Invariant Risk Minimization? A Sample Complexity Perspective, August 2022b. URL http://arxiv.org/abs/2010.16412. arXiv:2010.16412 [cs, stat].

Ehab A. AlBadawy, Ashirbani Saha, and Maciej A. Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical Physics*, 45(3):1150–1158, March 2018. ISSN 2473-4209. doi: 10.1002/mp.12752.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization, March 2020. URL http://arxiv.org/abs/1907.02893. arXiv:1907.02893 [cs, stat].

Sara Beery, Grant van Horn, and Pietro Perona. Recognition in Terra Incognita, July 2018. URL http://arxiv.org/abs/1807.04975. arXiv:1807.04975 [cs, q-bio].

C. Bekas, E. Kokiopoulou, and Y. Saad. An estimator for the diagonal of a matrix. *Applied Numerical Mathematics*, 57(11):1214–1229, November 2007. ISSN 0168-9274. doi: 10.1016/j.apnum.2007.01.003. URL https://www.sciencedirect.com/science/article/pii/S0168927407000244.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, May 2010. ISSN 1573-0565. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from Several Related Classification Tasks to a New Unlabeled Sample. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://papers.nips.cc/paper_files/paper/2011/hash/b571ecea16a9824023ee1af16897a582-Abstract.html.

Kuang-Hua Chang. Chapter 19 - Multiobjective Optimization and Advanced Topics. In Kuang-Hua Chang, editor, *e-Design*, pages 1105–1173. Academic Press, Boston, January 2015. ISBN 978-0-12-382038-9. doi: 10.1016/B978-0-12-382038-9.00019-3. URL https://www.sciencedirect.com/science/article/pii/B9780123820389000193.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional Map of the World, April 2018. URL http://arxiv.org/abs/1711.07846. arXiv:1711.07846 [cs].

Dengxin Dai and Luc Van Gool. Dark Model Adaptation: Semantic Image Segmentation from Daytime to Nighttime, October 2018. URL http://arxiv.org/abs/1810.02575. arXiv:1810.02575 [cs].

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs] version: 2.

Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased Metric Learning: On the Utilization of Multiple Datasets and Web Images for Softening Bias. In *2013 IEEE International Conference on Computer Vision*, pages 1657–1664, Sydney, Australia, December 2013. IEEE. ISBN 978-1-4799-2840-8. doi: 10.1109/ICCV.2013.208. URL http://ieeexplore.ieee.org/document/6751316/.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks, May 2016. URL http://arxiv.org/abs/1505.07818. arXiv:1505.07818 [cs, stat].

Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain Generalization for Object Recognition with Multi-task Autoencoders, August 2015. URL http://arxiv.org/abs/1508.07680. arXiv:1508.07680 [cs, stat].

Soumya Suvra Ghosal, Yifei Ming, and Yixuan Li. Are Vision Transformers Robust to Spurious Correlations?, March 2022. URL http://arxiv.org/abs/2203.09125. arXiv:2203.09125 [cs].

Ishaan Gulrajani and David Lopez-Paz. In Search of Lost Domain Generalization, July 2020. URL http://arxiv.org/abs/2007.01434. arXiv:2007.01434 [cs, stat].

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data, April 2018. URL http://arxiv.org/abs/1803.02324. arXiv:1803.02324 [cs].

M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. High-resolution global maps of 21st-century forest cover change. *Science (New York, N.Y.)*, 342(6160):850–853, November 2013. ISSN 1095-9203. doi: 10.1126/science.1244693.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant Causal Prediction for Nonlinear Models. *Journal of Causal Inference*, 6 (2):20170016, September 2018. ISSN 2193-3685, 2193-3677. doi: 10.1515/jci-2017-0016. URL https://www.degruyter.com/document/doi/10.1515/jci-2017-0016/html.

Sobhan Hemati, Guojun Zhang, Amir Estiri, and Xi Chen. Understanding Hessian Alignment for Domain Generalization. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18958–18968, Paris, France, October 2023. IEEE. ISBN 9798350307184. doi: 10.1109/ICCV51070.2023.01742. URL https://ieeexplore.ieee.org/document/10376731/.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. CyCADA: Cycle-Consistent Adversarial Domain Adaptation, December 2017. URL http://arxiv.org/abs/1711.03213. arXiv:1711.03213 [cs].

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does Distributionally Robust Supervised Learning Give Robust Classifiers?, July 2018. URL http://arxiv.org/abs/1611.02041. arXiv:1611.02041 [stat].

Yeping Hu, Xiaogang Jia, Masayoshi Tomizuka, and Wei Zhan. Causal-based Time Series Domain Generalization for Vehicle Intention Prediction, December 2021. URL http://arxiv.org/abs/2112.02093. arXiv:2112.02093 [cs, stat].

Zhuo Huang, Muyang Li, Li Shen, Jun Yu, Chen Gong, Bo Han, and Tongliang Liu. Winning Prize Comes from Losing Tickets: Improve Invariant Learning by Exploring Variant Parameters for Out-of-Distribution Generalization. *International Journal of Computer Vision*, 133(1):456–474, January 2025. ISSN 1573-1405. doi: 10.1007/s11263-024-02075-x. URL https://doi.org/10.1007/s11263-024-02075-x.

Pritish Kamath, Akilesh Tangella, Danica J. Sutherland, and Nathan Srebro. Does Invariant Risk Minimization Capture Invariance?, February 2021. URL http://arxiv.org/abs/2101.01134. arXiv:2101.01134 [cs, stat].

Masanori Koyama and Shoichiro Yamaguchi. Out-of-Distribution Generalization with Maximal Invariant Predictor. October 2020. URL https://openreview.net/forum?id=FzGiUKN4aBp.

Masanori Koyama and Shoichiro Yamaguchi. When is invariance useful in an Out-of-Distribution Generalization problem ?, November 2021. URL http://arxiv.org/abs/2008.01883. arXiv:2008.01883 [cs, stat].

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-Distribution Generalization via Risk Extrapolation (REx), February 2021. URL http://arxiv.org/abs/2003.00688. arXiv:2003.00688 [cs, stat].

Yann LeCun, Corinna Cortes, and Chris Burges. MNIST handwritten digit database, 2010. URL http://yann.lecun.com/exdb/mnist/.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, Broader and Artier Domain Generalization, October 2017. URL http://arxiv.org/abs/1710.03077. arXiv:1710.03077 [cs].

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain Generalization with Adversarial Feature Learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, Salt Lake City, UT, June 2018. IEEE. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00566. URL https://ieeexplore.ieee.org/document/8578664/.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild, September 2015. URL http://arxiv.org/abs/1411.7766. arXiv:1411.7766 [cs] version: 3.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning Transferable Features with Deep Adaptation Networks, May 2015. URL http://arxiv.org/abs/1502.02791. arXiv:1502.02791 [cs].

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain Generalization via Invariant Feature Representation, January 2013. URL http://arxiv.org/abs/1301.2115. arXiv:1301.2115 [cs, stat].

Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary, October 2020. URL http://arxiv.org/abs/2009.00329. arXiv:2009.00329 [cs, stat].

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment Matching for Multi-Source Domain Adaptation, August 2019. URL http://arxiv.org/abs/1812.01754. arXiv:1812.01754 [cs].

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals, November 2015. URL http://arxiv.org/abs/1501.01332. arXiv:1501.01332 [stat].

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant Gradient Variances for Out-of-Distribution Generalization. In *Proceedings of the 39th International Conference on Machine Learning*, pages 18347–18377. PMLR, June 2022. URL https://proceedings.mlr.press/v162/rame22a.html. ISSN: 2640-3498.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The Risks of Invariant Risk Minimization, March 2021. URL http://arxiv.org/abs/2010.05761. arXiv:2010.05761 [cs, stat].

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization, April 2020. URL http://arxiv.org/abs/1911.08731. arXiv:1911.08731 [cs, stat].

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. BREEDS: Benchmarks for Subpopulation Shift, August 2020. URL http://arxiv.org/abs/2008.04859. arXiv:2008.04859 [cs, stat].

Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do Image Classifiers Generalize Across Time?, December 2019. URL http://arxiv.org/abs/1906.02168. arXiv:1906.02168 [cs, stat].

Yuge Shi, Jeffrey Seely, Philip H. S. Torr, N. Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient Matching for Domain Generalization, July 2021. URL http://arxiv.org/abs/2104.09937. arXiv:2104.09937 [cs, stat].

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers, June 2022. URL http://arxiv.org/abs/2106.10270. arXiv:2106.10270 [cs].

Maryam Sultana, Muzammal Naseer, Muhammad Haris Khan, Salman Khan, and Fahad Shahbaz Khan. Self-Distilled Vision Transformer for Domain Generalization. pages 3068–3085, 2022. URL https://openaccess.thecvf.com/content/ACCV2022/html/Sultana_Self-Distilled_Vision_Transformer_for_Domain_Generalization_ACCV_2022_paper.html.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation Alignment for Deep Domain Adaptation, July 2016. URL http://arxiv.org/abs/1607.01719. arXiv:1607.01719 [cs].

David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, December 2019. ISSN 1361-8415. doi: 10.1016/j.media.2019.101544. URL https://www.sciencedirect.com/science/article/pii/S1361841519300799.

Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton van den Hengel. Evading the Simplicity Bias: Training a Diverse Set of Models Discovers Solutions with Superior OOD Generalization, September 2022. URL http://arxiv.org/abs/2105.05612. arXiv:2105.05612 [cs].

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial Discriminative Domain Adaptation, February 2017. URL http://arxiv.org/abs/1702.05464. arXiv:1702.05464 [cs].

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10 (5):988–999, September 1999. ISSN 1941-0093. doi: 10.1109/72.788640. URL https://ieeexplore.ieee.org/document/788640. Conference Name: IEEE Transactions on Neural Networks.

Christian Wachinger, Anna Rieckmann, Sebastian Pölsterl, and Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing. Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis*, 67:101879, January 2021. ISSN 1361-8423. doi: 10.1016/j.media.2020.101879.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical report, California Institute of Technology, 2011. URL https://paperswithcode.com/dataset/cub-200-2011.

Haoxiang Wang, Haozhe Si, Bo Li, and Han Zhao. Provable Domain Generalization via Invariant-Feature Subspace

Recovery, July 2022. URL http://arxiv.org/abs/2201.12919. arXiv:2201.12919 [cs, stat].

Haoxiang Wang, Gargi Balasubramaniam, Haozhe Si, Bo Li, and Han Zhao. Invariant-Feature Subspace Recovery: A New Class of Provable Domain Generalization Algorithms, November 2023. URL http://arxiv.org/abs/2311.00966. arXiv:2311.00966 [cs, stat].

Ross Wightman. PyTorch Image Models, 2019. URL https://github.com/google-research/vision_transformer. original-date: 2020-10-21T12:35:02Z.

Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central Moment Discrepancy (CMD) for Domain-Invariant Representation Learning, May 2019. URL http://arxiv.org/abs/1702.08811. arXiv:1702.08811 [cs, stat].

Guojun Zhang, Han Zhao, Yaoliang Yu, and Pascal Poupart. Quantifying and Improving Transferability in Domain Generalization, November 2021. URL http://arxiv.org/abs/2106.03632. arXiv:2106.03632 [cs, stat].

Nevin L. Zhang, Kaican Li, Han Gao, Weiyan Xie, Zhi Lin, Zhenguo Li, Luning Wang, and Yongxiang Huang. A Causal Framework to Unify Common Domain Generalization Approaches, July 2023. URL http://arxiv.org/abs/2307.06825. arXiv:2307.06825 [cs].

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On Learning Invariant Representations for Domain Adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7523–7532. PMLR, May 2019. URL https://proceedings.mlr.press/v97/zhao19a.html. ISSN: 2640-3498.

