CHAOS IS A LADDER: DOMAIN GENERALIZATION BEYOND INVARIANCE VIA REWEIGHTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Domain generalization (DG) aims to learn from multiple source domains and generalize to unseen target domains. Most DG methods pursue invariance: they seek a causal representation whose prediction rule is invariant across domains. This principle is effective when the causal mechanism is stable, but becomes restrictive when the domain itself modulates how causal content maps to the response. In this case, directly feeding domain style into the predictor can create misleading shortcuts, since style does not by itself cause the response. Yet the apparent chaos of multiple styles can become a ladder: style can locate the unseen target domain among source domains and guide which domain-dependent prediction rules should be trusted. We propose *Latent Adaptive Domain Disentanglement and Environment Reweighting* (LADDER), a fixed-model DG pipeline that learns causal/style representations, freezes the encoders, fits source-specific classifiers, and uses an unlabeled target-domain covariate set only at inference to compute weights over these fixed classifiers, with no target labels or model-state updates. We establish theoretical guarantees for source reweighting and validate LADDER on simulations, FMoW, and a location-grouped iWildCam protocol, with gains in overall and group-averaged accuracy.

1 INTRODUCTION

Domain generalization (DG) aims to learn predictors from multiple source domains that generalize to an unseen target domain (Blanchard et al., 2011; Zhou et al., 2022). This challenge arises in many applications, where the deployment environment may differ from the training environments in background, acquisition condition, geographic region, time, sensor, or population composition (Rothenhäusler et al., 2021). For example, iWildCam images across camera-trap locations share animal categories but vary in habitat and species prevalence (Beery et al., 2020). In this setting, empirical risk minimization (ERM) over pooled source data learns a single predictor across domains, which can be suboptimal when the target domain follows a different prediction rule. The challenge is therefore both to suppress spurious domain-specific correlations and to determine when domain information should guide prediction in an unseen domain (Zhu et al., 2025).

A dominant line of DG research addresses this challenge through invariance. The aim is to learn a causal representation whose relationship with the label remains invariant across source domains and transfers to unseen domains. In causal language, this line often aims to extract a domain-invariant causal representation, denoted by $\boldsymbol{Z}_c$, while removing the domain-specific style representation, denoted by $\boldsymbol{Z}_s^{(e)}$ for domain e (Peters et al., 2016; Rothenhäusler et al., 2021). Representative approaches include invariant risk minimization and closely related invariance regularizers (Arjovsky et al., 2019; Krueger et al., 2021; Ahuja et al., 2021; Rame et al., 2022). Although these methods differ in implementation, they share the same target: style representation is treated mainly as a nuisance factor, and the purpose of modeling it is to learn a better invariant causal representation. This influential view is appropriate when the prediction rule is invariant across domains.

However, the invariance assumption can be too restrictive for broader DG tasks. Recent studies question whether invariant methods consistently outperform strong ERM baselines, especially on realistic benchmarks with complex source-target shifts (Gulrajani & Lopez-Paz, 2021; Zhu et al., 2025). One reason is that the usual invariant formulation does not distinguish per-sample style features from domain-level style distributions. Figure 1 illustrates this distinction. The per-sample style

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

representation $\boldsymbol{Z}_s^{(e)}$ may not directly cause the label $Y^{(e)}$, and therefore feeding style into a pooled classifier can create shortcuts that impair generalization. At the same time, the conditional prediction rule η_e, i.e. $Y^{(e)} \sim \eta_e(\cdot \mid \boldsymbol{Z}_c)$, may vary across domains, so a single predictor may be insufficient. The covariate $\boldsymbol{X}^{(e)}$ is generated from both the causal representation $\boldsymbol{Z}_c$ and the style representation $\boldsymbol{Z}_s^{(e)}$, while the response $Y^{(e)}$ depends on $\boldsymbol{Z}_c$ through a domain-dependent rule. The invariant-rule setting is recovered as a special case when this $e \to Y^{(e)}$ path collapses to a common prediction rule across all domains. In this general case, domain-level style distributions guide the selection of domain-dependent prediction rules, with causal content providing the per-sample predictor input. For example, an iWildCam cloudy-forest target domain may have a prediction rule closer to that of a forest-like source domain, without using style as a direct label shortcut.

This perspective connects to work retaining useful domain-specific information. Environment-adaptive covariate selection studies when apparently spurious correlations may help, while structured DG exploits known domain groups and low-rank parameter structure (Zuo & Wang, 2026; Li & Zhang, 2023). Domain adaptation and test-time adaptation include approaches that update model parameters using target data (Zhang, 2021; Xiao & Snoek, 2024). In our DG protocol, unlabeled target covariates determine weights over source classifiers while all fitted model parameters remain fixed. Closer to our mechanism, Best Sources Forward uses per-sample source-membership probabilities to fuse source-specific classifiers; DAEL collaboratively trains source heads; Gated Domain Units form observation-dependent ensembles; and DSNR maps an unlabeled target distribution to a domain-specific nonparametric predictor (Mancini et al., 2018; Zhou et al., 2021; Foll et al., 2023; Zheng et al., 2026). We propose *Latent Adaptive Domain Disentanglement and Environment Reweighting* (LADDER), which learns causal/style encoders, fits source classifiers, caches source style distributions, and computes one Sinkhorn-KNN weight vector from an unlabeled target set. LADDER routes through style-distribution distances; standard MoE uses a learned routing gate (Shazeer et al., 2017).

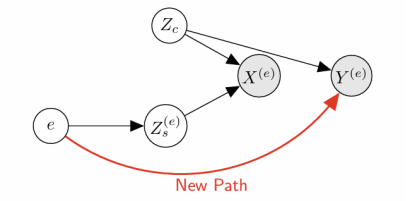

Figure 1: General structural causal model. For domain index e, $e \to Y^{(e)}$ denotes variation in $\eta_e(Y \mid \boldsymbol{Z}_c)$. LADDER uses domain-level style summaries for source weighting, not style as a per-sample classifier input.

Our contributions are three-fold. First, we propose LADDER, which estimates the target prediction rule by locally combining source-specific classifiers according to the target style distribution. Second, we establish theoretical guarantees for source reweighting and quantify how source coverage, finite-sample estimation, and representation error affect target prediction. Third, we validate LADDER on simulations and two real-world DG benchmarks, showing strong overall and group-averaged accuracy under shared-stack routing controls.

2 Methodology

Figure 2 illustrates LADDER training and inference. Source domains are encoded into causal representations for source-specific classifiers and style representations for source-domain fingerprints. At prediction, the fixed style encoder maps unlabeled target covariates to a style summary, compares it with stored source fingerprints, and changes only weights over fixed source classifiers.

2.1 Problem Set-up

We consider a DG problem in which source domains are indexed by $e \in \{1, \ldots, E\}$ and an unseen target domain is indexed by 0. Each domain induces a joint distribution P_e over observations and labels $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$. At training time we observe labeled samples $\mathcal{D}^{(e)} = \{(\boldsymbol{x}_i^{(e)}, y_i^{(e)})\}_{i=1}^{n_e}$ from the source domains. At inference, an unlabeled target covariate set $\mathcal{D}^{(0)} = \{\boldsymbol{x}_i^{(0)}\}_{i=1}^{n_0}$ is used only to estimate a style summary and source weights; encoders, classifiers, and normalization statistics remain fixed. Following the meta-distribution view of DG (Blanchard et al., 2011; Zhou

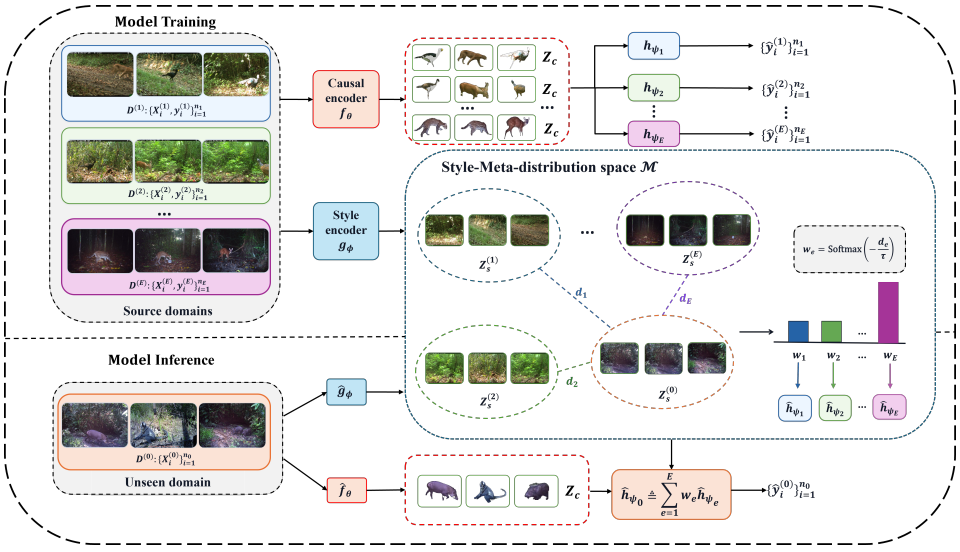

Figure 2: Overview of LADDER.

et al., 2022), we treat domains as points in a style-distribution space $\mathcal{M}$ and let $\Pi_{\mathcal{M}}$ denote their meta-distribution.

The key structural assumption is that each observation admits two learned roles, denoted $\boldsymbol{Z}_c$ and $\boldsymbol{Z}_s^{(e)}$. The causal and style representations are optimized for label prediction and source reweighting respectively; this division requires no statistical independence between representations or between style and labels. For a domain e, let $\nu_e \in \mathcal{P}(\mathcal{Z}_s)$ be its domain-level style distribution on the style space $\mathcal{Z}_s$; its empirical estimate $\widehat{\nu}_e$ is the style fingerprint used below. We write observations and labels abstractly as

$$\boldsymbol{X}^{(e)} = G(\boldsymbol{Z}_c, \boldsymbol{Z}_s^{(e)}, \boldsymbol{\varepsilon}_x), \qquad Y^{(e)} \sim \eta_e(\cdot \mid \boldsymbol{Z}_c),$$

where $\boldsymbol{\varepsilon}_x$ denotes noise, G is a shared observation mechanism, and η_e is the domain-dependent conditional prediction rule.

The smoothness principle underlying LADDER is that domains close in style distribution should have similar prediction rules:

$$d_{\mathcal{F}}(\eta_e, \eta_{e'}) \le L_\eta \, \mathcal{W}_1(\nu_e, \nu_{e'}), \tag{1}$$

where $d_{\mathcal{F}}$ measures task-level distance between conditional rules, $\mathcal{W}_1$ is the 1-Wasserstein distance, and L_η is a smoothness constant. In the parametric example, $d_{\mathcal{F}}$ is the parameter Frobenius distance.

A pooled predictor $\widehat{y} = h(f_{\boldsymbol{\theta}}(x), g_{\boldsymbol{\phi}}(x))$ may use style as a shortcut. LADDER computes source weights from distances between the target style fingerprint $\widehat{\nu}_0 = n_0^{-1} \sum_i \delta_{g_{\widehat{\boldsymbol{\phi}}}(x_i^{(0)})}$ and cached source fingerprints, where $\delta_{\boldsymbol{s}}$ denotes a Dirac point mass. It averages source-classifier logits on $\boldsymbol{Z}_c$ before Softmax (Eq. (5)), yielding distribution-dependent model averaging (Arpit et al., 2022).

2.2 Model Training: Encoders and Classifiers

Following disentanglement-based DG and information-bottleneck learning (Alemi et al., 2017; Zhang et al., 2022), model training learns $\boldsymbol{z}_c = f_{\boldsymbol{\theta}}(\boldsymbol{x})$ for label prediction and $\boldsymbol{z}_s = g_{\boldsymbol{\phi}}(\boldsymbol{x})$ for source reweighting, then fits source classifiers. We optimize:

$$\min_{\Theta} \mathcal{J} = \mathcal{L}_{\text{cls}} + \lambda_s \mathcal{L}_{\text{sd}} + \lambda_a \mathcal{L}_{\text{cd}}^{\text{GRL}} + \lambda_o \mathcal{L}_{\text{orth}} + \lambda_r \mathcal{L}_{\text{reg}}. \tag{2}$$

Here Θ collects the representation-stage parameters and $\lambda_s, \lambda_a, \lambda_o, \lambda_r$ are fixed loss weights. Equation (2) combines auxiliary label classification, style-domain classification, adversarial suppression of domain information in $\boldsymbol{Z}_c$, cross-covariance orthogonality, and representation regularization. Let ρ_y denote the shared auxiliary label classifier, and let ρ_s and ρ_c denote the auxiliary domain classifiers on $\boldsymbol{Z}_s^{(e)}$ and $\boldsymbol{Z}_c$, respectively. Writing $\widehat{\mathbb{E}}_e$ for the empirical average over source domain e

and $\ell_{\mathrm{CE}}(\boldsymbol{u}, y) = -\log(\mathrm{Softmax}(\boldsymbol{u})_y)$ for cross-entropy, $\mathcal{L}_{\mathrm{cls}} = E^{-1}\sum_e \widehat{\mathbb{E}}_e \ell_{\mathrm{CE}}(\rho_y(f_{\boldsymbol{\theta}}(x)), y)$, $\mathcal{L}_{\mathrm{sd}} = E^{-1}\sum_e \widehat{\mathbb{E}}_e \ell_{\mathrm{CE}}(\rho_s(g_{\boldsymbol{\phi}}(x)), e)$, and $\mathcal{L}_{\mathrm{cd}}^{\mathrm{GRL}} = E^{-1}\sum_e \widehat{\mathbb{E}}_e \ell_{\mathrm{CE}}(\rho_c(R_{\lambda_a}(f_{\boldsymbol{\theta}}(x))), e)$, where R_{λ_a} is a gradient-reversal layer.

We use $\mathcal{L}_{\mathrm{orth}} = (d_c d_s)^{-1}\|\widehat{\mathrm{Cov}}(f_{\boldsymbol{\theta}}(x), g_{\boldsymbol{\phi}}(x))\|_F^2$ to discourage redundant encoding; d_c, d_s denote the branch dimensions. The generic regularizer $\mathcal{L}_{\mathrm{reg}}$ may, for example, be a variational-bottleneck Kullback–Leibler (KL) term $\sum_{r\in\{c,s\}} d_r^{-1} D_{\mathrm{KL}}(q_r(z_r \mid x)\|p_r)$, where q_r and p_r are the branch posterior and prior, but LADDER does not depend on this choice. We then freeze both encoders and fit each final source classifier on fixed causal representations by

$$\widehat{\boldsymbol{\psi}}_e \in \arg\min_{\boldsymbol{\psi}_e} \widehat{\mathbb{E}}_e \ell_{\mathrm{CE}}(h_{\boldsymbol{\psi}_e}(f_{\widehat{\boldsymbol{\theta}}}(x)), y).$$

2.3 Model Inference: Reweighted Ensemble

For each source domain $e = 1, \ldots, E$, cache the low-dimensional style fingerprint

$$\widehat{\nu}_e = \frac{1}{n_e}\sum_{i=1}^{n_e} \delta_{g_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}_i^{(e)})}. \tag{3}$$

After these fingerprints are cached, source covariates need not be stored. For a new target domain, construct only the target style summary $\widehat{\nu}_0$ from target covariates, then compute d_e for each source domain. Theory uses $d_e = \mathcal{W}_1(\widehat{\nu}_e, \widehat{\nu}_0)$; real-data runs use a balanced debiased Sinkhorn approximation with Euclidean ground cost (Appendix A.3). The rates exclude entropic and solver errors. Let $\widehat{\mathcal{N}}_K(0)$ be the K source domains with the smallest distances. The target-adaptive weights are

$$w_e = \frac{\exp(-d_e/\tau)}{\sum_{j\in\widehat{\mathcal{N}}_K(0)}\exp(-d_j/\tau)}, \quad \text{for every } e \in \widehat{\mathcal{N}}_K(0), \tag{4}$$

where $1 \le K \le E$ is the neighbor count and $\tau > 0$ is the temperature. Set $w_e = 0$ outside $\widehat{\mathcal{N}}_K(0)$. With few target covariates or a noisy target fingerprint, larger τ and source neighborhoods make the rule approach a global source-weighted average. Each source classifier outputs pre-softmax logits, and the final target logit averages these logits:

$$\widehat{\boldsymbol{\ell}}(\boldsymbol{x}) = \sum_{e\in\widehat{\mathcal{N}}_K(0)} w_e\, h_{\widehat{\boldsymbol{\psi}}_e}\big(f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x})\big). \tag{5}$$

Labels are predicted from $\mathrm{Softmax}(\widehat{\boldsymbol{\ell}}(\boldsymbol{x}))$; for generalized linear model (GLM) classifiers, Eq. (5) equals the parameter average analyzed below. The weights w_e are computed once from $\widehat{\nu}_0$ and shared within the target domain. Algorithm 1 summarizes the inference stage. Runtime and storage are reported in Appendix Tables 12 and 13.

Algorithm 1 LADDER Inference Stage

Require: Fixed $f_{\widehat{\boldsymbol{\theta}}}, g_{\widehat{\boldsymbol{\phi}}}$; cached source fingerprints $\{\widehat{\nu}_e\}_{e=1}^E$; source classifiers $\{h_{\widehat{\boldsymbol{\psi}}_e}\}_{e=1}^E$; target covariates $\mathcal{D}^{(0)}$; neighbor count K; temperature τ.

1: Encode target styles and form $\widehat{\nu}_0 = n_0^{-1}\sum_{i=1}^{n_0}\delta_{g_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}_i^{(0)})}$.
2: Compute style-summary distances d_e for $e = 1, \ldots, E$.
3: Select $\widehat{\mathcal{N}}_K(0)$ as the K nearest source domains.
4: Compute weights $\{w_e\}_{e\in\widehat{\mathcal{N}}_K(0)}$ by Eq. (4).
5: Predict with fixed source classifiers by Eq. (5).
6: **return** Target predictions $\{\widehat{y}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{D}^{(0)}\}$.

3 Theoretical Guarantees

With fixed causal/style encoders and source classifiers, we analyze how well an unlabeled target style summary selects nearby source rules. The oracle analysis uses oracle representations; the generic analysis uses learned ones. Dimensions $d_c, d_s, d_{\mathcal{M}}$ denote causal, style, and style-manifold dimensions; $C = |\mathcal{Y}|$, $d_\beta = (C-1)d_c$, $n_{\min} := \min_{1\le e\le E} n_e$, and $a \vee b = \max\{a, b\}$.

3.1 Oracle Representation Guided Theory

For domain e, use the style distribution ν_e defined in Section 2.1, and let $\boldsymbol{\beta}_*^{(e)}$ be the oracle source parameter. For the oracle analysis, labels obey $\mathbb{P}(Y \mid \boldsymbol{Z}_c, e) = \mathrm{Softmax}(\boldsymbol{\beta}_*^{(e)} \boldsymbol{Z}_c)$ with $\boldsymbol{\beta}_*^{(e)} \in \mathbb{R}^{C \times d_c}$. For observation i in domain e, let $\boldsymbol{z}_{c,i}^{(e)}$ and $\boldsymbol{z}_{s,i}^{(e)}$ denote its oracle causal and style features. Set $\widehat{\nu}_e = n_e^{-1} \sum_i \delta_{\boldsymbol{z}_{s,i}^{(e)}}$; for $e = 1, \ldots, E$, $\widehat{\boldsymbol{\beta}}^{(e)} \in \arg\min_{\boldsymbol{\beta} \in \mathcal{B}} n_e^{-1} \sum_i \ell_{\mathrm{CE}}(\boldsymbol{\beta} \boldsymbol{z}_{c,i}^{(e)}, y_i^{(e)})$ is the source empirical-risk minimizer. We state compact assumptions here and give their full versions in the appendices.

Assumption 1 (Oracle latent structure). *Observations admit a centered linear oracle decomposition, $\boldsymbol{X}^{(e)} = \boldsymbol{A}_c \boldsymbol{Z}_c^{(e)} + \boldsymbol{A}_s \boldsymbol{Z}_s^{(e)} + \boldsymbol{\epsilon}^{(e)}$. The causal/noise covariances are domain-invariant, while $\boldsymbol{\Lambda}_e = \mathrm{Cov}(\boldsymbol{Z}_s^{(e)})$ is diagonal with jointly separating normalized cross-domain variance profiles.*

Assumption 2 (Smooth geometry). *There is an L_β-Lipschitz map Ψ^* from domain-level style distributions to GLM parameters, so that $\boldsymbol{\beta}_*^{(e)} = \Psi^*(\nu_e)$ and nearby style distributions have nearby prediction rules under $\mathcal{W}_1$. Style laws are compactly supported; sources are i.i.d. from a regular $d_{\mathcal{M}}$-dimensional space, with the target sampled independently.*

Assumption 3 (GLM regularity). *Causal features satisfy $\|\boldsymbol{Z}_c\|_2 \le R_c$ almost surely, and GLM parameters lie in a compact convex set $\mathcal{B} \subset \{\boldsymbol{\beta} : \mathbf{1}^\top \boldsymbol{\beta} = \mathbf{0}\}$ with $\|\boldsymbol{\beta}\|_F \le R_\beta$ and relative-interior oracle optima. For $\mathcal{R}_e(\boldsymbol{\beta}) = \mathbb{E}[\ell_{\mathrm{CE}}(\boldsymbol{\beta} \boldsymbol{Z}_c, Y) \mid e]$, these population cross-entropy risks are uniformly μ-strongly convex, for some $\mu > 0$, on $\mathcal{B}$. Single-sample Hessians are uniformly bounded and L_H-Lipschitz in $\boldsymbol{\beta}$ over $\mathcal{B}$; and centered-logit risks are locally μ_ℓ-strongly convex on bounded comparison logits and globally L_ℓ-smooth.*

These compact assumptions summarize style-covariance profile separability, smooth rule variation over style distributions, finite-sample source estimation, and empirical Wasserstein concentration.