Zangwei Zheng, Xiangyu Yue, Kai Wang, and Yang You. Prompt Vision Transformer for Domain Generalization, August 2022. URL http://arxiv.org/abs/2208.08914. arXiv:2208.08914 [cs].

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, June 2018. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2017.2723009. URL https://ieeexplore.ieee.org/document/7968387/.

Moment Alignment: Unifying Gradient and Hessian Matching for Domain Generalization (Supplementary Material)

Yuen Chen[1] **Haozhe Si**[1] **Guojun Zhang**[2] **Han Zhao**[1]

[1]University of Illinois at Urbana-Champaign
[2]University of Waterloo
{yuenc2, haozhes3, hanzhao}@illinois.edu, g39zhang@uwaterloo.ca

A PROOF OF PROPOSITIONS

Proposition 2 (upper bound on transfer measure). *Given $\mathcal{S} = \{\mu_i\}_{i=1}^K$ and some $\Gamma \subseteq \mathcal{H}$. Define $\mathcal{L}^*_{\mu_i} := \inf_{h\in\Gamma} \mathcal{L}_{\mu_i}(h)$ for all $i \in [K]$, $\mathcal{L}_{\mathcal{T}}{}^* := \inf_{h\in\Gamma} \mathcal{L}_{\mathcal{T}}(h)$, $\mu^* := \arg\min_\mu \max_{i\in[K]} \mathrm{T}_\Gamma(\mu_i\|\mu)$, and $\mathcal{L}_{\mathcal{S}}(h) := \mathcal{L}_{\mu^*}(h)$. Under Assumption 2, we have:*

$$\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \le \frac{1}{2}\max_{i\neq j} \mathrm{T}_\Gamma(\mu_j\|\mu_i) \tag{3}$$

Proof.

$$\begin{aligned}
\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) &= \mathrm{T}_\Gamma(\mu^*\|\mathcal{T}) \\
&= \mathrm{T}_\Gamma\left(\mu^*\middle\|\sum_{i=1}^K w_i\mu_i\right) \\
&= \sup_{h\in\Gamma}\sum_{i=1}^K w_i\mathcal{L}_{\mu_i}(h) - \mathcal{L}^*_{\mathcal{T}} - \left(\mathcal{L}_{\mu^*}(h) - \mathcal{L}^*_{\mu^*}\right) \\
&\le \sup_{h\in\Gamma}\sum_{i=1}^K w_i\left[\mathcal{L}_{\mu_i}(h) - \mathcal{L}^*_{\mathcal{T}} - \left(\mathcal{L}_{\mu^*}(h) - \mathcal{L}^*_{\mu^*}\right)\right] \\
&\le \sup_{h\in\Gamma}\sum_{i=1}^K w_i\left[\mathcal{L}_{\mu_i}(h) - \mathcal{L}^*_{\mu_i} - \left(\mathcal{L}_{\mu^*}(h) - \mathcal{L}^*_{\mu^*}\right)\right] \\
&\le \sum_{i=1}^K w_i \sup_{h\in\Gamma}\left[\mathcal{L}_{\mu_i}(h) - \mathcal{L}_{\mu_i^*} - \left(\mathcal{L}_{\mu^*}(h) - \mathcal{L}^*_{\mu^*}\right)\right] \\
&= \sum_{i=1}^K w_i\mathrm{T}_\Gamma(\mu^*\|\mu_i) \\
&\le \max_{i\in[k]}\mathrm{T}_\Gamma(\mu^*\|\mu_i)
\end{aligned} \tag{12}$$

On the other hand, let $j_{max} := \arg\max_{j\in[K]} \mathrm{T}_\Gamma(\mu_j\|\mu^*)$ and for any fixed i, let μ_{mid} be such that $\mathrm{T}_\Gamma(\mu_{j_{max}}\|\mu_i) = \mathrm{T}_\Gamma(\mu_{j_{max}}\|\mu_{mid}) + \mathrm{T}_\Gamma(\mu_{mid}\|\mu_i)$. By the definition μ^* we have:

$$\begin{aligned}
\frac{1}{2}\max_{i\neq j}\mathrm{T}_\Gamma(\mu_j\|\mu_i) \ge& \frac{1}{2}\mathrm{T}_\Gamma(\mu_{j_{max}}\|\mu_i) \quad \forall i\in[K] \\
=& \mathrm{T}_\Gamma(\mu_{mid}\|\mu_i) \quad \forall i\in[K], \quad \exists\mu_{mid} \\
\ge& \mathrm{T}_\Gamma(\mu^*\|\mu_i) \quad \forall i\in[K]
\end{aligned} \tag{13}$$

Combine Eq. (12) and Eq. (13), we have:

$$\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq \max_{i\in[k]} \mathrm{T}_\Gamma\left(\mu^*\|\mu_i\right) \leq \frac{1}{2}\max_{i\neq j} \mathrm{T}_\Gamma\left(\mu_j\|\mu_i\right) \tag{14}$$

□

Proposition 3 (**target error bound – multiple source domains**). *Given $\Gamma \subseteq \mathcal{H}$, for any $h \in \Gamma$, the target error is bounded by:*

$$\mathcal{L}_\mathcal{T}(h) \leq \mathcal{L}_\mathcal{S}(h) + \mathcal{L}^*_\mathcal{T} - \mathcal{L}^*_\mathcal{S} + \frac{1}{2}\max_{i\neq j} \mathrm{T}_\Gamma\left(\mu_j\|\mu_i\right) \tag{4}$$

Proof. Apply Proposition 1 on μ^* and $\mathcal{T}$, we have:

$$\begin{aligned}
\mathcal{L}_\mathcal{T}(h) \leq& \mathcal{L}_{\mu^*}(h) + \mathcal{L}^*_\mathcal{T} - \mathcal{L}^*_{\mu^*} + \mathrm{T}_\Gamma(\mu^*\|\mathcal{T}) \\
=& \mathcal{L}_\mathcal{S}(h) + \mathcal{L}^*_\mathcal{T} - \mathcal{L}^*_\mathcal{S} + \mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \\
\leq& \mathcal{L}_\mathcal{S}(h) + \mathcal{L}^*_\mathcal{T} - \mathcal{L}^*_\mathcal{S} + \frac{1}{2}\max_{i\neq j} \mathrm{T}_\Gamma\left(\mu_j\|\mu_i\right)
\end{aligned} \tag{15}$$

The second line follows from the definition of μ^* and the last line follows from Proposition 2. □

B PROOF OF THEOREMS

Theorem 1 (**moment alignment under IRM**). *Given K source domains $\mathcal{S} = \{\mu_i\}_{i=1}^K$ and a target domain $\mathcal{T} \in conv(\mu_1, \ldots, \mu_k)$, assume the losses $\mathcal{L}_{\mu_i}\ \forall i \in [K]$ are ν-strongly convex w.r.t. the classifier head and M-times differentiable. Under the IRM assumption (Assumption 1), let $\boldsymbol{\theta}^*$ be the optimal invariant predictor, $\Gamma = \arg\min(\mathcal{L}_\mathcal{S}, \delta_\mathcal{S})_\mathcal{H} := \left\{\boldsymbol{\theta} \mid h_{\boldsymbol{\theta}\in\mathcal{H}} : \max_{i\in[K]}\left(\mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}\right) - \mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}^*\right)\right) \leq \delta_\mathcal{S}\right\}$, and $\delta = \frac{2\delta_\mathcal{S}}{\nu}$, we have:*

$$\begin{aligned}
\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq \max_{i\neq j}\Bigg(&\sum_{n=2}^{N}\frac{1}{n!}\delta^{\frac{n}{2}}\|\nabla^n_{\boldsymbol{\theta}}\mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}^*\right) \\
&- \nabla^n_{\boldsymbol{\theta}}\mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}^*\right)\|^n_F\Bigg) + o(\delta^{\frac{N}{2}})
\end{aligned} \tag{5}$$

for any integer $2 \leq N \leq M$. $\nabla^{\mathbf{n}}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$ is an n^{th} order tensor with dimension $d\times\cdots\times d$ (n times) where $\nabla^{\mathbf{n}}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})_{(k_1,\cdots,k_n)} = \frac{\partial^n\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_{k_1}\ldots\partial\boldsymbol{\theta}_{k_n}}$.

Proof. From the ν-strong convexity of $\mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}\right)$, we can write for all i

$$\begin{aligned}
&\mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}\right) \geq \mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}^*\right) + \underbrace{\left(\boldsymbol{\theta}-\boldsymbol{\theta}^*\right)^\top\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}^*\right)}_{=0} + \frac{\nu}{2}\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\right\|_2^2 \\
\implies &\mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}\right) - \mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}^*\right) \geq \frac{\nu}{2}\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\right\|_2^2
\end{aligned} \tag{16}$$

So for any $\boldsymbol{\theta} \in \Gamma$

$$\frac{\nu}{2}\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\right\|_2^2 \leq \max_{i\in[n]}\left(\mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}\right) - \mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}^*\right)\right) \leq \delta_\mathcal{S} \tag{17}$$

If we define set $\mathcal{F}_2$ as

$$\mathcal{F}_2 = \left\{\boldsymbol{\theta} : \frac{\nu}{2}\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\right\|_2^2 \leq \delta_\mathcal{S}\right\} \tag{18}$$

then Eq. (17) implies $\Gamma \subset \mathcal{F}_2$.

From Proposition 2, we have

$$\begin{aligned}
\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq& \frac{1}{2}\max_{i\neq j}\mathrm{T}_\Gamma\left(\mu_j\|\mu_i\right) \\
=& \frac{1}{2}\max_{i\neq j}\sup_{\boldsymbol{\theta}\in\Gamma}\left(\mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}\right) - \mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}^*\right) - \left(\mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}\right) - \mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}^*\right)\right)\right) \\
=& \frac{1}{2}\max_{i\neq j}\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2\leq\delta}\left(\mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}\right) - \mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}^*\right) - \left(\mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}\right) - \mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}^*\right)\right)\right)
\end{aligned} \tag{19}$$

To bound the terms inside the supremum, which is the difference in excess risk of μ_j and μ_i, we write the Taylor expansion of $\mathcal{L}_{\mu_i}(\boldsymbol{\theta})$ around $\boldsymbol{\theta}^*$:

$$\begin{aligned}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) &= \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) + \underbrace{(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)}_{=0} + \sum_{n=2}^{N} \frac{1}{n!} \nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_{\mathtt{i}}}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^{\otimes \mathtt{n}} + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^N) \\ \Longrightarrow \mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) &= \sum_{n=2}^{N} \frac{1}{n!} \nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_{\mathtt{i}}}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^{\otimes \mathtt{n}} + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^N)\end{aligned} \tag{20}$$

Here $\otimes n$ denote the n^{th}-order tensor product, where $(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^{\otimes \mathtt{n}}{}_{(k_1,\cdots,k_n)}$ is the product of $\left(\boldsymbol{\theta}_{k_1}-\boldsymbol{\theta}^*_{k_1}\right),\ldots,\left(\boldsymbol{\theta}_{k_n}-\boldsymbol{\theta}^*_{k_n}\right)$.