Theorem 1. *Under Assumption 1, let $\boldsymbol{W}_s$ be a population style-extraction matrix, $\widehat{\boldsymbol{Z}}_s^{(e)} = \boldsymbol{W}_s \boldsymbol{X}^{(e)}$, and $\boldsymbol{M} := \boldsymbol{W}_s \boldsymbol{A}_s$ be invertible. Suppose the extractor diagonalizes the domain-varying style covariance, i.e., $\boldsymbol{M} \boldsymbol{\Lambda}_e \boldsymbol{M}^\top$ is diagonal for every source domain $e = 1, \ldots, E$. Then $\boldsymbol{M} = \boldsymbol{P}\boldsymbol{D}$ for a permutation matrix $\boldsymbol{P}$ and nonsingular diagonal matrix $\boldsymbol{D}$, and $\widehat{\boldsymbol{Z}}_s^{(e)} = \boldsymbol{B} \boldsymbol{Z}_c^{(e)} + \boldsymbol{P}\boldsymbol{D}\boldsymbol{Z}_s^{(e)} + \widetilde{\boldsymbol{\epsilon}}^{(e)}$, where $\boldsymbol{B} = \boldsymbol{W}_s \boldsymbol{A}_c$ and $\widetilde{\boldsymbol{\epsilon}}^{(e)} = \boldsymbol{W}_s \boldsymbol{\epsilon}^{(e)}$. Thus $\boldsymbol{W}_s \boldsymbol{A}_s$ is identified up to permutation/scaling, while the extracted representation may contain causal leakage. Moreover, under supplemental mean-separation/completeness conditions, $\boldsymbol{B} = 0$.*

Theorem 2 (Target Parameter Error). *Under Assumptions 2–3, and $n_{\min} \gtrsim \mu^{-2}\{d_\beta \log(1 + R_\beta L_H/\mu) + \log(d_\beta E)\}$ with $d_\beta = (C-1)d_c$, the reweighted estimator $\widehat{\boldsymbol{\beta}}^{(0)} = \sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e \widehat{\boldsymbol{\beta}}^{(e)}$ satisfies*

$$\mathbb{E}\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\|_F \le \mathcal{O}\left(\frac{R_c}{\mu} \sqrt{\frac{d_\beta + \log E}{n_{\min}}} + \gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s) + (K/E)^{1/d_{\mathcal{M}}} \right).$$

Here $\gamma(n, d) = n^{-1/(d \vee 2)} \ell_d(n)$ and $\Gamma(n, E, d) = ((1 + \log E)/n)^{1/(d \vee 2)} \widetilde{\ell}_d(n)$, where $\ell_2(n) = \widetilde{\ell}_2(n) = \log(1 + n)$ and $\ell_d(n) = \widetilde{\ell}_d(n) = 1$ for $d \neq 2$ (Fournier & Guillin, 2015).

Remark 1 (Source coverage trade-off). *Theorem 2's terms have three direct roles: $\frac{R_c}{\mu}\sqrt{(d_\beta + \log E)/n_{\min}}$ comes from fitting fixed source classifiers, $\gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s)$ comes from estimating target/source fingerprints, and $(K/E)^{1/d_{\mathcal{M}}}$ measures the distance from the target style law to the selected source neighborhood. More sources reduce the interpolation term when $n_{\min}$ is large enough to keep the worst source fingerprint reliable; otherwise, empirical distribution error dominates. For fixed dimensions, this bound vanishes when $n_0 \to \infty$, $E \to \infty$, $K/E \to 0$, and $(1 + \log E)\widetilde{\ell}_{d_s}(n_{\min})^{d_s \vee 2} = o(n_{\min})$.*

Theorem 3 (Target Excess Risk). *Under the assumptions of Theorem 2,*

$$\mathbb{E}\{\mathcal{R}_0(\widehat{\boldsymbol{\beta}}^{(0)}) - \mathcal{R}_0(\boldsymbol{\beta}_*^{(0)})\} \le \mathcal{O}\left(R_c^2 \left[\frac{R_c}{\mu} \sqrt{\frac{d_\beta + \log E}{n_{\min}}} + \gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s) + (K/E)^{1/d_{\mathcal{M}}} \right]^2 \right).$$

Theorem 3 turns the parameter bound into excess risk: bounded causal features scale logit error by R_c, and local smoothness gives a quadratic rate in the sum of the source-estimation, empirical-distribution, and interpolation terms.

To compare LADDER with PooledJoint and a fixed prediction rule motivated by IRM's invariant-prediction principle (Arjovsky et al., 2019), we consider a special case in which a scalar coordinate $c \sim \pi_c$ represents latent domain position. Define

$$f^*(\boldsymbol{Z}_c, c) = (\boldsymbol{\beta}_{\text{base}} + c\boldsymbol{\beta}_{\text{shift}})\boldsymbol{Z}_c, \qquad \boldsymbol{\Pi}_C = \boldsymbol{I}_C - C^{-1}\mathbf{1}\mathbf{1}^\top,$$

and for any predictor f, $\mathcal{R}_c(f) = \mathbb{E}[\ell_{\text{CE}}(f(\boldsymbol{Z}_c, \boldsymbol{Z}_s), Y_c)]$, where $Y_c \sim \text{Softmax}(f^*(\boldsymbol{Z}_c, c))$. The bounded additive pooled class is

$$\mathcal{H}_{\text{pool}}(R_W) = \{\, f = \boldsymbol{W}_c\boldsymbol{Z}_c + \boldsymbol{W}_s\boldsymbol{Z}_s : \|\boldsymbol{W}_c\|_F, \|\boldsymbol{W}_s\|_F \le R_W \,\}.$$

For target coordinate c_0, write $f_{\text{inv}} = f^*(\cdot, c_{\text{inv}})$ for fixed prediction, where c_{inv} does not depend on the target domain and satisfies $c_{\text{inv}} \neq c_0$. Write $f_{\text{L}} = f^*(\cdot, \widehat{c}_{\text{L}})$ for oracle LADDER, where $\widehat{c}_{\text{L}} = \sum_e w_e c_e$ and c_e is the coordinate of source e.

Assumption 4 (Comparison shift model). *Let c be a bounded scalar domain coordinate with $\text{Var}_{\pi_c}(c) > 0$. The causal representation is centered with $\mathbb{E}\boldsymbol{Z}_c\boldsymbol{Z}_c^\top = \boldsymbol{\Sigma}_c \succ 0$ and is independent of $(c, \boldsymbol{Z}_s)$. For the target coordinate, $c_{\text{inv}} \neq c_0$.*

Corollary 1 (Pooled-joint misspecification). *Under Assumptions 3 and 4,*

$$\inf_{f \in \mathcal{H}_{\text{pool}}(R_W)} \mathbb{E}_{c\sim\pi_c}\{\mathcal{R}_c(f) - \mathcal{R}_c(f^*)\} \ge \frac{\mu_\ell}{2}\,\text{Var}_{c\sim\pi_c}(c)\|\boldsymbol{\Pi}_C\boldsymbol{\beta}_{\text{shift}}\boldsymbol{\Sigma}_c^{1/2}\|_F^2.$$

Under nontrivial rule variation, every additive pooled predictor has positive domain-averaged excess risk: a style-dependent offset cannot modulate the causal decision matrix.

Corollary 2 (Comparison with fixed prediction). *Under Assumptions 3 and 4, if $\|\boldsymbol{\Pi}_C\boldsymbol{\beta}_{\text{shift}}\boldsymbol{\Sigma}_c^{1/2}\|_F > 0$, then for target coordinate c_0,*

$$\frac{\mathcal{R}_{c_0}(f_{\text{L}}) - \mathcal{R}_{c_0}(f^*)}{\mathcal{R}_{c_0}(f_{\text{inv}}) - \mathcal{R}_{c_0}(f^*)} \le \frac{L_\ell}{\mu_\ell}\frac{(c_0 - \widehat{c}_{\text{L}})^2}{(c_0 - c_{\text{inv}})^2}.$$

Consequently, within this comparison model, LADDER *has lower target excess risk than the fixed predictor whenever $|c_0 - \widehat{c}_{\text{L}}| < \sqrt{\mu_\ell/L_\ell}|c_0 - c_{\text{inv}}|$.*

Corollary 2 gives a sufficient condition for oracle source reweighting to improve on the fixed rule (Peters et al., 2016; Rosenfeld et al., 2021). Under the consistency regime in Remark 1, Theorem 3 gives vanishing expected excess risk for the estimated predictor, whereas this fixed rule retains positive excess risk when $c_{\text{inv}} \neq c_0$ and the centered rule shift is nonzero.

3.2 Generic Representation Guided Theory

We next quantify representation error in the parameter and excess risk bounds.

After a fixed coordinate alignment, for $e = 0, \ldots, E$ and $i = 1, \ldots, n_e$, write the learned representations as $\widetilde{\boldsymbol{z}}_{c,i}^{(e)} = \boldsymbol{z}_{c,i}^{(e)} + \boldsymbol{\xi}_{c,i}^{(e)}$ and $\widetilde{\boldsymbol{z}}_{s,i}^{(e)} = \boldsymbol{z}_{s,i}^{(e)} + \boldsymbol{\xi}_{s,i}^{(e)}$, where $\boldsymbol{\xi}_{c,i}^{(e)}$ and $\boldsymbol{\xi}_{s,i}^{(e)}$ are representation errors.

Assumption 5 (Second-moment representation error). *For a representation-error scale $\epsilon_{\text{rep}} \ge 0$, uniformly over domains and observations, $\mathbb{E}\|\boldsymbol{\xi}_{c,i}^{(e)}\|_2^2 \vee \mathbb{E}\|\boldsymbol{\xi}_{s,i}^{(e)}\|_2^2 \le \epsilon_{\text{rep}}^2$. For some $\mu_{\text{rep}} > 0$, each perturbed source empirical risk is μ_{rep}-strongly convex on $\mathcal{B}$. For a fresh target observation, conditional on the fitted objects, oracle representations, and representation errors, the label law remains $\text{Softmax}(\boldsymbol{\beta}_*^{(0)}\boldsymbol{Z}_c^{(0)})$.*

Write $\widetilde{\mathcal{R}}_0$ for target cross-entropy risk on learned causal features, evaluated on a fresh observation, and $\mathcal{R}_0$ for the corresponding risk on oracle features. Set $L_{\nabla z} = \sqrt{2} + \frac{1}{2}R_\beta R_c$.