Similarly, expanding $\mathcal{L}_{\mu_j}(\boldsymbol{\theta})$ around $\boldsymbol{\theta}^*$, we have:

$$\mathcal{L}_{\mu_j}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) = \sum_{n=2}^{N} \frac{1}{n!} \nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_{\mathtt{j}}}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^{\otimes \mathtt{n}} + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^N) \tag{21}$$

These two equations together give an upper bound on the difference in excess risk of domain j and i:

$$\begin{aligned}&\mathcal{L}_{\mu_j}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - (\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)) \\ =&\sum_{n=2}^{N} \frac{1}{n!} \left(\nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_{\mathtt{j}}}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_{\mathtt{i}}}(\boldsymbol{\theta}^*)\right)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^{\otimes \mathtt{n}} + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^N) \\ \leq&\sum_{n=2}^{N} \frac{1}{n!} \left\|\nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_{\mathtt{j}}}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_{\mathtt{i}}}(\boldsymbol{\theta}^*)\right\|_F \left\|(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^{\otimes \mathtt{n}}\right\|_F + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^N) \\ \leq&\sum_{n=2}^{N} \frac{1}{n!} \left\|\nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_{\mathtt{j}}}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_{\mathtt{i}}}(\boldsymbol{\theta}^*)\right\|_F \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^n + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^N)\end{aligned} \tag{22}$$

Taking the supremum over $\boldsymbol{\theta} \in \mathcal{F}_2$ on both sides, for any $i \neq j$, we have:

$$\begin{aligned}&\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2 \leq \delta} \left(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - (\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*))\right) \\ \leq&\sum_{n=2}^{N} \frac{1}{n!} \delta^{\frac{n}{2}} \left\|\nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_{\mathtt{j}}}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_{\mathtt{i}}}(\boldsymbol{\theta}^*)\right\|_F + o(\delta^{\frac{N}{2}})\end{aligned} \tag{23}$$

Finally, by taking the maximum over i and j, $i \neq j$, on both sides, we can bound the transfer measure as follows:

$$\begin{aligned}\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq&\frac{1}{2} \max_{i \neq j} \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2 \leq \delta} \left(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - (\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*))\right) \\ \leq&\max_{i \neq j} \sum_{n=2}^{N} \frac{1}{n!} \delta^{\frac{n}{2}} \left\|\nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_{\mathtt{j}}}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_{\mathtt{i}}}(\boldsymbol{\theta}^*)\right\|_F + o(\delta^{\frac{N}{2}})\end{aligned} \tag{24}$$

□

Theorem 3 (moment alignment). *Given K source domains $\mathcal{S} = \{\mu_i\}_{i=1}^K$ and target domain $\mathcal{T} \in conv(\mu_1,\ldots,\mu_k)$. Assume loss $\mathcal{L}_{\mu_i}$, $\forall i \in [K]$ are ν-strongly convex and M-times differentiable w.r.t. the classifier head. Let $\mathcal{P}(\{\mathcal{L}_{\mu_i}\}_{i=1}^K) := \{\boldsymbol{\theta} : \max_{i\in[K]}(\boldsymbol{\theta'}-\boldsymbol{\theta})^\top \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i}(\boldsymbol{\theta}) \geq 0,\ \forall \boldsymbol{\theta'} \in \Gamma\}$ (a set of weakly Pareto optimal points for the objectives $\{\mathcal{L}_{\mu_i}\}_{i=1}^K$), and let $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\mathcal{P}} \max_{i\in[K]} \|\nabla \mathcal{L}_{\mu_i}(\boldsymbol{\theta})\|_2$, $\Gamma := \left\{\boldsymbol{\theta} \in \mathcal{H} : \max_{i\in[K]} \left(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)\right) \leq \delta_{\mathcal{S}}\right\}$, and $\delta = \frac{2\delta_{\mathcal{S}}}{\nu}$, we have:*

$$\begin{aligned}\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq&\frac{1}{2} \max_{i \neq j} \Bigg(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*_j) \\ &- (\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*_i)) \\ &+ \sum_{n=1}^{N} \frac{1}{n!} \delta^{\frac{n}{2}} \left\|\nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_{\mathtt{j}}}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_{\mathtt{i}}}(\boldsymbol{\theta}^*)\right\|_F \Bigg) \\ &+ o(\delta^{\frac{N}{2}})\end{aligned} \tag{7}$$

where $\boldsymbol{\theta}_i^*$ is the minimizer of $\mathcal{L}_{\mu_i}(\boldsymbol{\theta})$. Furthermore, suppose Assumption 3 holds with $g > 0$:

$$\begin{aligned}\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq & \delta^{\frac{1}{2}} g + \frac{1}{2}\max_{i\neq j}\Bigg(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}_j^*\right) \\ & - \left(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*)\right) \\ & + \sum_{n=2}^{N}\frac{1}{n!}\delta^{\frac{n}{2}}\left\|\nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)\right\|_F\Bigg) \\ & + o(\delta^{\frac{N}{2}})\end{aligned} \tag{8}$$

Proof. From the ν-strong convexity of $\mathcal{L}_{\mu_i}(\boldsymbol{\theta})$, we can write for all i

$$\begin{aligned}& \mathcal{L}_{\mu_i}(\boldsymbol{\theta}) \geq \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) + (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) + \frac{\nu}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2 \\ \Longrightarrow\ & \delta_{\mathcal{S}} \geq \max_{i\in[K]}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) \geq \underbrace{\max_{i\in[K]}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)}_{\geq 0 \text{ since } \boldsymbol{\theta}^* \text{ is weak Pareto Optimal}} + \frac{\nu}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2 \geq \frac{\nu}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2\end{aligned} \tag{25}$$

Now define $\mathcal{F}_2$ as

$$\mathcal{F}_2 = \left\{\boldsymbol{\theta} : \max_{i\in[K]}\frac{\nu}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2 \leq \delta_{\mathcal{S}}\right\} = \left\{\boldsymbol{\theta} : \max_{i\in[K]}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2 \leq \frac{2\delta_{\mathcal{S}}}{\nu} = \delta\right\} \tag{26}$$

From Eq. (25), we have $\Gamma \subseteq \mathcal{F}_2$, and thus $\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq \mathrm{T}_{\mathcal{F}_2}(\mathcal{S}\|\mathcal{T})$.

$$\begin{aligned}\mathrm{T}_{\mathcal{F}_2}(\mathcal{S}\|\mathcal{T}) \leq & \frac{1}{2}\max_{i\neq j}\mathrm{T}_{\mathcal{F}_2}(\mu_j\|\mu_i) \\ = & \frac{1}{2}\max_{i\neq j}\sup_{\boldsymbol{\theta}\in\mathcal{F}_2}\left(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}_j^*\right) - \left(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*)\right)\right)\end{aligned} \tag{27}$$

We write the Taylor expansion of $\mathcal{L}_{\mu_i}(\boldsymbol{\theta})$ and $\mathcal{L}_{\mu_j}(\boldsymbol{\theta})$ around $\boldsymbol{\theta}^*$ as done in the proof of Theorem 1:

$$\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) = \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) + (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) + \sum_{n=2}^{N}\frac{1}{n!}\nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^{\otimes\mathtt{n}} + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^N)$$

$$\mathcal{L}_{\mu_j}(\boldsymbol{\theta}) = \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) + (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) + \sum_{n=2}^{N}\frac{1}{n!}\nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^{\otimes\mathtt{n}} + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^N)$$

Combining the two equations above, we have:

$$\begin{aligned}& \mathcal{L}_{\mu_j}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}_j^*\right) - \left(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*)\right) \\ \leq & \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) + (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) + \sum_{n=2}^{N}\frac{1}{n!}\nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^{\otimes\mathtt{n}} - \mathcal{L}_{\mu_i}\left(\boldsymbol{\theta}_j^*\right) \\ & - \left(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) + (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) + \sum_{n=2}^{N}\frac{1}{n!}\nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^{\otimes\mathtt{n}} - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*)\right) + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^N) \\ \leq & \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}_j^*\right) - \left(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*)\right) \\ & + \sum_{n=1}^{N}\frac{1}{n!}\left\|\nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)\right\|_F\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^n + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^N)\end{aligned}$$

Taking the supremum over $\boldsymbol{\theta} \in \mathcal{F}_2$ on both sides, for any $i \neq j$, we have:

$$\begin{aligned}& \sup_{\boldsymbol{\theta}\in\mathcal{F}_2}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}_j^*\right) - \left(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*)\right) + \sum_{n=1}^{N}\frac{1}{n!}\left\|\nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)\right\|_F\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^n + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^N) \\ \leq & \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}_j^*\right) + \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*) + \sum_{n=1}^{N}\frac{1}{n!}\delta^{\frac{n}{2}}\left\|\nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}^{\mathtt{n}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)\right\|_F + o(\delta^{\frac{N}{2}})\end{aligned}$$

Finally, maximizing over $i \neq j$ on both sides, the transfer measure is bounded by:

$$\begin{aligned}
\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq & \frac{1}{2} \max_{i \neq j} \sup_{\boldsymbol{\theta} \in \mathcal{F}_2} \left(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - (\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)) \right) \\
\leq & \frac{1}{2} \max_{i \neq j} \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}_j^*\right) + \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*) \\
& + \sum_{n=2}^{N} \frac{1}{n!} \delta^{\frac{n}{2}} \left\| \nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_\mathtt{j}}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}^{\mathtt{n}} \mathcal{L}_{\mu_\mathtt{i}}(\boldsymbol{\theta}^*) \right\|_F + o(\delta^{\frac{N}{2}})
\end{aligned} \tag{28}$$

We have proved the first part of the theorem. Now suppose there exists a constant g such that $\min_{\boldsymbol{\theta} \in \mathcal{H}} \max_{i \in [K]} \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i}(\boldsymbol{\theta})\|_2 \leq g$, we can further upper bound the first-order term ($n = 1$) by g:

$$\begin{aligned}
\delta^{\frac{1}{2}} \left\| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) \right\|_F = & \delta^{\frac{1}{2}} \left\| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) \right\|_2 \\
\leq & \delta^{\frac{1}{2}} \left\| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) \right\|_2 + \left\| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) \right\|_2 \\
\leq & 2\delta^{\frac{1}{2}} g
\end{aligned} \tag{29}$$

Replacing the first-order term in Eq. (28) with this upper bound completes the proof. □

Definition 4 (weakly Pareto optimal (Chang, 2015))**.** *A point $\boldsymbol{\theta} \in \Gamma$ is weakly Pareto optimal iff $\nexists$ another point $\boldsymbol{\theta}' \in \Gamma$ such that $\mathcal{L}_{\mu_i}(\boldsymbol{\theta}') < \mathcal{L}_{\mu_i}(\boldsymbol{\theta}) \quad \forall i$.*

Lemma 1. *For convex $\{\mathcal{L}_{\mu_i}\}_{i=1}^K$, $\boldsymbol{\theta} \in \mathcal{P}(\{\mathcal{L}_{\mu_i}\}_{i=1}^K)$ iff $\boldsymbol{\theta}$ is weakly Pareto optimal.*

Proof. ($\Longrightarrow$) Let $\boldsymbol{\theta} \in \mathcal{P}(\{\mathcal{L}_{\mu_i}\}_{i=1}^K)$, by convexity, we have for all $\boldsymbol{\theta}' \in \Gamma$ and $i \in [K]$,

$$\mathcal{L}_{\mu_i}(\boldsymbol{\theta}') \geq \mathcal{L}_{\mu_i}(\boldsymbol{\theta}) + (\boldsymbol{\theta}' - \boldsymbol{\theta})^\top \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i}(\boldsymbol{\theta}) \geq \mathcal{L}_{\mu_i}(\boldsymbol{\theta}),$$

where the gradient term is non-negative by $\boldsymbol{\theta} \in \mathcal{P}(\{\mathcal{L}_{\mu_i}\}_{i=1}^K)$. Thus, $\boldsymbol{\theta}$ is weakly Pareto optimal as for all $\boldsymbol{\theta}' \in \Gamma$ there is some $i \in [K]$ such that $\mathcal{L}_{\mu_i}(\boldsymbol{\theta}') \geq \mathcal{L}_{\mu_i}(\boldsymbol{\theta})$.
($\Longleftarrow$) Suppose for contradiction that $\boldsymbol{\theta} \notin \mathcal{P}(\{\mathcal{L}_{\mu_i}\}_{i=1}^K)$, Then there exists some $\boldsymbol{\theta}'$ such that $(\boldsymbol{\theta}' - \boldsymbol{\theta})^\top \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i}(\boldsymbol{\theta}) < 0$ for all $i \in [K]$. Using the Taylor expansion,

$$\mathcal{L}_{\mu_i}(\boldsymbol{\theta}') = \mathcal{L}_{\mu_i}(\boldsymbol{\theta}) + (\boldsymbol{\theta}' - \boldsymbol{\theta})^\top \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i}(\boldsymbol{\theta}) + \mathcal{O}\left(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2^2\right).$$