Corollary 3 (Generic Representation Bound). *Under the assumptions of Theorem 2 and Assumption 5, let E be fixed. For $e = 1, \ldots, E$, let $\widetilde{\boldsymbol{\beta}}^{(e)}$ minimize the perturbed empirical cross-entropy*

risk on $\mathcal{B}$, and let $\widetilde{\nu}_e = n_e^{-1}\sum_{i=1}^{n_e}\delta_{\widetilde{\boldsymbol{z}}_{s,i}^{(e)}}$ for $e = 0,\dots,E$. Compute $\widetilde{\mathcal{N}}_K(0)$ and weights $\widetilde{w}_e$ from these measures by Eq. (4), and set $\widetilde{\boldsymbol{\beta}}^{(0)} = \sum_{e\in\widetilde{\mathcal{N}}_K(0)}\widetilde{w}_e\widetilde{\boldsymbol{\beta}}^{(e)}$. Then the parameter error satisfies

$$\mathbb{E}\|\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\|_F \le \mathcal{O}\Bigg(\frac{R_c}{\mu}\sqrt{\frac{d_\beta + \log E}{n_{\min}}} + \gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s) + (K/E)^{1/d_{\mathcal{M}}} + \left(\frac{L_{\nabla z}}{\mu_{\text{rep}}} + L_\beta\right)\epsilon_{\text{rep}}\Bigg),$$

and the target excess risk satisfies

$$\mathbb{E}\{\widetilde{\mathcal{R}}_0(\widetilde{\boldsymbol{\beta}}^{(0)}) - \mathcal{R}_0(\boldsymbol{\beta}_*^{(0)})\} \le \mathcal{O}\Bigg(R_c^2\Bigg[\frac{R_c}{\mu}\sqrt{\frac{d_\beta + \log E}{n_{\min}}} + \gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s) + (K/E)^{1/d_{\mathcal{M}}}\Bigg]^2 + \left[R_c^2\left(\frac{L_{\nabla z}}{\mu_{\text{rep}}} + L_\beta\right)^2 + R_\beta^2\right]\epsilon_{\text{rep}}^2\Bigg).$$

For fixed E, representation error enters linearly in parameter error and quadratically in risk; as $\epsilon_{\text{rep}} \to 0$, the oracle bounds are recovered, with the source-coverage term retained. The growing-source consistency regime is given in Remark 1.

4 Experiments

We evaluate LADDER in simulations, FMoW-WILDS, and iWildCam-WILDS. Simulation results with centered parameter errors and ablations appear in Appendix A.2.1; the main text focuses on real-data evidence. We compare with source-trained baselines ERM, IRM (Arjovsky et al., 2019), EQRM (Eastwood et al., 2022), DSNR (Zheng et al., 2026), RDM (Nguyen et al., 2024), and ARM-CML (Zhang et al., 2021); LADDER routing ablations (uniform weighting, $K = 1$, sample gating, and, on iWildCam, group gating); and, in the simulations, PooledJoint and an oracle target classifier. Baseline implementations, target-summary protocols, and metrics are detailed in Appendix A.

4.1 Real Data 1: FMoW-WILDS

FMoW-WILDS (Koh et al., 2021; Christie et al., 2018) is a 62-way satellite land-use classification benchmark built from Functional Map of the World images, where distribution shifts arise across time and geographic regions. We index source domains by year and test whether style-based reweighting improves overall OOD accuracy and regional robustness. Metrics are test accuracy, worst-/average-region accuracy (Worst/Avg.), and ID-to-test drop.

Table 1: FMoW-WILDS: mean ± SD over five seeds (accuracies in %; Drop in percentage points).

Method	Test ↑	Worst ↑	Avg. ↑	Drop ↓
ERM	46.68 ± 0.96	30.68 ± 1.92	45.49 ± 1.30	9.14 ± 0.77
IRM	46.68 ± 0.54	29.06 ± 1.79	45.39 ± 0.74	8.88 ± 0.73
EQRM	46.85 ± 0.55	28.89 ± 1.61	45.15 ± 0.54	9.09 ± 0.71
DSNR	45.52 ± 0.87	27.03 ± 1.89	43.93 ± 0.87	9.78 ± 0.61
RDM	47.15 ± 1.19	29.56 ± 1.68	45.22 ± 1.36	9.27 ± 0.70
ARM-CML	47.89 ± 0.35	29.53 ± 1.49	46.32 ± 0.80	$\mathbf{8.56 \pm 0.20}$
LADDER (uniform)	49.65 ± 0.32	30.46 ± 1.24	48.55 ± 0.45	8.69 ± 0.22
LADDER ($K = 1$)	49.12 ± 0.53	30.71 ± 0.97	48.08 ± 0.65	8.70 ± 0.71
LADDER (sample gate)	49.98 ± 0.31	30.67 ± 1.18	48.67 ± 0.58	9.01 ± 0.25
LADDER	$\mathbf{50.19 \pm 0.34}$	$\mathbf{31.46 \pm 0.74}$	$\mathbf{49.05 \pm 0.54}$	8.87 ± 0.53

All methods use ImageNet-pretrained ResNet-50 (He et al., 2016; Deng et al., 2009) and five seeds. ERM, IRM, EQRM, and RDM predict each example independently of other target examples. DSNR estimates one domain index per evaluation split from unlabeled covariates. ARM-CML conditions predictions on a learned mean context from the same split. The LADDER ablations share its frozen encoders and source classifiers: LADDER (uniform) assigns equal weights, LADDER ($K = 1$) uses the nearest source, and LADDER (sample gate) uses the frozen style-domain head for per-sample source-membership weights. LADDER and its $K = 1$ variant use split-level Sinkhorn style summaries. Following the WILDS evaluator, worst-region accuracy excludes the Other bucket. We compute average-region accuracy over the same nonempty regions.

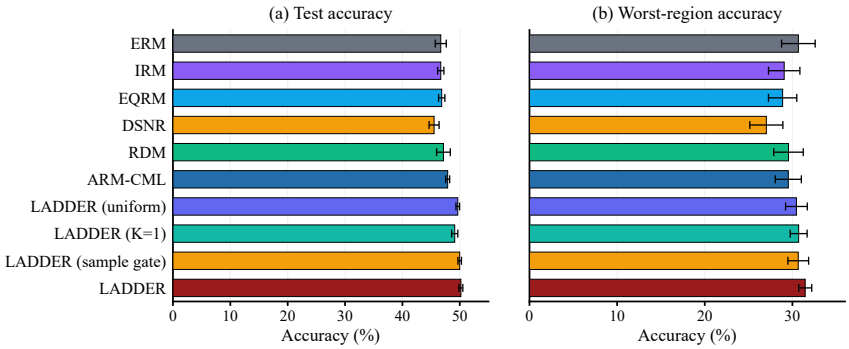

Figure 3: FMoW-WILDS. (a) Test accuracy and (b) worst-region accuracy: mean ± SD over five seeds.

Table 1 and Figure 3(a,b) show that the frozen representation and source-classifier stack improves FMoW over DG baselines. LADDER exceeds ARM-CML in test accuracy (50.19 versus 47.89), whereas ARM-CML has a smaller ID-to-test drop (8.56 versus 8.87). LADDER (uniform), LADDER (sample gate), and LADDER are numerically close. The small gains over sample gating have paired-seed intervals that include zero (Appendix Table 10); LADDER ($K = 1$) is weaker. Appendix Figure 12 visualizes style representations and Sinkhorn source reweighting. Appendix Figure 13 reports positive associations of style distance with classifier transfer loss and class-prototype distance. Full split-wise results, trade-off/source-weight plots, and t-SNE views appear in Appendix Figures 9, 10, and 12.

4.2 Real Data 2: iWildCam-WILDS

iWildCam-WILDS (Koh et al., 2021; Beery et al., 2020) stresses a different failure mode: the target split comprises heterogeneous camera-trap locations. It is a 182-way wildlife species classification benchmark collected from camera traps, with severe long-tailed class and location frequencies. This dataset fits source reweighting because locations may differ in background, illumination, species composition, and sensor style.

All methods use ImageNet-pretrained ResNet-50 (He et al., 2016; Deng et al., 2009), 64 source locations, class-balanced cross-entropy with parameter 0.999 (Cui et al., 2019), and five seeds. We report test accuracy, macro F1, and filtered average-group accuracy over the 35 locations with at least 50 samples; raw and filtered worst-location metrics remain nearly degenerate. ERM, IRM, EQRM, and RDM predict independently per example; DSNR estimates one domain index from unlabeled covariates for the full evaluation split. LADDER uses camera-location IDs to partition unlabeled target covariates into domains, constructing one Sinkhorn style summary per domain, also used by its $K = 1$ ablation; encoders and classifiers remain fixed. ARM-CML uses the same location groups to average its learned contexts. The LADDER (group gate) ablation averages per-sample source-membership weights within each location; LADDER (uniform) uses equal weights; LADDER (sample gate) computes weights per example. Appendix A.2.3 details summary construction and the single whole-split summary comparison.

Table 2: iWildCam-WILDS: mean ± SD over five seeds (all metrics in %).

Method	Test ↑	Macro F1 ↑	Filt. Avg. ↑
ERM	42.26 ± 5.48	11.23 ± 1.35	33.29 ± 7.61
IRM	34.59 ± 10.11	6.20 ± 1.08	23.42 ± 9.71
EQRM	40.75 ± 6.62	10.45 ± 1.37	32.22 ± 8.07
DSNR	47.36 ± 7.96	12.34 ± 1.12	39.44 ± 2.07
RDM	39.36 ± 7.67	8.97 ± 1.78	31.09 ± 5.37
ARM-CML	54.08 ± 1.03	$\mathbf{14.74 \pm 0.58}$	49.52 ± 0.84
LADDER (uniform)	44.42 ± 1.03	4.53 ± 0.25	34.13 ± 0.78
LADDER ($K = 1$)	57.96 ± 2.73	11.62 ± 0.90	48.41 ± 0.44
LADDER (sample gate)	58.00 ± 3.05	14.40 ± 0.70	48.16 ± 1.23
LADDER (group gate)	59.22 ± 3.10	12.00 ± 0.86	48.29 ± 1.98
LADDER	$\mathbf{61.71 \pm 1.76}$	11.61 ± 0.65	$\mathbf{50.01 \pm 0.84}$

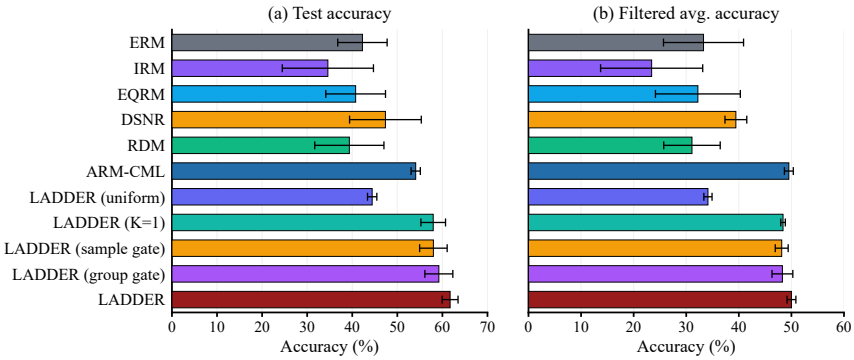

Figure 4: iWildCam-WILDS: (a) test and (b) filtered average-group accuracy (mean ± SD, five seeds). LADDER uses one target summary per camera location; its group gate ablation shares weights within each location. Whole-split summary comparison: Appendix A.2.3.

Table 2 and Figure 4(a,b) show that LADDER obtains the highest test and filtered average-group accuracy, numerically exceeding LADDER (group gate) by about 2.5 and 1.7 points, respectively; ARM-CML attains the highest macro F1 (14.74), while LADDER has higher test accuracy (61.71 versus 54.08) and slightly higher filtered average-group accuracy (50.01 versus 49.52). Together with the sample gate (14.40 macro F1), this indicates a trade-off between overall/location-averaged accuracy and class-balanced performance. The protocol-matched comparison with LADDER (group gate) supports distribution-level routing beyond location sharing. Appendix Figure 15 reports the overall versus filtered location-averaged accuracy trade-off and full routing comparison, including macro F1 and the whole-split summary.

5 Discussion

LADDER uses the target style distribution to combine nearby source prediction rules. Its effectiveness depends on whether style proximity reflects rule similarity and whether the sources adequately cover the target. Our empirical decomposition uses $\boldsymbol{Z}_c$ for prediction and $\boldsymbol{Z}_s^{(e)}$ for domain-level routing, without requiring statistical independence between representations. Future work could extend LADDER with source weighting based on uncertainty, extrapolation beyond convex source combinations, and representations from foundation models for richer visual and multimodal DG.

AI USE STATEMENT

Generative AI tools assisted with language editing, code scaffolding, and the review and revision of mathematical arguments and manuscript text. The authors reviewed the AI-assisted material and take responsibility for the final manuscript.

REFERENCES

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, 2021.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

Martin Arjovsky, Leon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019.

Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. In *Advances in Neural Information Processing Systems*, 2022.

Sara Beery, Elijah Cole, and Arvi Gjoka. The iWildCam 2020 competition dataset. In *Seventh Fine-Grained Visual Categorization Workshop at CVPR*, 2020.

Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer, 2015.

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*, 2011.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

Cian Eastwood, Alexander Robey, Julius von Kügelgen, Hamed Hassani, Shashank Singh, George J. Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. In *Advances in Neural Information Processing Systems*, 2022.

Simon Foll, Alina Dubatovka, Eugen Ernst, Siu Lun Chau, Martin Maritsch, Patrik Okanovic, Gudrun Thaeter, Joachim M. Buhmann, Felix Wortmann, and Krikamol Muandet. Gated domain units for multi-source domain generalization. *Transactions on Machine Learning Research*, 2023.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3–4):707–738, 2015. doi: 10.1007/s00440-014-0583-7.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5815–5826, 2021.

Sai Li and Linjun Zhang. Multi-dimensional domain generalization with low-rank structures, 2023.

Massimiliano Mancini, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Best sources forward: Domain generalization through source-specific nets. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 1353–1357. IEEE, 2018. doi: 10.1109/ICIP.2018.8451318.

Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman and Hall, 2 edition, 1989.

Toan Nguyen, Kien Do, Bao Duong, and Thin Nguyen. Domain generalisation via risk distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012, 2016.

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*, 2022.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

Dominik Rothenhäusler, Peter Bühlmann, Nicolai Meinshausen, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B*, 83 (2):215–246, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Zehao Xiao and Cees G. M. Snoek. Beyond model adaptation at test time: A survey, 2024.

Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P. Xing. Towards principled disentanglement for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Youshan Zhang. A survey of unsupervised domain adaptation for visual recognition, 2021.

Siming Zheng, Yuanyuan Lin, Yong Zhou, and Jian Huang. Domain-specific nonparametric regression for domain generalization. *Journal of the American Statistical Association*, 2026. doi: 10.1080/01621459.2026.2658866.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021. doi: 10.1109/TIP.2021.3112012.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Yilun Zhu, Naihao Deng, Naichen Shi, Aditya Gangrade, and Clayton Scott. Domain generalization: A tale of two ERMs, 2025. arXiv:2510.04441.

Shuozhi Zuo and Yixin Wang. Environment-adaptive covariate selection: Learning when to use spurious correlations for out-of-distribution prediction, 2026. arXiv:2601.02322.

Appendix A provides experimental details and additional results; Appendix B gives full assumptions and auxiliary results; Appendix C contains the proofs.

A Additional Experimental Details and Results

A.1 Dataset Details

FMoW-WILDS is a 62-class satellite land-use classification benchmark. The images come from the Functional Map of the World collection, and the WILDS split induces distribution shifts across acquisition years and geographic regions. In our experiments, years define source domains for source reweighting, and regions define the group metric used for robustness evaluation. The Other bucket is excluded from worst-region accuracy following the WILDS evaluator, and from our average-region statistic. Overall accuracy includes all examples.

For an evaluation split $\mathcal{S}$, test accuracy is the fraction of correctly classified examples. Let $\mathcal{R}$ be the retained FMoW regions after excluding "Other", and let Acc_r be the accuracy within region r. The region-aware metrics and distribution-shift drop are

$$\mathrm{Worst} = \min_{r\in\mathcal{R}} \mathrm{Acc}_r, \qquad \mathrm{Avg.} = \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}} \mathrm{Acc}_r, \qquad \mathrm{Drop} = \mathrm{Acc}_{\mathrm{ID\ val}} - \mathrm{Acc}_{\mathrm{OOD\ test}}\,.$$

Thus, Worst and Avg. are respectively the minimum and unweighted mean of the retained region accuracies, while ID-to-test drop measures the decrease from ID validation accuracy to OOD test accuracy.

iWildCam-WILDS is a 182-class wildlife species classification benchmark collected by camera traps. Domains are camera locations, and both class frequencies and location frequencies are highly long-tailed. This makes raw worst-location accuracy extremely brittle: a few target locations contain only a handful of examples, so a single error can dominate the raw worst-domain metric. Moreover, naturally tiny target locations are entangled with other effects, including unseen camera locations, severe class imbalance, and incomplete class support within a location. We therefore use filtered average-group accuracy for the main real-data robustness evaluation and reserve target-sample-size sensitivity for synthetic simulations where the target-summary sample size can be varied without changing the label support or the environment mechanism.

For iWildCam-WILDS, test accuracy is computed over the full evaluation split, macro F1 is the unweighted mean of the F1 scores over classes present in the evaluation labels, and filtered average-group accuracy is

$$\mathrm{FiltAvg} = \frac{1}{|\mathcal{G}_{50}|}\sum_{g\in\mathcal{G}_{50}} \mathrm{Acc}_g, \qquad \mathcal{G}_{50} = \{g : n_g \geq 50\},$$

where g indexes target camera locations, n_g is the number of evaluation examples at location g, and Acc_g is its accuracy.

Information-use boundary. In the benchmarks, the unlabeled summary covariates are taken from the evaluation split or camera location whose examples are predicted. The theoretical population risk evaluates a fresh observation from the same target domain after the summary has been formed.

The experiments use source-only training; all learned model parameters and source normalization statistics remain fixed during evaluation. For LADDER, this fixes both encoders and all source classifiers. At inference, LADDER encodes unlabeled target covariates $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^{n_0}$ into a domain-level style fingerprint $\widehat{\nu}_0$. Distances to cached source fingerprints determine the source-neighbor set and source-classifier weights. Test labels are used only to compute evaluation metrics. On iWildCam, location metadata additionally partitions unlabeled covariates for the group-level routing variants; predictions use image representations and the resulting source weights. Optimization is confined to source-data training.

Real-data domain construction and target summaries. On FMoW-WILDS, year metadata defines 11 retained source domains after source-domain filtering, and one unlabeled style summary is

formed for each evaluation split without using target-domain metadata. Overall accuracy is split-level; region metadata is used only for the region-aware metrics, which exclude "Other".

On iWildCam-WILDS, camera-location metadata defines 64 retained source domains. LADDER forms one unlabeled style summary per target camera location, treating each location as a target domain. Location IDs define the summary groups, and each image is classified using the weights assigned to its location. The alternative LADDER (split summary) forms one summary for the whole validation or test split and uses location metadata only for evaluation. Both variants follow the source-only training and fixed-model inference protocol above. Filtered average-group accuracy averages the target locations with at least 50 examples.

A.2 Baselines and Dataset-Specific Results

Reference baselines. ERM, IRM, EQRM, and RDM predict each example independently of the other target examples. DSNR conditions predictions on a domain index estimated from an unlabeled target set. In the simulation, we use a linear DSNR implementation with the empirical target-domain covariate mean as its index. On FMoW and iWildCam, the classification implementation averages the outputs of a source-trained index network over an unlabeled subset of each evaluation split and shares the resulting index across that split. Its trained parameters and normalization statistics remain fixed. The iWildCam location-grouping option used by LADDER does not change DSNR's split-level index.

ARM-CML (Zhang et al., 2021) learns a context network c_η jointly with a predictor h_θ through source-domain episodes. For an unlabeled target set $\mathcal{U}_g$, it uses

$$\bar{\boldsymbol{c}}_g = \frac{1}{|\mathcal{U}_g|} \sum_{\boldsymbol{x} \in \mathcal{U}_g} c_\eta(\boldsymbol{x}), \qquad \widehat{p}(\boldsymbol{y} \mid \boldsymbol{x}) = \mathrm{softmax}(h_\theta([\boldsymbol{x}, \bar{\boldsymbol{c}}_g])) \,.$$

Our adaptation follows the authors' context-network implementation, with source-trained parameters and batch-normalization statistics fixed at inference. The primary target sets are the evaluation split on FMoW, each camera location on iWildCam, and each target separately in the simulation.

LADDER routing ablations. We use four LADDER routing ablations in the real-data tables. LADDER (uniform) reuses the same frozen representation and source-classifier stack as LADDER but assigns equal weights,

$$\widehat{\boldsymbol{\ell}}_{\mathrm{uni}}(\boldsymbol{x}) = \frac{1}{E} \sum_{e=1}^{E} h_{\widehat{\boldsymbol{\psi}}_e}\left(f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x})\right) .$$

LADDER ($K = 1$) uses the same target summary and distance geometry as LADDER but performs hard $K = 1$ routing,

$$e^* = \arg\min_e d_e, \qquad \widehat{\boldsymbol{\ell}}_{\mathrm{near}}(\boldsymbol{x}) = h_{\widehat{\boldsymbol{\psi}}_{e^*}}\left(f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x})\right) .$$

LADDER (sample gate) is a BSF-style per-sample control that uses the frozen auxiliary style-domain classifier ρ_s to form source-membership weights,

$$\widehat{\boldsymbol{\ell}}_{\mathrm{sg}}(\boldsymbol{x}) = \sum_{e=1}^{E} \pi_e(\boldsymbol{x}) h_{\widehat{\boldsymbol{\psi}}_e}\left(f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x})\right) , \qquad \boldsymbol{\pi}(\boldsymbol{x}) = \mathrm{softmax}\Big(\rho_s(g_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))/T_{\mathrm{sg}}\Big) .$$

We fix $T_{\mathrm{sg}} = 1$ for all datasets and seeds. On iWildCam, LADDER (group gate) uses the same target grouping metadata as LADDER but averages LADDER (sample gate) weights within each target group,

$$\bar{\pi}_{g,e} = \frac{1}{|\mathcal{D}_g|} \sum_{\boldsymbol{x} \in \mathcal{D}_g} \pi_e(\boldsymbol{x}), \qquad \widehat{\boldsymbol{\ell}}_{\text{sg-group}}(\boldsymbol{x}) = \sum_{e=1}^{E} \bar{\pi}_{g,e} h_{\widehat{\boldsymbol{\psi}}_e}\left(f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x})\right) , \quad \boldsymbol{x} \in \mathcal{D}_g .$$

Table 5 evaluates learned summary-conditioned alternatives in the simulation.