Choosing a scaled step $\boldsymbol{\theta}^* = \boldsymbol{\theta} + \delta(\boldsymbol{\theta}' - \boldsymbol{\theta})$ to still satisfy $(\boldsymbol{\theta}^* - \boldsymbol{\theta})^\top \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i}(\boldsymbol{\theta}) < 0$. For small enough $\delta \to 0$, $\mathcal{O}\left(\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2^2\right)$ becomes negligible and

$$\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) \to \mathcal{L}_{\mu_i}(\boldsymbol{\theta}) + (\boldsymbol{\theta}^* - \boldsymbol{\theta})^\top \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i}(\boldsymbol{\theta}) < \mathcal{L}_{\mu_i}(\boldsymbol{\theta}),$$

and $\boldsymbol{\theta}$ is not weakly Pareto optimal. By contrapositive, we have shown that if $\boldsymbol{\theta}$ is weakly Pareto optimal, $\boldsymbol{\theta} \in \mathcal{P}(\{\mathcal{L}_{\mu_i}\}_{i=1}^K)$. □

Note that the convexity assumption is necessary, as we can construct a simple one-dimensional counterexample where $\theta \in \mathcal{P}(\{\mathcal{L}_{\mu_i}\}_{i=1}^K)$ but not weakly Pareto optimal. We consider two functions $\mathcal{L}_1(\theta)$ and $\mathcal{L}_2(\theta)$ given by:

$$\mathcal{L}_1(\theta) = \sin\theta, \ \mathcal{L}_2(\theta) = \theta^3, \ \Gamma = \mathbb{R}.$$

At $\theta = 0$, we compute the gradients:

$$\begin{aligned}
\mathcal{L}_1'(0) &= \cos(0) = 1, \\
\mathcal{L}_2'(0) &= 0.
\end{aligned}$$

Thus, for any $\theta' \in \mathbb{R}$:

$$\begin{aligned}
(\theta' - 0)\mathcal{L}_1'(0) &= \theta', \\
(\theta' - 0)\mathcal{L}_2'(0) &= 0.
\end{aligned}$$

Taking the maximum, we have:

$$\max\{\theta', 0\} \geq 0, \quad \forall \theta'.$$

therefore $\theta = 0 \in \mathcal{P}(\mathcal{L}_1, \mathcal{L}_2)$.

However, $\theta = 0$ is not weakly Pareto optimal. Consider $\theta^* = -0.5$. The function values at θ^* are:

$$\mathcal{L}_1(\theta^*) = \sin(-0.5) = -\frac{\sqrt{2}}{2} < \mathcal{L}_1(0) = 0,$$
$$\mathcal{L}_2(\theta^*) = (-0.5)^3 = -0.125 < \mathcal{L}_2(0) = 0.$$

Since there exists θ^* such that both $\mathcal{L}_i(\theta^*) < \mathcal{L}_i(\theta)$ for all i, $\theta = 0$ is not weakly Pareto optimal.

C PROOF OF COROLLARIES

Corollary 2 (hessian alignment under IRM). *Under the same setup as in Theorem 1, we have:*

$$\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq \frac{1}{2}\delta \max_{i \neq j} \|\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*) - \mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\|_F + o(\delta) \tag{6}$$

where $\mathbf{H}(\boldsymbol{\theta})$ *denotes the Hessian matrix of* $\mathcal{L}(\boldsymbol{\theta})$.

Proof. We use the same $\mathcal{F}_2$ as in the proof of Theorem 1 and have $\Gamma \subset \mathcal{F}_2$.

From Proposition 2, we have

$$\begin{aligned}
\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq & \frac{1}{2}\max_{i \neq j} \mathrm{T}_\Gamma(\mu_j\|\mu_i) \\
= & \frac{1}{2}\max_{i \neq j} \sup_{\boldsymbol{\theta} \in \Gamma} \left(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - (\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*))\right) \\
= & \frac{1}{2}\max_{i \neq j} \sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \leq \delta} \left(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - (\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*))\right)
\end{aligned} \tag{30}$$

To bound the terms inside the supremum, which is the difference in excess risk of μ_j and μ_i, we write the Taylor expansion of $\mathcal{L}_{\mu_i}(\boldsymbol{\theta})$ around $\boldsymbol{\theta}^*$:

$$\begin{aligned}
\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) &= \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) + \underbrace{(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)}_{=0} + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + o(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2) \\
\implies \mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) &= \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + o(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2)
\end{aligned} \tag{31}$$

Similarly, expand $\mathcal{L}_{\mu_j}(\boldsymbol{\theta})$ around $\boldsymbol{\theta}^*$, we have:

$$\mathcal{L}_{\mu_j}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + o(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2) \tag{32}$$

These two equations together give an upper bound on the difference in excess risk of domain j and i:

$$\begin{aligned}
& \mathcal{L}_{\mu_j}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - (\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)) \\
= & \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + o(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2) \\
= & \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \left(\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*) - \mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\right)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + o(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2) \\
\leq & \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \|\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*) - \mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\|_2 + o(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2)
\end{aligned} \tag{33}$$

Taking the supremum over $\boldsymbol{\theta} \in \mathcal{F}_2$ on both sides, for any $i \neq j$, we have:

$$\sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \leq \delta} \left(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - (\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*))\right) \leq \frac{1}{2}\delta \|\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*) - \mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\|_2 + o(\delta)$$

Finally, by taking the maximum over i and j, $i \neq j$, on both sides, we can bound the transfer measure as follows:

$$\begin{aligned}\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq & \frac{1}{2}\max_{i\neq j}\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2\leq\delta}\left(\mathcal{L}_{\mu_j}(\boldsymbol{\theta})-\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*)-(\mathcal{L}_{\mu_i}(\boldsymbol{\theta})-\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*))\right)\\ \leq & \frac{1}{2}\delta\max_{i\neq j}\|\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*)-\mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\|_2+o(\delta)\end{aligned} \tag{34}$$

□

Corollary 4 (hessian alignment). *Under the same setup as in Theorem 3, we have:*

$$\begin{aligned}\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq & \frac{1}{2}\max_{i\neq j}\Bigg(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*)-\mathcal{L}_{\mu_j}(\boldsymbol{\theta}_j^*)\\ & -(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)-\mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*))\\ & +\delta^{\frac{1}{2}}\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*)-\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*)\right\|_2\\ & +\frac{1}{2}\delta\left\|\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*)-\mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\right\|_F\Bigg)+o(\delta)\end{aligned} \tag{9}$$

where $\boldsymbol{\theta}_i^*$ is the minimizer of $\mathcal{L}_{\mu_i}(\boldsymbol{\theta})$.

Furthermore, suppose Assumption 3 holds with $g > 0$:

$$\begin{aligned}\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq & \delta^{\frac{1}{2}}g+\frac{1}{2}\max_{i\neq j}\Bigg(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*)-\mathcal{L}_{\mu_j}(\boldsymbol{\theta}_j^*)\\ & -(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)-\mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*))\\ & +\frac{1}{2}\delta\left\|\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*)-\mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\right\|_F\Bigg)+o(\delta)\end{aligned} \tag{10}$$

Proof. Use the same $\mathcal{F}_2$ as in the proof of Theorem 3, we have $\Gamma \subseteq \mathcal{F}_2$, and $\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq \mathrm{T}_{\mathcal{F}_2}(\mathcal{S}\|\mathcal{T})$.

$$\begin{aligned}\mathrm{T}_{\mathcal{F}_2}(\mathcal{S}\|\mathcal{T}) \leq & \frac{1}{2}\max_{i\neq j}\mathrm{T}_{\mathcal{F}_2}(\mu_j\|\mu_i)\\ = & \frac{1}{2}\max_{i\neq j}\sup_{\boldsymbol{\theta}\in\mathcal{F}_2}\left(\mathcal{L}_{\mu_j}(\boldsymbol{\theta})-\mathcal{L}_{\mu_j}(\boldsymbol{\theta}_j^*)-(\mathcal{L}_{\mu_i}(\boldsymbol{\theta})-\mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*))\right)\end{aligned} \tag{35}$$

We write the Taylor expansion of $\mathcal{L}_{\mu_i}(\boldsymbol{\theta})$ and $\mathcal{L}_{\mu_j}(\boldsymbol{\theta})$ around $\boldsymbol{\theta}^*$:

$$\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) = \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)+(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)+\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\mathbf{H}_{\mu_i}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)+o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2) \tag{36}$$

$$\mathcal{L}_{\mu_j}(\boldsymbol{\theta}) = \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*)+(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*)+\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\mathbf{H}_{\mu_j}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)+o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2) \tag{37}$$

Combining the two equations above, we have:

$$\begin{aligned}& \mathcal{L}_{\mu_j}(\boldsymbol{\theta})-\mathcal{L}_{\mu_j}(\boldsymbol{\theta}_j^*)-(\mathcal{L}_{\mu_i}(\boldsymbol{\theta})-\mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*))\\ = & \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*)-\mathcal{L}_{\mu_j}(\boldsymbol{\theta}_j^*)+\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)-\mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*)+(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\left(\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*)-\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)\right)\\ & +\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\left(\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*)-\mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\right)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)+o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2)\\ \leq & \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*)-\mathcal{L}_{\mu_j}(\boldsymbol{\theta}_j^*)+\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)-\mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*)+\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*)-\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)\|_2\\ & +\frac{1}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2\left(\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*)-\mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\right)+o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2)\end{aligned}$$

Taking the supremum over $\boldsymbol{\theta} \in \mathcal{F}_2$ on both sides, for any $i \neq j$, we have:

$$
\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\mathcal{F}_2} \mathcal{L}_{\mu_j}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}_j^*\right) - \left(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*)\right) \\
=&\sup_{\boldsymbol{\theta}\in\mathcal{F}_2} \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}_j^*\right) + \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*) + \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2 \|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)\|_2 \\
&\quad + \frac{1}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2 \|\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*) - \mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\|_2 + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2) \\
\leq&\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}_j^*\right) + \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*) + \delta^{\frac{1}{2}} \|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)\|_2 \\
&+ \frac{1}{2}\delta \|\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*) - \mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\|_2 + o(\delta)
\end{aligned}
$$

Finally, maximizing over $i \neq j$ on both sides, the transfer measure is bounded by:

$$
\begin{aligned}
\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq& \frac{1}{2}\max_{i\neq j}\sup_{\boldsymbol{\theta}\in\mathcal{F}_2}\left(\mathcal{L}_{\mu_j}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \left(\mathcal{L}_{\mu_i}(\boldsymbol{\theta}) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)\right)\right) \\
\leq& \frac{1}{2}\max_{i\neq j} \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}_j^*\right) + \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*) \\
&+ \delta^{\frac{1}{2}} \|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*)\|_2 + \frac{1}{2}\delta\|\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*) - \mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\|_2 + o(\delta)
\end{aligned} \tag{38}
$$

Suppose a constant upper bound g on the maximum gradient norm exists, replacing the first-order term with $2\delta^{\frac{1}{2}}g$, we get:

$$
\begin{aligned}
\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}) \leq& \frac{1}{2}\max_{i\neq j} \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}_j^*\right) + \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*) + 2\delta^{\frac{1}{2}}g \\
&+ \frac{1}{2}\delta\|\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*) - \mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\|_2 + o(\delta) \\
\leq& \delta^{\frac{1}{2}}g + \frac{1}{2}\max_{i\neq j} \mathcal{L}_{\mu_j}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_j}\left(\boldsymbol{\theta}_j^*\right) + \mathcal{L}_{\mu_i}(\boldsymbol{\theta}^*) - \mathcal{L}_{\mu_i}(\boldsymbol{\theta}_i^*) \\
&+ \frac{1}{2}\delta\|\mathbf{H}_{\mu_j}(\boldsymbol{\theta}^*) - \mathbf{H}_{\mu_i}(\boldsymbol{\theta}^*)\|_2 + o(\delta) \quad \square
\end{aligned} \tag{39}
$$