A.2.1 SYNTHETIC SIMULATION

Data generation and metrics. Each repetition draws two matrices in $\mathbb{R}^{3\times 10}$ with independent entries $(\boldsymbol{\beta}_{\text{base}})_{kj} \sim \mathcal{N}(0, 0.6^2)$ and $(\boldsymbol{\beta}_{\text{shift}})_{kj} \sim \mathcal{N}(0, 0.4^2)$, shared across domains within that repetition. The nine source coordinates are $c_e = -5+1.25(e-1)$ for $e = 1, \ldots, 9$; the interpolation and extrapolation targets have coordinates 2 and 6. At coordinate c, the independent causal and style vectors follow

$$\boldsymbol{Z}_c \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{10}), \qquad \boldsymbol{Z}_s \sim \mathcal{N}(\mathbf{0}, \operatorname{diag}(3 + 0.4c,\ 3 - 0.4c,\ 64\boldsymbol{I}_6)) .$$

The observation is $\boldsymbol{X} = (\boldsymbol{Z}_c^\top, \boldsymbol{Z}_s^\top)^\top$, and $Y \mid \boldsymbol{Z}_c, c \sim \text{Categorical}(\text{Softmax}((\boldsymbol{\beta}_{\text{base}} + c\boldsymbol{\beta}_{\text{shift}})\boldsymbol{Z}_c))$. Each domain contains 1000 examples. The simulation supplies the latent features directly to LADDER, while the reference baselines receive $\boldsymbol{X}$; PooledJoint receives the concatenated latent vectors. Accuracy is the fraction of correct class predictions. For probabilities $\widehat{p}_{ik}$ on n target examples, the Brier score is $n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{3}(\widehat{p}_{ik} - \mathbf{1}\{y_i = k\})^2$. The parameter error for the effective causal classifier matrix is

$$\left\| \boldsymbol{\Pi}_C \left(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)} \right) \right\|_F , \qquad \boldsymbol{\Pi}_C = \boldsymbol{I}_C - C^{-1}\mathbf{1}\mathbf{1}^\top .$$

This removes the common class shift that leaves Softmax probabilities unchanged. Parameter error is evaluated after fitting; it does not alter training or predictions. Tables report means and population standard deviations across repetitions. The main comparison, KNN sweep, and additional pooled ablation use separate sets of 100 repetitions.

Table 3: Simulation results over 100 repetitions (mean ± standard deviation; accuracy in %). Param. is the centered parameter error; it is not defined for nonlinear ARM-CML.

Target A: interpolation, $c = 2$			
Method	Accuracy ↑	Brier ↓	Parameter error ↓
ERM	55.56 ± 4.85	0.553 ± 0.041	3.82 ± 0.74
IRM	45.91 ± 3.35	0.626 ± 0.018	4.15 ± 0.74
EQRM	54.38 ± 5.41	0.570 ± 0.039	3.95 ± 0.77
DSNR	55.47 ± 5.00	0.554 ± 0.041	3.82 ± 0.74
RDM	55.61 ± 4.56	0.555 ± 0.038	3.86 ± 0.76
PooledJoint	55.56 ± 4.86	0.553 ± 0.041	3.82 ± 0.74
ARM-CML	67.64 ± 5.52	0.505 ± 0.084	–
LADDER	$\mathbf{72.88 \pm 10.28}$	$\mathbf{0.379 \pm 0.112}$	$\mathbf{2.64 \pm 0.76}$
Oracle	80.80 ± 3.30	0.267 ± 0.042	0.00 ± 0.00
Target B: extrapolation, $c = 6$			
Method	Accuracy ↑	Brier ↓	Parameter error ↓
ERM	41.86 ± 4.92	0.673 ± 0.042	10.48 ± 1.83
IRM	38.68 ± 2.39	0.668 ± 0.009	10.59 ± 1.84
EQRM	41.26 ± 4.20	0.662 ± 0.032	10.52 ± 1.84
DSNR	41.59 ± 5.27	0.676 ± 0.044	10.48 ± 1.83
RDM	42.01 ± 3.44	0.664 ± 0.023	10.49 ± 1.85
PooledJoint	41.87 ± 4.93	0.673 ± 0.042	10.48 ± 1.83
ARM-CML	82.79 ± 2.98	0.273 ± 0.047	–
LADDER	$\mathbf{86.23 \pm 3.70}$	$\mathbf{0.234 \pm 0.037}$	$\mathbf{8.15 \pm 1.75}$
Oracle	91.80 ± 1.48	0.117 ± 0.020	0.00 ± 0.00

Figure 5 reports the controlled simulation results. The top and bottom rows separate the interpolation and extrapolation targets. This controlled SCM isolates the source-reweighting mechanism rather than the sufficient identifiability conditions in Section B. The style representation is 8-dimensional, there are 9 source domains, and each domain contains 1000 samples; the main LADDER setting uses $K = 4$ source neighbors. Two coordinates encode the environment coordinate through opposite covariance shifts, while the remaining coordinates are high-variance nuisance variation. In the main comparison, LADDER reaches mean accuracies of 72.88% and 86.23% on the interpolation and extrapolation targets, respectively. It has the highest mean accuracy and lowest mean Brier score

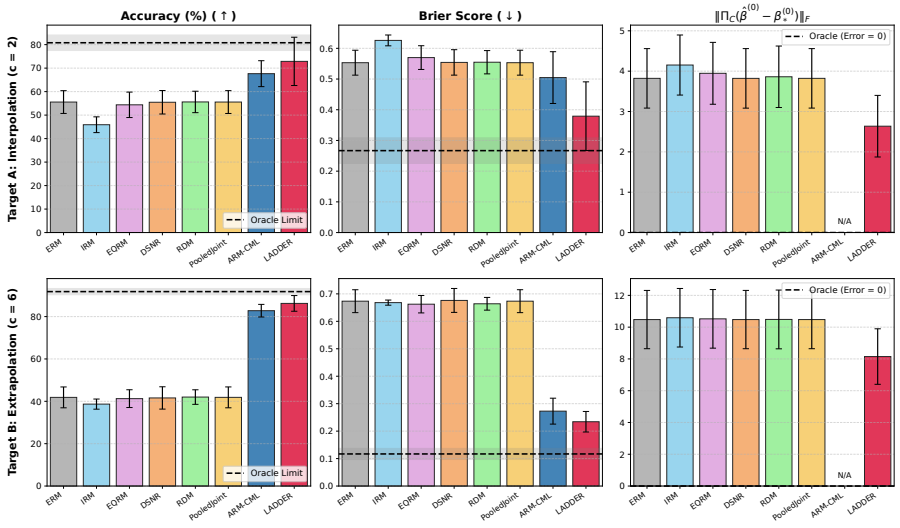

Figure 5: Simulation results over 100 independent repetitions. Bars and error bars show means and standard deviations; parameter errors use class-centered classifier matrices and are not defined for nonlinear ARM-CML.

among the main non-oracle methods on both targets, and the lowest centered parameter error among those with linear causal coefficients. ARM-CML reaches 67.64% and 82.79% accuracy, with Brier scores 0.505 and 0.273. Its 100 repetitions use the same data-generating mechanism but independent draws from the previously retained baselines.

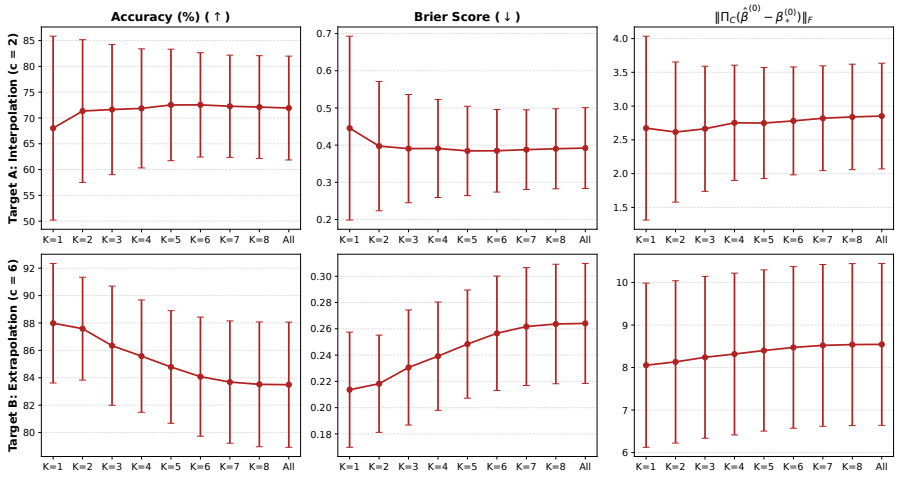

Figure 6: Synthetic simulation KNN source-selection ablation over $K \in \{1, \ldots, 8\}$ and "All". The main setting uses $K = 4$, while "All" removes the KNN truncation and reweights all available source classifiers.

Figure 6 studies the local source-selection radius under the same covariance-shift style protocol. For the interpolation target, mean accuracy ranges from 71.34% to 72.54% across $K = 2, \ldots, 8$ and the "All" variant. For extrapolation, $K = 1$ reaches 87.98%, compared with 83.49% when all sources are reweighted. Small neighborhoods focus on the boundary source domains closest to the target mechanism. We retain $K = 4$ as the main setting: it lies in the stable interpolation region and preserves local source selection for extrapolation.

The target-summary ablation uses five independent repetitions of data generation and source training. Within each repetition, it resamples subsets of the same 1000 target covariates, using 5000,

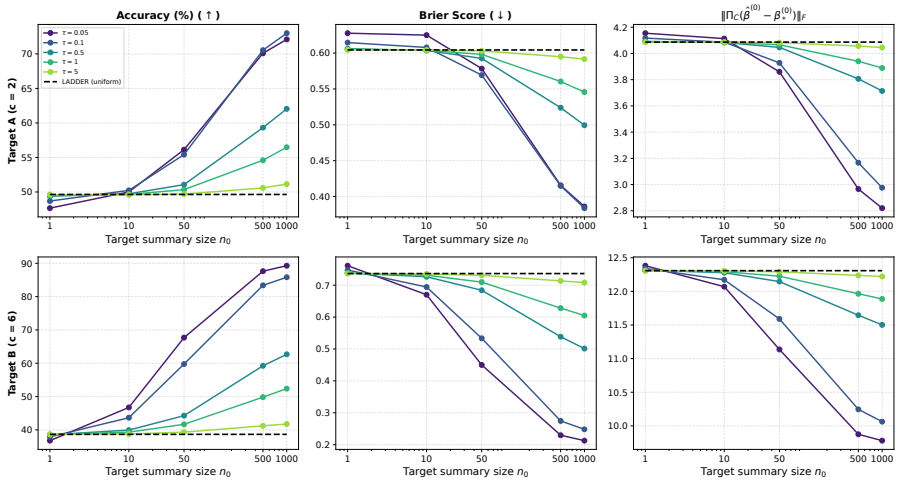

Figure 7: Synthetic simulation ablation for target-summary size and temperature $\tau \in \{0.05, 0.1, 0.5, 1, 5\}$. Rows correspond to interpolation and extrapolation targets; columns report accuracy, Brier score, and centered parameter error. Curves show mean results over target-summary resamples pooled across five training repetitions. LADDER (uniform) uniformly averages all source classifiers.