D OTHER TRANSFER MEASURES

Proposition 4 (upper bounds on symmetric and realizable transfer measures). *Given $\mathcal{S} = \{\mu_i\}_{i=1}^K$ and some $\Gamma \subseteq \mathcal{H}$. Define $\mathcal{L}_{\mu_i}^* := \inf_{h\in\Gamma}\mathcal{L}_{\mu_i}(h)$ for all $i \in [K]$, $\mathcal{L}_{\mathcal{T}}{}^* := \inf_{h\in\Gamma}\mathcal{L}_{\mathcal{T}}(h)$, $\mu^* := \arg\min_\mu \max_{i\in[K]} \mathrm{T}_\Gamma(\mu_i\|\mu)$, and $\mathcal{L}_{\mathcal{S}}(h) := \mathcal{L}_{\mu^*}(h)$. Under Assumption 2, we have:*

$$
\begin{aligned}
\mathrm{T}_\Gamma(\mathcal{S},\mathcal{T}) &\leq \frac{1}{2}\max_{i\neq j}\mathrm{T}_\Gamma(\mu_j\|\mu_i) \\
\mathrm{T}_\Gamma^r(\mathcal{S},\mathcal{T}) &\leq \frac{1}{2}\max_{i\neq j}\mathrm{T}_\Gamma^{\mathrm{r}}(\mu_j,\mu_i)
\end{aligned} \tag{40}
$$

Proof. We first prove an upper bound on symmetric transfer measure $\mathrm{T}_\Gamma(\mathcal{S},\mathcal{T})$.

From Definition 3 and Eq. (3), we have:

$$
\begin{aligned}
\mathrm{T}_\Gamma(\mathcal{S},\mathcal{T}) &:= \max\left\{\mathrm{T}_\Gamma(\mathcal{S}\|\mathcal{T}), \mathrm{T}_\Gamma(\mathcal{T}\|\mathcal{S})\right\} \\
&\leq \max\left\{\frac{1}{2}\max_{i\neq j}\mathrm{T}_\Gamma(\mu_j\|\mu_i), \frac{1}{2}\max_{i\neq j}\mathrm{T}_\Gamma(\mu_i\|\mu_j)\right\} \\
&= \frac{1}{2}\max_{i\neq j}\mathrm{T}_\Gamma(\mu_j\|\mu_i)
\end{aligned} \tag{41}
$$

Now we prove an upper bound on realizable transfer measure $\mathrm{T}_\Gamma^{\mathrm{r}}(\mathcal{S},\mathcal{T})$

First, define $\mu^* := \arg\min_\nu \max_{i\in[K]} \mathrm{T}_\Gamma(\mu_i, \mu)$, and since $\mathcal{T}$ is a convex combination of distribution in $\mathcal{S}$:

$$\begin{aligned}
\mathrm{T}^{\mathrm{r}}_\Gamma(\mathcal{S}, \mathcal{T}) &= \sup_{h\in\Gamma} |\mathcal{L}_\mathcal{T}(h) - \mathcal{L}_\mathcal{S}(h)| \\
&= \sup_{h\in\Gamma} \left| \sum_{i=1}^K w_i \mathcal{L}_{\mu_i}(h) - \mathcal{L}_{\mu^*}(h) \right| \\
&= \sup_{h\in\Gamma} \left| \sum_{i=1}^K w_i \left[\mathcal{L}_{\mu_i}(h) - \mathcal{L}_{\mu^*}(h)\right] \right| \\
&\leq \sum_{i=1}^K w_i \sup_{h\in\Gamma} |\mathcal{L}_{\mu_i}(h) - \mathcal{L}_{\mu^*}(h)| \\
&= \sum_{i=1}^K w_i \mathrm{T}^{\mathrm{r}}_\Gamma(\mu_i, \mu^*) \\
&\leq \max_{i\in[k]} \mathrm{T}^{\mathrm{r}}_\Gamma(\mu_i, \mu^*)
\end{aligned} \tag{42}$$

Similar to one-sided transfer measure, let $j_{max} := \arg\max_{j\in[K]} \mathrm{T}^{\mathrm{r}}_\Gamma(\mu_j, \mu^*)$, and for any fixed , let μ_{mid} be such that $\mathrm{T}^{\mathrm{r}}_\Gamma(\mu_{j_{max}}, \mu_i) = \mathrm{T}^{\mathrm{r}}_\Gamma(\mu_{j_{max}}, \mu_{mid}) + \mathrm{T}^{\mathrm{r}}_\Gamma(\mu_{mid}, \mu_i)$. By the definition μ^* we have:

$$\begin{aligned}
\frac{1}{2}\max_{i\neq j} \mathrm{T}^{\mathrm{r}}_\Gamma(\mu_j, \mu_i) \geq& \frac{1}{2}\mathrm{T}^{\mathrm{r}}_\Gamma(\mu_{j_{max}}, \mu_i) \quad \forall i \in [K] \\
=& \mathrm{T}^{\mathrm{r}}_\Gamma(\mu_{mid}, \mu_i) \quad \forall i \in [K], \exists \mu_{mid} \\
\geq& \mathrm{T}^{\mathrm{r}}_\Gamma(\mu^*, \mu_i) \quad \forall i \in [K]
\end{aligned} \tag{43}$$

Finally, combining Eq. (42) and Eq. (43), we have:

$$\mathrm{T}^{\mathrm{r}}_\Gamma(\mathcal{S}, \mathcal{T}) \leq \frac{1}{2}\max_{i\neq j} \mathrm{T}^{\mathrm{r}}_\Gamma(\mu_j, \mu_i) \tag{44}$$

□

E MORE RELATED WORK

Domain Generalization. The goal of domain generalization is to learn a predictor using labeled data from multiple source domains that generalize well to related but unseen target domains (Blanchard et al., 2011; Muandet et al., 2013). The standard baseline for DG is Empirical Risk Minimization (ERM) (Vapnik, 1999), which minimizes the average loss across training domains. However, ERM does not generalize well under distribution shifts in the presence of spurious correlation in data (Arjovsky et al., 2020). Various approaches have been proposed to address the shortcomings of ERM. Below we discuss some approaches relevant to this work, Invariant Risk Minimization, gradient matching, and hessian matching.

Invariant Risk Minimization. The Invariant Risk Minimization (IRM) principle (Arjovsky et al., 2020) proposes jointly learning a feature extractor and a classifier such that the optimal classifier remains consistent across different training environments. The IRM objective, by definition, is non-convex and bi-level, so the authors proposed IRMv1, a regularized objective in place of the bi-level one. Later, we make the connection between our proposed loss objective and IRMv1. Followup works (Rosenfeld et al., 2021; Ahuja et al., 2022b; Wang et al., 2022, 2023; Krueger et al., 2021; Ahuja et al., 2022a, 2020; Kamath et al., 2021) showed that IRM and its variants do not improve over ERM unless the test domain are similar enough to the training domains.

Gradient Matching. Gradient matching methods seek alignment between domain-level gradients. For instance, IGA (Koyama and Yamaguchi, 2020) penalizes large Euclidean distances between gradients, Fish (Shi et al., 2021) increases the gradient inner products, and AND-Mask (Parascandolo et al., 2020) only updates the parameters whose gradients are of the same sign across all environments. Despite their good performance, Hemati et al. (2023) showed that aligning domain-level gradients does not guarantee small generalization loss to the test domain.

Hessian Matching. Most relevant to our approach, a recent line of DG works align the domain level Hessians w.r.t. the classifier head to promote consistency (Parascandolo et al., 2020) across domains. Due to the complexity of computing the Hessian matrices, prior works find Hessian approximations instead. CORAL Sun and Saenko (2016) minimizes the

difference in feature covariance matrices between source and target domains, which is approximately Hessian matching. Fishr (Rame et al., 2022) uses domain-level gradient variance as its hessian approximation. The idea of aligning gradients and Hessian simultaneously was first proposed by Hemati et al. (2023), who also discussed what attributes are aligned by gradients and Hessian matching.

Domain-Invariant Feature Learning. Initially proposed by Ben-David et al. (2010), invariant representation learning seeks various types of invariance across domains. For instance, Ganin et al. (2016); Li et al. (2018); Tzeng et al. (2017); Hoffman et al. (2017) employ adversarial training, whereas Muandet et al. (2013); Long et al. (2015) uses kernel method, Huang et al. (2025) seeks invariant parameters, and Peng et al. (2019); Zellinger et al. (2019); Sun and Saenko (2016) match the feature moments for domain adaptation. In particular, Sun and Saenko (2016) introduces CORAL, which matches the covariance between features in the source and target domains and achieves state-of-the-art performance as evaluated by Gulrajani and Lopez-Paz (2020) and Hemati et al. (2023). Most of the invariant representation learning methods are originally for domain adaptation, where one has access to unlabelled data from the test domain. In the case of multi-domain generalization, these methods can be adopted by finding invariance across training domains. Nevertheless, Zhao et al. (2019) shows that matching the features is insufficient for DG.

F CONNECTION BETWEEN CMA AND EXISTING METHODS

By alignment of both gradient and Hessians in closed form, CMA implicitly integrates multiple existing algorithms. Below we build such connections.

F.1 CMA AS INVARIANT RISK MINIMIZATION

We draw connections between IRM and CMA objectives. Fixing a feature extractor and letting the classifier head be parameterized by θ, the IRMv1 objective in Arjovsky et al. (2020) is:

$$\mathcal{L}_{\text{IRM}} := \mathcal{L}_{\text{ERM}} + \lambda \frac{1}{K} \sum_{i=1}^{K} \| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i} (\boldsymbol{\theta}) \|_2^2 \tag{IRMv1}$$

On the other hand, we can rewrite the gradient variance regularization in Eq. (CMA) as

$$\frac{1}{K} \sum_{i=1}^{K} \| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i} (\boldsymbol{\theta}) - \overline{\nabla_{\boldsymbol{\theta}} \mathcal{L} (\boldsymbol{\theta})} \|_2^2 = \frac{1}{K} \sum_{i=1}^{K} \| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i} (\boldsymbol{\theta}) \|_2^2 - \| \frac{1}{K} \sum_{j=1}^{K} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_j} (\boldsymbol{\theta}) \|_2^2 \tag{45}$$

The second term on the right-hand side, the norm of the average gradients, is small for a classifier $\boldsymbol{\theta}^*$ well-trained on $\mathcal{L}_{\text{ERM}}$, and the first term resembles the regularization in Eq. (IRMv1). Therefore, penalizing large gradient variance can be seen as enforcing the learned classier $\boldsymbol{\theta}$ to be invariant across domains. Under the same assumptions as in Theorem 1, at the optimal invariant predictor $\boldsymbol{\theta}^*$, the norm of the average of gradients is zero, making the gradient variance term in Eq. (CMA) exactly the gradient penalty in Eq. (IRMv1). By setting $\beta = 0$ in Eq. (CMA), we recover Eq. (IRMv1).

F.2 CMA AS GRADIENT MATCHING

While multiple version of gradient matching losses have been proposed (Shi et al., 2021; Koyama and Yamaguchi, 2020; Parascandolo et al., 2020), we focus on the most recent one proposed by Shi et al. (2021), defined as:

$$\mathcal{L}_{\text{GM}} := \mathcal{L}_{\text{ERM}} + \lambda \frac{1}{K} \left(\sum_{i=1}^{K} \| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i} (\boldsymbol{\theta}) \|_2^2 - \left\| \sum_{j=1}^{K} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_j} (\boldsymbol{\theta}) \right\|_2^2 \right) \tag{GM}$$

Comparing the second term with Eq. (45), and ignoring the constant factor λ, the difference is $\frac{K-1}{K^2} \| \sum_{j=1}^{K} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_j} (\boldsymbol{\theta}) \|_2^2$. When an invariant optimal predictor $\boldsymbol{\theta}^*$ exists, this difference vanishes, and setting $\beta = 0$ in Eq. (CMA) recovers Eq. (GM).