1000, 1000, 300, and 300 inference draws for $n_0 \in \{1, 10, 50, 500, 1000\}$, respectively. It uses all nine source classifiers, held fixed across inference draws, and varies $\tau \in \{0.05, 0.1, 0.5, 1, 5\}$. These inference draws are not additional training repetitions.

Figure 7 varies n_0 and τ while holding the source classifiers and domain mechanisms fixed. Two of the eight style coordinates carry environment information; the remaining six contain high-variance nuisance variation. At $n_0 = 1$, mean accuracy remains close to uniform source averaging on both targets; at $n_0 = 10$ and $\tau = 0.1$, it is already higher, with larger gains at larger n_0.

As n_0 increases, the target style summary becomes more informative for navigation. With $n_0 = 1000$ and $\tau = 0.1$, mean accuracies are 73.00% and 85.82% on the two targets, compared with 49.64% and 38.66% for uniform source averaging. Smaller temperatures help select the relevant source classifiers, especially for extrapolation. Thus τ should be interpreted as a temperature parameter controlling how strongly LADDER trusts empirical style-distribution distances: a large temperature such as $\tau = 5$ is a conservative fallback when target summaries are tiny, whereas smaller temperatures become useful once the target distribution is estimated with enough samples.

Table 4 compares the same fitted ARM-CML model using separate target contexts, a context pooled across both targets, and shuffled groups preserving group sizes and summary caps. Separate contexts improve extrapolation accuracy but reduce interpolation accuracy relative to pooled or shuffled contexts.

Table 4: ARM-CML simulation context controls (mean ± population SD, 100 repetitions).

Target	Context	Accuracy (%) ↑	Brier ↓
Interpolation	Split	73.87 ± 4.15	0.420 ± 0.066
Interpolation	True groups	67.64 ± 5.52	0.505 ± 0.084
Interpolation	Shuffled groups	73.99 ± 4.01	0.419 ± 0.065
Extrapolation	Split	77.14 ± 4.47	0.362 ± 0.072
Extrapolation	True groups	82.79 ± 2.98	0.273 ± 0.047
Extrapolation	Shuffled groups	77.24 ± 4.33	0.360 ± 0.070

Additional style-conditioned pooled baselines. The main simulation includes PooledJoint as an additive linear classifier on $(\boldsymbol{Z}_c, \boldsymbol{Z}_s^{(e)})$. This instantiates the additive pooled class in Corollary 1;

more expressive style-conditioned pooled predictors are examined below. To test whether the main simulation gap is caused only by this additive restriction, Table 5 and Figure 8 add three stronger pooled variants. Pooled-Interaction appends a domain-level style summary and explicit products between the style summary and $\boldsymbol{Z}_c$; Pooled-MLP uses a nonlinear pooled network on $\boldsymbol{Z}_c$, $\boldsymbol{Z}_s^{(e)}$, and the style summary; and StyleHyperNet maps the style summary to classifier parameters. All three baselines use the same source data and unlabeled target style samples as LADDER. Their domain summaries concatenate coordinate-wise means and log variances, with normalization statistics fitted on source domains. LADDER computes its quantile-based distances from the empirical target style samples.

Table 5: Supplementary style-conditioned pooled ablation in the controlled SCM over 100 repetitions. Accuracy is in %; Param. is the centered parameter error. Pooled-MLP has no single linear target classifier, so parameter error is not reported.

Target	Method	Acc. ↑	Brier ↓	Param. ↓
Interpolation, $c = 2$	PooledJoint	54.84 ± 4.49	0.561 ± 0.037	3.822 ± 0.735
	Pooled-Interaction	61.74 ± 19.60	0.534 ± 0.288	2.975 ± 1.734
	Pooled-MLP	58.21 ± 18.52	0.646 ± 0.331	—
	StyleHyperNet	64.46 ± 19.60	0.505 ± 0.300	2.709 ± 1.818
	LADDER	72.57 ± 11.01	0.383 ± 0.119	2.556 ± 0.998
	Oracle	80.86 ± 3.10	0.267 ± 0.041	0.000 ± 0.000
Extrapolation, $c = 6$	PooledJoint	41.94 ± 4.36	0.669 ± 0.034	10.559 ± 1.823
	Pooled-Interaction	85.01 ± 10.43	0.214 ± 0.138	5.769 ± 2.535
	Pooled-MLP	78.57 ± 11.01	0.322 ± 0.179	—
	StyleHyperNet	84.05 ± 11.60	0.230 ± 0.175	4.286 ± 2.454
	LADDER	85.74 ± 3.75	0.239 ± 0.036	8.258 ± 1.726
	Oracle	91.88 ± 1.66	0.117 ± 0.023	0.000 ± 0.000

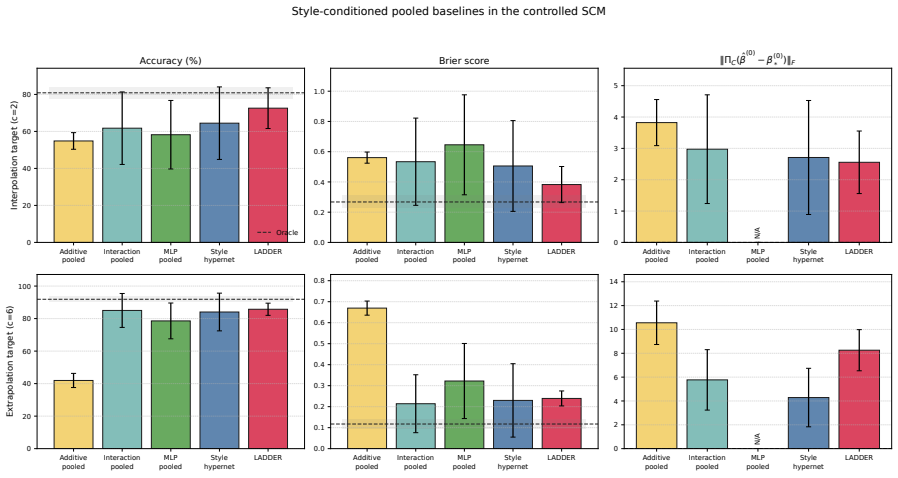

Figure 8: Supplementary style-conditioned pooled ablation in the controlled SCM. Stronger pooled predictors close much of the additive PooledJoint gap, especially in extrapolation, but they also have substantially larger variance across repetitions. LADDER remains competitive in extrapolation and gives the strongest interpolation performance among non-oracle methods.

LADDER gives the best mean interpolation accuracy, Brier score, and parameter recovery among non-oracle methods. In extrapolation, it has the highest mean accuracy and lower accuracy variability than the expressive pooled alternatives, while Pooled-Interaction and StyleHyperNet have lower mean Brier scores and parameter errors.

Table 6: FMoW-WILDS results over five random seeds. Accuracies are reported in %; ID-to-test accuracy drop is reported in percentage points.

Method	Test ↑	Worst ↑	Avg. ↑	Drop ↓
ERM	46.68 ± 0.96	30.68 ± 1.92	45.49 ± 1.30	9.14 ± 0.77
IRM	46.68 ± 0.54	29.06 ± 1.79	45.39 ± 0.74	8.88 ± 0.73
EQRM	46.85 ± 0.55	28.89 ± 1.61	45.15 ± 0.54	9.09 ± 0.71
DSNR	45.52 ± 0.87	27.03 ± 1.89	43.93 ± 0.87	9.78 ± 0.61
RDM	47.15 ± 1.19	29.56 ± 1.68	45.22 ± 1.36	9.27 ± 0.70
ARM-CML	47.89 ± 0.35	29.53 ± 1.49	46.32 ± 0.80	$\mathbf{8.56 \pm 0.20}$
LADDER (uniform)	49.65 ± 0.32	30.46 ± 1.24	48.55 ± 0.45	8.69 ± 0.22
LADDER ($K = 1$)	49.12 ± 0.53	30.71 ± 0.97	48.08 ± 0.65	8.70 ± 0.71
LADDER (sample gate)	49.98 ± 0.31	30.67 ± 1.18	48.67 ± 0.58	9.01 ± 0.25
LADDER	$\mathbf{50.19 \pm 0.34}$	$\mathbf{31.46 \pm 0.74}$	$\mathbf{49.05 \pm 0.54}$	8.87 ± 0.53

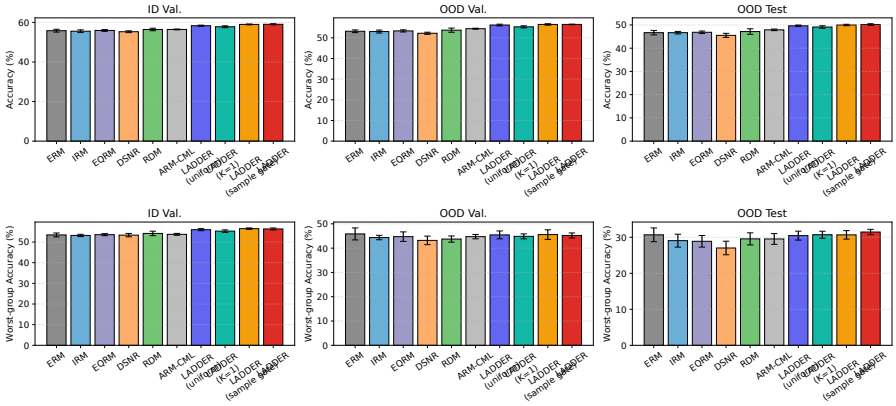

Figure 9: FMoW-WILDS accuracy and worst-region accuracy on the ID validation, OOD validation, and OOD test splits.

A.2.2 FMoW-WILDS

Figures 9 and 10 provide the FMoW-WILDS diagnostic view. They report both aggregate accuracy and region-aware robustness, together with the source-classifier weights used by LADDER for the representative target split. Figure 11 further isolates the effect of the KNN source-selection radius.

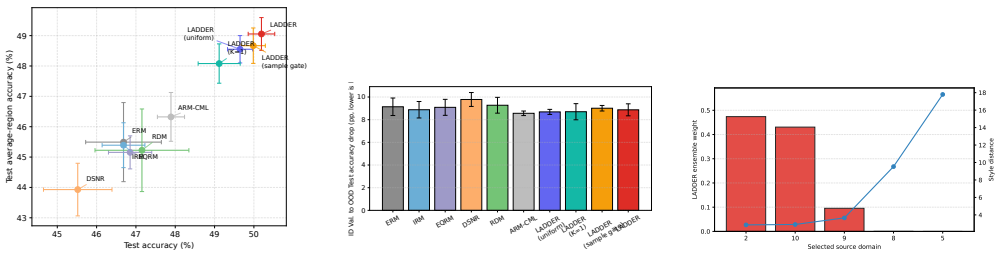

Figure 10: FMoW-WILDS accuracy/average-region trade-off, OOD degradation, and representative LADDER source-classifier weights.