F.3 CMA AS HESSIAN MATCHING

We first compare CMA with Fishr (Rame et al., 2022), a state-of-the-art DG algorithm based on Hessian matching. The principle behind Hessian matching is to match the domain-level Hessian matrices by minimizing the objective:

$$\mathcal{L}_{\text{HM}} := \mathcal{L}_{\text{ERM}} + \lambda \frac{1}{K} \sum_{i=1}^{K} \|\mathbf{H}_{\mu_i} - \overline{\mathbf{H}}\|_F^2 \tag{HM}$$

Rame et al. (2022) achieves this by approximating the Hessian matrices with their diagonals. In contrast, we proposed to compute the Hessian matrices analytically. Thus, by setting $\alpha = 0$, Eq. (CMA) is the closed-form of the Fishr objective.

Next, we compare CMA with the two objectives proposed in Hemati et al. (2023), namely HGP and Hutchinson's method (eq. (18) and eq. (23) in Hemati et al. (2023)):

$$\mathcal{L}_{\text{HGP}} = \mathcal{L}_{\text{ERM}} + \frac{1}{K} \sum_{i=1}^{K} \alpha \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i} - \overline{\nabla_{\boldsymbol{\theta}} \mathcal{L}}\|_2^2 + \beta \|\mathbf{H}_{\mu_i} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i} - \overline{\mathbf{H} \nabla_{\boldsymbol{\theta}} \mathcal{L}}\|_2^2 \tag{HGP}$$

where $\overline{\mathbf{H} \nabla_{\boldsymbol{\theta}} \mathcal{L}} = \frac{1}{K} \sum_{i=1}^{K} \mathbf{H}_{\mu_i} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i}$ is the average Hessian-gradient product.

$$\mathcal{L}_{\text{Hutchinson}} = \mathcal{L}_{\text{ERM}} + \frac{1}{K} \sum_{i=1}^{K} \alpha \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mu_i} - \overline{\nabla_{\boldsymbol{\theta}} \mathcal{L}}\|_2^2 + \beta \|\mathbf{D}_{\mu_i} - \overline{\mathbf{D}}\|_2^2 \tag{Hutchinson}$$

where $\mathbf{D}_{\mu_i}$ is the Hessian diagonal estimated by Hutchinson's method (Bekas et al., 2007). Like CMA, HGP, and Hutchinson match the first and second moment across domains. Unlike CMA, HGP approximates the second-order penalties with Hessian-gradient products, while Hutchinson's method estimates them with Hessian diagonals which themselves are estimated by sampling. In other words, Eq. (CMA) is the closed form of Eq. (HGP) and Eq. (Hutchinson).

G GRADIENT AND HESSIAN DERIVATIONS

G.1 CROSS-ENTROPY LOSS

G.1.1 Gradient

Given the logistic regression model without a bias term, parameterized by $\boldsymbol{\theta} = \{\mathbf{w}_1, \ldots, \mathbf{w}_C\}$, where $\mathbf{w}_c \in \mathbb{R}^d$ for all $c \in [C]$, and the prediction $p_c = \frac{e^{\mathbf{w}_c^\top \mathbf{x}}}{\sum_{j=1}^{C} e^{\mathbf{w}_j^\top \mathbf{x}}}$, the cross-entropy loss for a single example $(\mathbf{x}, y)$ is defined as:

$$\ell(\boldsymbol{\theta}) = -\sum_{c=1}^{C} y_c \log(p_c)$$

To find the gradient of the loss w.r.t. $\mathbf{w}_k$, we compute:

$$\begin{aligned}
\nabla_{\mathbf{w}_k} \ell(\boldsymbol{\theta}) &= -\sum_{c=1}^{C} y_c \nabla_{\mathbf{w}_k} \log(p_c) \\
&= -\sum_{c \neq k}^{C} y_c \nabla_{\mathbf{w}_k} \log(p_c) - y_k \nabla_{\mathbf{w}_k} \log(p_k) \\
&= \sum_{c \neq k}^{C} y_c p_k \mathbf{x} - y_k \mathbf{x}(1 - p_k) \\
&= (1 - y_k) p_k \mathbf{x} - y_k \mathbf{x}(1 - p_k) \\
&= (p_k - y_k)\mathbf{x}
\end{aligned}$$

From the second to the third equality, we use the facts that

$$\nabla_{\mathbf{w}_k} p_c = \begin{cases} p_k \left(1 - p_k\right) \mathbf{x}, & \text{if } c = k \\ -p_c p_k \mathbf{x}, & \text{if } c \neq k \end{cases}$$

$$\nabla_{\mathbf{w}_k} \log\left(p_c\right) = \begin{cases} \left(1 - p_k\right) \mathbf{x}, & \text{if } c = k \\ -p_k \mathbf{x}, & \text{if } c \neq k \end{cases}$$

G.1.2 Hessian

To find the Hessian matrix, we compute the second-order partial derivatives. We consider two cases:

Case 1: $k = c$:

$$\begin{aligned} \nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_k} \ell(\boldsymbol{\theta}) &= \nabla_{\mathbf{w}_k} \left(\left(p_k - y_k\right) \mathbf{x}\right) \\ &= \nabla_{\mathbf{w}_k} p_k \mathbf{x} \\ &= p_k \left(1 - p_k\right) \mathbf{x}\mathbf{x}^\top \end{aligned}$$

Case 2: $k \neq c$:

$$\begin{aligned} \nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_c} \ell(\boldsymbol{\theta}) &= \nabla_{\mathbf{w}_k} \left(\left(p_c - y_c\right) \mathbf{x}\right) \\ &= \nabla_{\mathbf{w}_k} p_c \mathbf{x} \\ &= -p_c p_k \mathbf{x}\mathbf{x}^\top \end{aligned}$$

Combining these results, we write the Hessian matrix as:

$$\mathbf{H} = \left(\operatorname{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top\right) \otimes \left(\mathbf{x}\mathbf{x}^\top\right)$$

Where:

- $\operatorname{diag}(\mathbf{p}) \in \mathbb{R}^{C \times C}$ is the diagonal matrix with elements of $\mathbf{p}$, $p_1, \ldots, p_C$, on the diagonal.
- $\mathbf{x}\mathbf{x}^\top \in \mathbb{R}^{d \times d}$.
- $\otimes$ denotes the Kronecker product.

G.1.3 Higher Order Derivatives of Logistic Regression Classifier

We show by induction that the n^{th} order derivative of the cross-entropy loss w.r.t. the weight vector $\mathbf{w}$ of a binary-logistic regression classifier is:

$$\nabla_{\mathbf{w}}^{\mathtt{n}} \ell(\boldsymbol{\theta}) = Q_n(p) \mathbf{x}^{\otimes \mathtt{n}} \tag{46}$$

where $Q_n(p)$ is some scalar-valued polynomial function of p.

Proof. By induction.

Base Case ($n = 1$):

For $n = 1$, the gradient of the cross-entropy loss ℓ w.r.t. $\mathbf{w}$ is:

$$\nabla_{\mathbf{w}}^{1} \ell(\boldsymbol{\theta}) = (p - y)\mathbf{x}$$

This matches the form $Q_1(p)\mathbf{x}^{\otimes 1}$ for $Q_1(p) = p - y$. We have that the base case holds.

Inductive Step: Assume Eq. (46) holds for some n

$$\nabla_{\mathbf{w}}^{\mathtt{n}} \ell(\boldsymbol{\theta}) = Q_n(p) \mathbf{x}^{\otimes \mathtt{n}}$$

we need to show that it also holds for $(n+1)$:

$$\nabla_{\mathbf{w}}^{\mathtt{n}+1}\ell(\boldsymbol{\theta}) = Q_n\left(p\right)\mathbf{x}^{\otimes(\mathtt{n}+1)}$$

By the product rule:

$$\begin{aligned}\nabla_{\mathbf{w}}^{\mathtt{n}+1}\ell(\boldsymbol{\theta}) =& \nabla_{\mathbf{w}} Q_n\left(p\right)\mathbf{x}^{\otimes\mathtt{n}} \\ =& \left[\nabla_{\mathbf{w}} Q_n\left(p\right)\right]\mathbf{x}^{\otimes\mathtt{n}}\end{aligned}$$

And by chain rule:

$$\nabla_{\mathbf{w}} Q_n\left(p\right) = \left[\nabla_p Q_n\left(p\right)\right]\nabla_{\mathbf{w}} p$$

The first gradient is the derivative of a polynomial function of p, which is again a polynomial function of p. The second term, as we have seen in Appendix G.1.2, is $p(1-p)\mathbf{x}$. Now putting everything together, we have

$$\begin{aligned}\nabla_{\mathbf{w}}^{\mathtt{n}+1}\ell(\boldsymbol{\theta}) =& \left[\nabla_{\mathbf{w}} Q_n\left(p\right)\right]\mathbf{x}^{\otimes\mathtt{n}} \\ =& \left[\nabla_{\mathbf{w}} Q_n\left(p\right)\right]p(1-p)\mathbf{x}\mathbf{x}^{\otimes\mathtt{n}} \\ =& Q_{n+1}\left(p\right)\mathbf{x}^{\otimes\mathtt{n}+1}\end{aligned}$$

which completes the induction. □

G.1.4 Memory-Efficient Hessian Frobenius Norm

Note that to obtain the Frobenius of the hessian, we do not need to compute the Kroncker product explicitly:

$$\|\mathbf{H}\|_F^2 = \operatorname{tr}\left(\operatorname{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top\right)\operatorname{tr}\left(\mathbf{x}\mathbf{x}^\top\right)$$

To compute the Hessian regularization

$$\frac{1}{K}\sum_{i=1}^{K}\|\mathbf{H}_{\mu_i}\left(\boldsymbol{\theta}\right) - \overline{\mathbf{H}\left(\boldsymbol{\theta}\right)}\|_F^2$$

without saving the $dC \times dC$ Kroncker product, we first expand the Frobenius norm:

$$\begin{aligned}\|\mathbf{H}_{\mu_i}(\boldsymbol{\theta}) - \overline{\mathbf{H}(\boldsymbol{\theta})}\|_F^2 &= \|\mathbf{H}_{\mu_i}(\boldsymbol{\theta})\|_F^2 + \left\|\frac{1}{K}\sum_{j=1}^{K}\mathbf{H}_{\mu_j}(\boldsymbol{\theta})\right\|_F^2 - 2\left\langle\mathbf{H}_{\mu_i}(\boldsymbol{\theta}), \frac{1}{K}\sum_{j=1}^{K}\mathbf{H}_{\mu_j}(\boldsymbol{\theta})\right\rangle_F \\ &= \|\mathbf{H}_{\mu_i}(\boldsymbol{\theta})\|_F^2 + \frac{1}{K^2}\left\|\sum_{j=1}^{K}\mathbf{H}_{\mu_j}(\boldsymbol{\theta})\right\|_F^2 - \frac{2}{K^2}\sum_{j=1}^{K}\left\langle\mathbf{H}_{\mu_i}(\boldsymbol{\theta}), \mathbf{H}_{\mu_j}(\boldsymbol{\theta})\right\rangle_F \\ &= \|\mathbf{H}_{\mu_i}(\boldsymbol{\theta})\|_F^2 + \frac{1}{K^2}\sum_{j,l=1}^{K}\langle\mathbf{H}_{\mu_j}(\boldsymbol{\theta}), \mathbf{H}_{\mu_l}(\boldsymbol{\theta})\rangle_F - \frac{2}{K^2}\sum_{j=1}^{K}\langle\mathbf{H}_{\mu_i}(\boldsymbol{\theta})\mathbf{H}_{\mu_j}(\boldsymbol{\theta})\rangle_F\end{aligned}$$

We need $\left\langle\mathbf{H}_{\mu_i}(\boldsymbol{\theta}), \mathbf{H}_{\mu_j}(\boldsymbol{\theta})\right\rangle_F$ for all $i, j \in [K]$. For the ease of notation, we denote the two environmental Hessians as $\mathbf{H}^{e_1}, \mathbf{H}^{e_2}$, $\mathcal{E}_e$ as the indices of points in environment e, and $\mathbf{H}_i$ as the Hessian of the sample i.

$$\begin{aligned}\langle H^{e_1}, H^{e_2}\rangle_F &= \frac{1}{|\mathcal{E}_{e_1}||\mathcal{E}_{e_2}|}\sum_{i\in\mathcal{E}_{e_1}}\sum_{j\in\mathcal{E}_{e_2}}\langle\mathbf{H}_i, \mathbf{H}_j\rangle_F \\ &= \frac{1}{|\mathcal{E}_{e_1}||\mathcal{E}_{e_2}|}\sum_{i\in\mathcal{E}_{e_1}}\sum_{j\in\mathcal{E}_{e_2}}\operatorname{tr}(\mathbf{H}_i\mathbf{H}_j) \\ &= \frac{1}{|\mathcal{E}_{e_1}||\mathcal{E}_{e_2}|}\sum_{i\in\mathcal{E}_{e_1}}\sum_{j\in\mathcal{E}_{e_2}}\operatorname{tr}\left((\operatorname{diag}(\mathbf{p}^{(i)}) - \mathbf{p}^{(i)}\mathbf{p}^{(i)\top})\otimes\mathbf{x}^{(i)}\mathbf{x}^{(i)\top}((\operatorname{diag}(\mathbf{p}^{(j)}) - \mathbf{p}^{(j)}\mathbf{p}^{(j)\top})\otimes\mathbf{x}^{(j)}\mathbf{x}^{(j)\top}\right) \\ &= \frac{1}{|\mathcal{E}_{e_1}||\mathcal{E}_{e_2}|}\sum_{i\in\mathcal{E}_{e_1}}\sum_{j\in\mathcal{E}_{e_2}}\operatorname{tr}\left(\operatorname{diag}(\mathbf{p}^{(i)} - \mathbf{p}^{(i)}\mathbf{p}^{(i)\top})\operatorname{diag}(\mathbf{p}^{(j)} - \mathbf{p}^{(j)}\mathbf{p}^{(j)\top})\right)\operatorname{tr}\left(\mathbf{x}^{(i)}\mathbf{x}^{(i)\top})(\mathbf{x}^{(j)}\mathbf{x}^{(j)\top}\right)\end{aligned}$$

The last expression only involves matrices of dimensions $C \times C$ and $d \times d$.

However, this memory-efficient method requires computing the trace for all pairs of Hessians, $\mathbf{H}_i$ and $\mathbf{H}_j$, where $(i, j) \in (\mathcal{E}_{e_1}, \mathcal{E}_{e_2})$ for each combination of environments $e_1, e_2 \in [K]$.

G.2 MEAN-SQUARED ERROR LOSS

We derive the gradient and Hessian of the mean-squared error (MSE) loss. Given a linear regression model parameterized by $\mathbf{w} \in \mathbb{R}^d$, where the prediction is $\hat{y} = \mathbf{w}^\top \mathbf{x}$, the mean-squared error loss for a single example $(\mathbf{x}, y)$ is defined as:

$$\ell(\mathbf{w}) = \frac{1}{2}(\hat{y} - y)^2 = \frac{1}{2}(\mathbf{w}^\top \mathbf{x} - y)^2$$

G.2.1 Gradient

To find the gradient of the loss w.r.t. $\mathbf{w}$, we compute:

$$\begin{aligned}
\nabla_{\mathbf{w}} \ell(\boldsymbol{\theta}) &= \nabla_{\mathbf{w}} \frac{1}{2}(\mathbf{w}^\top \mathbf{x} - y)^2 \\
&= (\mathbf{w}^\top \mathbf{x} - y) \nabla_{\mathbf{w}} (\mathbf{w}^\top \mathbf{x} - y) \\
&= (\mathbf{w}^\top \mathbf{x} - y)\mathbf{x} \\
&= (\hat{y} - y)\mathbf{x}
\end{aligned}$$

G.2.2 Hessian

To find the Hessian matrix, we compute the second-order partial derivatives:

$$\begin{aligned}
\mathbf{H}(\mathbf{x}) = \nabla_{\mathbf{w}} \nabla_{\mathbf{w}} \ell(\boldsymbol{\theta}) &= \nabla_{\mathbf{w}} (\hat{y} - y)\mathbf{x} \\
&= \nabla_{\mathbf{w}} (\mathbf{w}^\top \mathbf{x} - y)\mathbf{x} \\
&= \mathbf{x}\mathbf{x}^\top
\end{aligned}$$

Note that the second-order derivative of MSE loss is a constant matrix w.r.t. $\mathbf{w}$, so higher-order derivatives are tensors with all zeros.

H EXPERIMENTAL DETAILS

H.1 LINEAR PROBING

For the linear probing experiments in Section 6, we conduct a grid search for both α and β in Eq. (CMA) over the set {1, 10, 100, 1000, 2000, 5000, 10000}. We also implement penalty annealing, wherein the gradient and Hessian penalties are initially set to zero and activated only after a predetermined number of updates. This approach ensures that the classifier, to which further regularization is subsequently applied, already achieves a small ERM loss. For Fishr, we perform a grid search over the suggested hyperparameter ranges by Rame et al. (2022). The grid search is first conducted using a single random seed. From this, the top five performing sets of hyperparameters are chosen. These sets are then evaluated using four additional random seeds. Lastly, we report the test performance of the hyperparameter set that demonstrates the highest worst-group validation accuracy over five runs. Here we summarize the best hyperparameters found for Fishr and CMA. We train on Waterbirds for 300 epochs, CelebA for 50 epochs, and MultiNLI for 3 epochs. Since ISR projected features have small dimensions (we follow the implementation in Wang et al. (2022) and choose 100), the experiment is computationally efficient to run, taking five days on four RTX 6000 GPUs.

H.1.1 Datasets

Waterbirds Sagawa et al. (2020): This is an image dataset, where each image is a combination of a bird image from the CUB (Wah et al., 2011) and a background image from the Place dataset (Zhou et al., 2018). Each combined image is labelled

Table 6: Best hyperparameters for Fishr and CMA on each dataset.

Algorithm	Parameter	**Waterbirds**	**CelebA**	**MultiNLI**
Fishr	regularization strength λ	100	10	10000
	ema γ	0.945	0.9225	0.99
	annealing iterations	2800	12000	600
CORAL	regularization strength γ	0.256	0.16	0.45
CMA	gradient regularization strength α	10	5000	5000
	hessian regularization strength β	1000	100	1
	annealing iterations	2100	4000	0

with class $y \in \mathcal{Y} = \{\textit{waterbird}, \textit{landbird}\}$ and environment $e \in \mathcal{E} = \{\textit{water\_background}, \textit{land\_background}\}$. Each (y, e) pair forms a group, for a total of 4 groups $\mathcal{G} = \mathcal{Y} \times \mathcal{E}$. There are 4795 training samples, and the smallest group has 56.

CelebA Liu et al. (2015): This is an image dataset composed of celebrity faces. Following Sagawa et al. (2020) and Wang et al. (2022, 2023), we consider a hair color classification task ($\mathcal{Y} = \{\textit{blond}, \textit{dark}\}$) with gender as spurious feature ($\mathcal{E} = \{\textit{male}, \textit{female}\}$). The four groups are formed by $\mathcal{G} = \mathcal{Y} \times \mathcal{E}$. There are 162k training samples, and the smallest group, males with blond hair, has 1387 samples.

MultiNLI (Williams et al., 2018): This is a text dataset for natural language inference. Each sample is composed of one hypothesis and one premise, and the task is to determine whether the given premise entails, is neutral with, or contradicted by the hypothesis ($\mathcal{Y} = \{\textit{contradiction}, \textit{neutral}, \textit{entailment}\}$). The spurious attribute is the presence of negation words, for example, "no", "nobody", "never", and "nothing" ($\mathcal{E} = \{\textit{no\_negation}, \textit{negation}\}$). The presence of negation words spuriously correlated with $y = contradiction$ (Gururangan et al., 2018). There are six groups formed by $\mathcal{G} = \mathcal{Y} \times \mathcal{E}$, for a total of 206175 samples in the training set. The smallest group, entailment with negations, contains 1521 examples.

H.2 FINE-TUNING

For the fine-tuning experiments in Section 6, we employ two model selection strategies from DomainBed (Gulrajani and Lopez-Paz, 2020): test-domain model selection and training-domain model selection. In test-domain model selection, we select the best hyperparameters based on a validation set that follows the same distribution as the test data. On the other hand, for training-domain model selection, the best hyperparameters are chosen based on performance across holdout sets from the training domains. Contrary to the original DomainBed setup, which randomizes batch sizes, we standardize the batch size to 64 for ColoredMNIST and RotatedMNIST, and to 32 for real image datasets. For each algorithm, we randomly search for 5 sets of hyperparameters and 3 runs each. The experiments take around 10 days on 4 RTX 6000 GPUs.

Despite the original DomainBed codebase recommending a search over 20 sets of hyperparameters per algorithm, per dataset, and per test domain, we restricted our search to only 5 sets due to time and resource constraints. Even with this limitation, our approach required running 1260 experiments. While this reduced number of searches means the algorithms might not have achieved their full potential, this limitation applies equally to all algorithms, ensuring a fair comparison. As our experiments are intended as proof-of-concept rather than comprehensive evaluations, we argue that the results in this section are sufficient to validate the effectiveness of our algorithm.

In the main text, we follow Rame et al. (2022) and report the test-domain validation performance. In practice, test-domain model selection is more realistic compared to training-domain model selection, as practitioners are unlikely to deploy a model without validating it with at least some small-scale data from the target domain. Additionally, as discussed in Rame et al. (2022) and Teney et al. (2022), by the definition of distribution shift, one cannot expect a model selected on a validation set sampled from the same distribution as the training set to generalize to an unseen test distribution. For completeness, we also present the training-domain validation performance in Appendix H.3.

H.2.1 Datasets

Colored MNIST (Arjovsky et al., 2020): This is an image dataset derived from the MNIST handwritten digit classification dataset (LeCun et al., 2010). The task is to identify whether a digit is in 0-4 or 5-9 ($\mathcal{Y} = \{0-4, 5-9\}$). The digits are

colored red or blue. The environments contain colored digits correlated differently ($\mathcal{E} = \{+90\%, +80\%, -90\%\}$) with the target label. In the first environment, the green color has a 90% correlation with class 5-9; similar correlations apply in the other two environments. Additionally, there is a 25% chance of label flipping. The dataset contains 70,000 examples of dimensions (2, 28, 28) categorized into 2 classes.

Rotated MNIST (Ghifary et al., 2015): This is another variant of MNIST, where each environment $e \in \mathcal{E} = \{0, 15, 30, 45, 60, 75\}$ is composed of digits rotated by e degrees. The dataset contains 70,000 examples of dimensions (1, 28, 28) and 10 classes.

PACS (Li et al., 2017): This is a 7-class classification dataset, where each image is either photo, art painting, cartoon, or sketch ($\mathcal{E} = \{photo, art\_painting, cartoon, sketch\}$). There are 9,991 samples, each with dimensions (3, 224, 224).

VLCS (Fang et al., 2013): This is a 5 class images dataset with images from environments $\mathcal{E} = \{Caltech101, LabelMe, SUN09, VOC2007\}$. There are 10,729 samples of dimension (3, 224, 224).

Terra Incognita (Beery et al., 2018): This is a dataset of photographs taken from various locations, each corresponds to one environment ($\mathcal{E} = \{L100, L38, L43, L46\}$). The DomainBed benchmark includes a subset of Terra Incognita, comprising 24,788 samples with dimensions (3, 224, 224) across 10 classes.

H.3 DOMAINBED RESULTS

H.3.1 Model selection: training-domain validation set

ColoredMNIST

Algorithm	+90%	+80%	-90%	Avg
ERM	72.2 ± 0.2	72.9 ± 0.2	10.1 ± 0.1	51.7
CORAL	71.7 ± 0.4	73.2 ± 0.1	10.2 ± 0.1	51.7
Fishr	72.6 ± 0.3	73.3 ± 0.1	10.6 ± 0.2	52.2
CMA	71.4 ± 0.3	72.8 ± 0.1	10.0 ± 0.2	51.4

RotatedMNIST

Algorithm	0	15	30	45	60	75	Avg
ERM	95.3 ± 0.2	98.6 ± 0.1	99.1 ± 0.1	98.9 ± 0.0	98.9 ± 0.0	96.1 ± 0.2	97.8
CORAL	95.7 ± 0.2	98.7 ± 0.1	99.0 ± 0.0	99.0 ± 0.0	99.0 ± 0.0	96.5 ± 0.0	98.0
Fishr	95.6 ± 0.3	98.5 ± 0.1	99.1 ± 0.1	99.0 ± 0.1	99.0 ± 0.1	96.4 ± 0.0	97.9
CMA	95.2 ± 0.2	98.4 ± 0.2	98.9 ± 0.0	98.9 ± 0.0	98.9 ± 0.1	96.5 ± 0.2	97.8

VLCS

Algorithm	C	L	S	V	Avg
ERM	97.1 ± 0.1	62.3 ± 0.3	71.9 ± 0.7	77.2 ± 0.4	77.2
CORAL	96.3 ± 0.1	64.5 ± 0.4	72.4 ± 0.3	72.4 ± 1.7	76.4
Fishr	96.4 ± 0.6	63.3 ± 0.9	74.8 ± 0.6	76.2 ± 0.4	77.7
CMA	96.1 ± 0.6	63.2 ± 0.4	73.5 ± 0.4	78.9 ± 0.3	77.9

PACS

Algorithm	A	C	P	S	Avg
ERM	80.2 ± 0.6	75.4 ± 0.2	95.9 ± 0.8	66.6 ± 0.3	79.5
CORAL	81.6 ± 0.6	74.9 ± 0.8	95.4 ± 0.6	64.9 ± 0.6	79.2
Fishr	83.1 ± 1.0	74.8 ± 0.5	97.2 ± 0.2	68.7 ± 0.8	81.0
CMA	83.3 ± 0.3	76.4 ± 0.2	96.1 ± 0.1	66.3 ± 0.7	80.5

TerraIncognita

Algorithm	L100	L38	L43	L46	Avg
ERM	48.2 ± 2.1	17.8 ± 2.3	37.8 ± 1.0	34.2 ± 0.5	34.5
CORAL	39.1 ± 2.1	12.4 ± 2.1	36.0 ± 1.4	30.6 ± 0.9	29.5
Fishr	47.2 ± 2.1	16.5 ± 1.6	39.9 ± 1.9	33.2 ± 0.7	34.2
CMA	45.8 ± 3.3	19.0 ± 1.2	37.7 ± 0.3	33.4 ± 1.0	34.0

Averages

Algorithm	ColoredMNIST	RotatedMNIST	VLCS	PACS	TerraIncognita	Avg
ERM	51.7 ± 0.1	97.8 ± 0.1	77.2 ± 0.2	79.5 ± 0.3	34.5 ± 0.4	68.1
CORAL	51.7 ± 0.1	98.0 ± 0.0	76.4 ± 0.5	79.2 ± 0.1	29.5 ± 1.1	67.0
Fishr	52.2 ± 0.1	97.9 ± 0.1	77.7 ± 0.4	81.0 ± 0.3	34.2 ± 0.9	68.6
CMA	51.4 ± 0.0	97.8 ± 0.0	77.9 ± 0.1	80.5 ± 0.2	34.0 ± 0.7	68.3

H.3.2 Model selection: test-domain validation set (oracle)

ColoredMNIST

Algorithm	+90%	+80%	-90%	Avg
ERM	68.1 ± 1.1	70.5 ± 0.7	25.0 ± 1.9	54.5
CORAL	68.2 ± 0.9	72.0 ± 0.8	26.9 ± 0.1	55.7
Fishr	73.9 ± 0.3	73.5 ± 0.2	38.5 ± 5.2	62.0
CMA	70.9 ± 0.6	72.2 ± 0.2	44.3 ± 2.9	62.5

RotatedMNIST

Algorithm	0	15	30	45	60	75	Avg
ERM	95.2 ± 0.3	98.5 ± 0.1	98.9 ± 0.1	98.9 ± 0.1	99.0 ± 0.1	96.2 ± 0.2	97.8
CORAL	95.8 ± 0.1	98.7 ± 0.1	98.9 ± 0.0	99.2 ± 0.1	99.1 ± 0.0	96.5 ± 0.1	98.0
Fishr	95.7 ± 0.2	98.7 ± 0.0	99.0 ± 0.1	99.1 ± 0.1	98.8 ± 0.2	96.4 ± 0.0	97.9
CMA	95.7 ± 0.2	98.8 ± 0.1	98.9 ± 0.1	98.9 ± 0.0	98.9 ± 0.1	95.9 ± 0.6	97.9

VLCS

Algorithm	C	L	S	V	Avg
ERM	96.4 ± 0.1	62.3 ± 1.0	72.1 ± 0.6	76.7 ± 0.3	76.9
CORAL	95.8 ± 0.3	63.1 ± 0.3	71.2 ± 0.3	73.5 ± 0.2	75.9
Fishr	96.0 ± 0.8	64.0 ± 0.1	73.5 ± 0.7	76.4 ± 0.6	77.5
CMA	95.8 ± 0.4	65.0 ± 0.5	70.6 ± 2.4	78.1 ± 0.3	77.4

PACS

TerraIncognita

H.3.3 Additional Baselines

We compare CMA against additional baselines on ColoredMNIST and RotatedMNIST datasets and discuss the results using test-domain model selection. From Table 7, we observe that CMA has the second-highest average accuracy. Note that VREx surpasses CMA on ColoredMNIST but has a substantially larger variance (4.6) compared to CMA (0.9).

Algorithm	A	C	P	S	Avg
ERM	81.2 ± 0.9	73.4 ± 0.9	96.1 ± 0.6	70.3 ± 0.5	80.2
CORAL	80.6 ± 0.6	74.9 ± 0.2	95.9 ± 0.4	69.4 ± 0.2	80.2
Fishr	83.6 ± 0.6	74.9 ± 1.0	97.4 ± 0.3	70.1 ± 0.5	81.5
CMA	82.8 ± 0.7	76.7 ± 1.3	97.3 ± 0.2	69.5 ± 0.7	81.6

Algorithm	L100	L38	L43	L46	Avg
ERM	50.2 ± 0.4	25.0 ± 1.9	36.3 ± 1.6	34.5 ± 0.1	36.5
CORAL	43.1 ± 3.2	21.4 ± 2.7	37.5 ± 0.6	32.1 ± 0.5	33.6
Fishr	49.9 ± 2.1	23.2 ± 1.8	41.4 ± 1.2	34.7 ± 0.7	37.3
CMA	47.5 ± 3.4	44.7 ± 2.4	29.0 ± 3.2	32.4 ± 0.9	38.4

Table 7: Model selection: test-domain validation set

Algorithm	ColoredMNIST	RotatedMNIST	Avg
ERM	54.5 ± 0.2	97.8 ± 0.1	76.2
CORAL	55.7 ± 0.5	98.0 ± 0.0	76.9
Fishr	62.0 ± 1.7	97.9 ± 0.0	80.0
GroupDRO	59.6 ± 0.3	98.0 ± 0.1	78.8
DANN	53.5 ± 0.7	97.4 ± 0.0	75.5
CDANN	53.6 ± 0.4	97.6 ± 0.0	75.6
VREx	66.1 ± 4.6	97.8 ± 0.0	**82.0**
SelfReg	53.8 ± 0.8	98.0 ± 0.1	75.9
CMA	62.5 ± 0.9	97.9 ± 0.1	<u>80.2</u>

Table 8: Model selection: Training-domain validation set

Algorithm	ColoredMNIST	RotatedMNIST	Avg
ERM	51.7 ± 0.1	97.8 ± 0.1	75.8
CORAL	51.7 ± 0.1	98.0 ± 0.0	74.9
Fishr	52.2 ± 0.1	97.9 ± 0.1	75.1
GroupDRO	51.9 ± 0.1	97.9 ± 0.1	74.9
DANN	51.7 ± 0.0	97.6 ± 0.2	74.6
CDANN	51.9 ± 0.2	97.8 ± 0.0	74.8
VREx	51.7 ± 0.1	97.7 ± 0.1	74.7
SelfReg	51.7 ± 0.0	98.1 ± 0.1	74.9
CMA	51.4 ± 0.0	97.8 ± 0.0	74.6

H.4 COMPARISON TO HGP

We compare CMA with the HGP algorithm (Hemati et al., 2023). Both algorithms align the gradients and Hessians, so we expect their performances to be similar. We do not compare CMA with Hutchinson's in (Hemati et al., 2023) due to the time costs incurred by sampling-based Hessian estimation.

H.4.1 Linear Probing

As shown in Appendix H.4.1, the two algorithms have comparable performance overall, except for the CelebA dataset. A potential explanation for this discrepancy is the differences in Hessian computations. Note that the original HGP does not apply penalty annealing. We added penalty annealing to both methods to eliminate differences caused by this factor, allowing us to focus on differences in the loss objectives.

Table 9: Test accuracy (%) with standard error. Each experiment is repeated over 5 random seeds.

Method	**Waterbirds (CLIP ViT-B/32)**		**CelebA (CLIP ViT-B/32)**		**MultiNLI (BERT)**	
	Average	**Worst-Group**	**Average**	**Worst-Group**	**Average**	**Worst-Group**
HGP	90.47 ± 0.06	86.48 ± 0.12	75.14 ± 0.12	71.68 ± 0.18	80.72 ± 0.62	69.35 ± 0.68
CMA	90.11 ± 0.17	86.16 ± 0.29	77.87 ± 0.04	74.16 ± 0.10	81.30 ± 0.25	69.72 ± 0.66

H.4.2 Fine-tuning

We also run fine-tuning experiments on HGP, strictly following the implementation in the code released by Hemati et al. (2023). The hyperparameter search scheme in DomainBed leads to more uncertainty and the implementation of HGP in Hemati et al. (2023) does not employ penalty annealing. Together with the differences in the Hessian computation, all of these factors potentially lead to the differences in the performance of CMA and HGP.

Table 10: Model selection: test-domain validation set

Algorithm	**ColoredMNIST**	**RotatedMNIST**	**VLCS**	**PACS**	**TerraIncognita**	**Avg**
HGP	55.8 ± 0.2	97.8 ± 0.1	76.5 ± 1.2	79.8 ± 0.2	29.6 ± 0.9	67.9
CMA	62.5 ± 0.9	97.9 ± 0.1	77.4 ± 0.8	81.6 ± 0.3	38.4 ± 1.2	71.5

Table 11: Model selection: training-domain validation set

Algorithm	**ColoredMNIST**	**RotatedMNIST**	**VLCS**	**PACS**	**TerraIncognita**	**Avg**
HGP	51.8 ± 0.0	97.9 ± 0.1	75.8 ± 1.0	77.5 ± 1.0	28.6 ± 0.8	66.3
CMA	51.4 ± 0.0	97.8 ± 0.0	77.9 ± 0.1	80.5 ± 0.2	34.0 ± 0.7	68.3