The FMoW-WILDS KNN ablation supports the inference-stage reweighting mechanism used by LADDER; its $K = 1$ endpoint corresponds to hard nearest-source classifier routing, while the navigation ablation in simulation includes the uniform source-classifier average. Moving from the very local $K = 1$ setting to moderate neighborhoods improves both test accuracy and average-region accuracy, and the main setting $K = 5$ gives the strongest mean performance. Removing

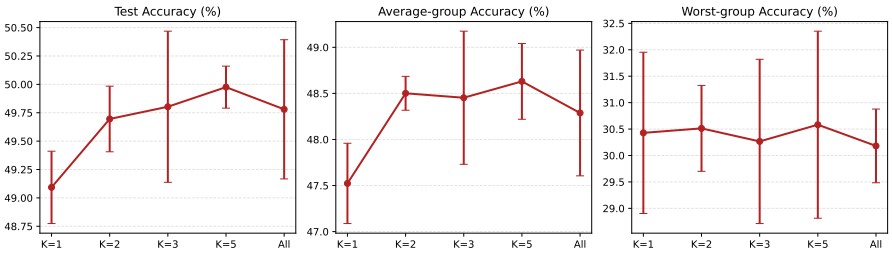

Figure 11: FMoW-WILDS KNN source-selection ablation over five random seeds. The main setting uses $K = 5$, while "All" removes the KNN truncation and reweights all retained source domains. The sweep includes $K \in \{1, 2, 3, 5\}$ and shows that local source selection is stable across nearby neighborhood sizes. The sweep uses up to 8192 covariates per source or target summary, compared with 1024 in the main table.

the neighborhood restriction remains competitive but slightly dilutes the nearest source classifiers, especially on the region-aware metrics.

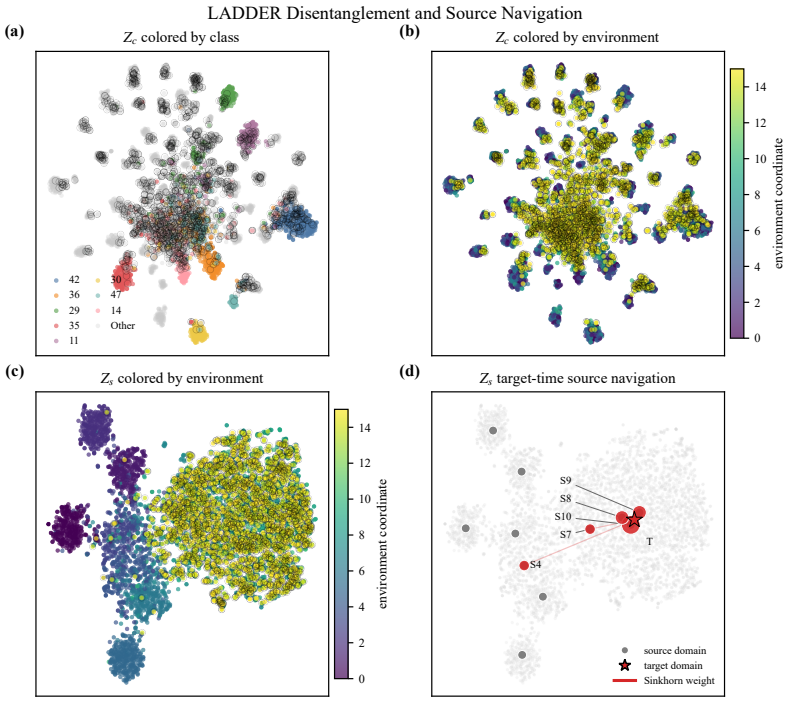

Figure 12: Full FMoW-WILDS t-SNE visualization (van der Maaten & Hinton, 2008). Panels (a,b) show the causal representation $\boldsymbol{Z}_c$ colored by class and domain, respectively; panels (c,d) show the style representation $\boldsymbol{Z}_s^{(e)}$ colored by domain and Sinkhorn source reweighting.

Figure 12 visualizes the domain structure in the style representation used for source reweighting. The style representation $\boldsymbol{Z}_s^{(e)}$ preserves domain geometry, while the causal representation $\boldsymbol{Z}_c$ remains the input used by the source-specific classifiers. The saved t-SNE coordinates are held fixed while navigation weights are recomputed using the FMoW inference settings: balanced debiased Sinkhorn distance with $p = 1$, blur 0.05, scaling 0.8, $K = 5$, and $\tau = 0.5$. This visualization uses the fixed seed-0 feature cache, containing 3000 source and 3000 target examples; the displayed embedding contains 5000 examples. Source summaries contain all 272–276 cached examples per source year. One target summary pools the test split, including both target years, and uses 1024 seed-sampled covariates. Style coordinates are standardized by the pooled source mean and popu-

lation standard deviation, with a 10^{-6} standard-deviation floor. The target star is the mean of the displayed target coordinates. Thus the routing settings match the main FMoW protocol, while the visualization cache is separate from the five-seed evaluation.

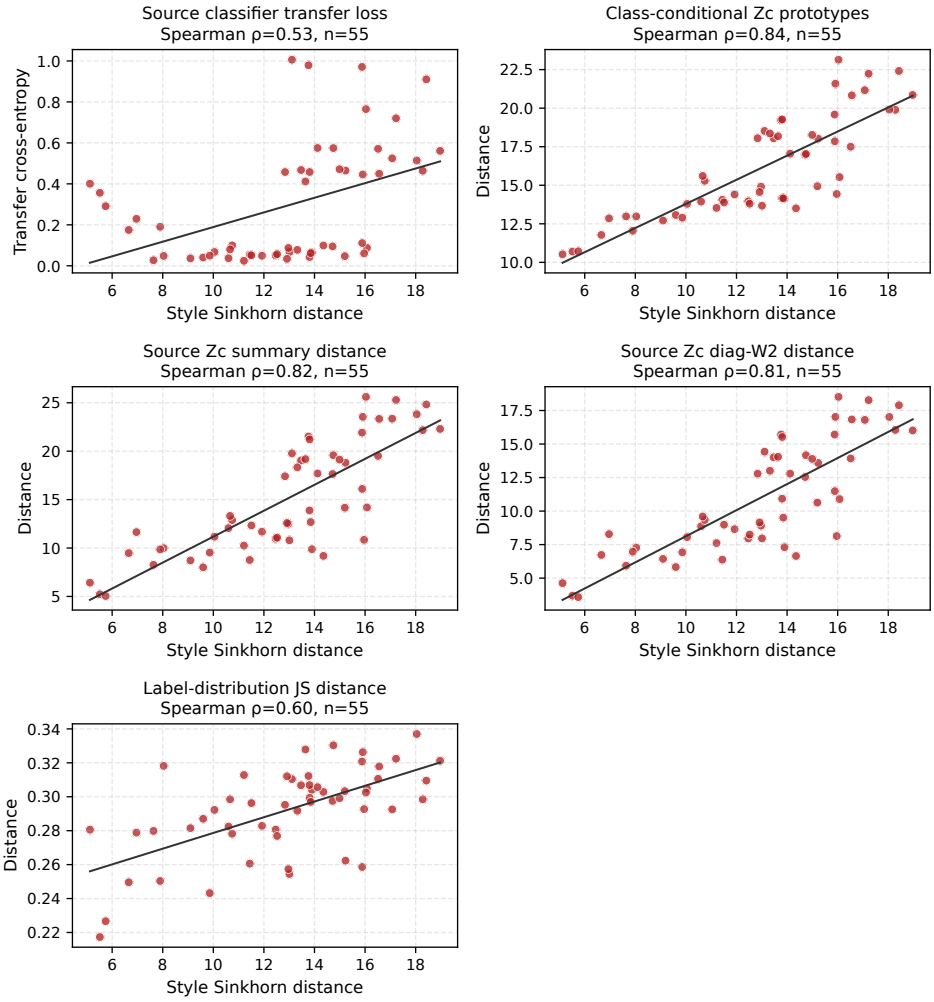

Figure 13: FMoW-WILDS source-geometry diagnostic over all 55 source-domain pairs. The horizontal axis is the debiased Sinkhorn distance between normalized style samples from the diagnostic run. The vertical axes report bidirectional source-classifier transfer loss, class-conditional $\boldsymbol{Z}_c$ prototype distance, two $\boldsymbol{Z}_c$ summary distances, and label-distribution distance.

Figure 13 tests whether the learned style distribution is only a domain identifier or also a quantitative signal for classifier reweighting. The second panel measures the distance between class prototypes in $\boldsymbol{Z}_c$. Each prototype is the mean representation of one class in one source domain; for a domain pair, we average the Euclidean distances between prototypes over classes with at least three cached examples in each domain. The first panel relates source-style distance to the symmetric transfer loss $d^{\mathrm{tr}}_{e,e'} = \frac{1}{2}(L_{e\to e'} + L_{e'\to e})$, where $L_{e\to e'}$ is the cross-entropy of the classifier trained on source e when evaluated on source-domain e' examples. The representative diagnostic uses cached features from the seed-0 diagnostic run, selected independently of the observed correlation value. Its 3000 source examples cover 11 domains with 272–276 examples per domain. We standardize style coordinates using these source examples and compute balanced debiased Sinkhorn distances with $p = 1$, blur 0.05, and scaling 0.8. The cached features and transfer losses come from the same diagnostic run; this cache is separate from the main five-seed evaluation and uses fewer samples per domain. Across the 55 unordered source-domain pairs, the Sinkhorn style distance correlates with bidirectional source-classifier transfer loss (Spearman $\rho = 0.53$) and class-conditional $\boldsymbol{Z}_c$ prototype

distance ($\rho = 0.84$). These pairwise correlations are descriptive because domain pairs share source domains and should not be read as independent-pair significance tests. The prototype diagnostic associates style proximity with similarity of class means in $\boldsymbol{Z}_c$, while the transfer-loss diagnostic provides complementary evidence about source-classifier transferability. Both diagnostics use source data with fixed encoders, classifiers, and normalization statistics.

A source-only probe diagnostic gives a complementary view. Probe features are seed-controlled subsamples with at most 4096 examples per source year, standardized from the probe training split, and split into an $80/20$ source-only train/held-out split. Each probe is a linear Softmax head trained for 50 epochs with AdamW (learning rate 10^{-2}, weight decay 10^{-4}, batch size 1024). On held-out FMoW source examples, a linear probe predicts the year domain more accurately from $\boldsymbol{Z}_s^{(e)}$ than from $\boldsymbol{Z}_c$ ($63.6 \pm 0.4\%$ vs. $56.2 \pm 0.7\%$), while label probes remain high for both branches ($\boldsymbol{Z}_c$: $96.5 \pm 0.3\%$; $\boldsymbol{Z}_s^{(e)}$: $95.9 \pm 0.3\%$). We therefore do not interpret $\boldsymbol{Z}_s^{(e)}$ as a label-free nuisance. To control for class composition, we also train class-conditional domain probes on classes represented in at least 3 source years with at least 20 sampled examples per class-year pair (46.8 ± 0.8 eligible classes; majority baseline $18.8 \pm 0.5\%$). Conditioned on class, $\boldsymbol{Z}_s^{(e)}$ still predicts the source year better than $\boldsymbol{Z}_c$ ($56.8 \pm 0.6\%$ vs. $41.3 \pm 1.0\%$), a 15.5 ± 1.4 point gap. These diagnostics support the operational claim that $\boldsymbol{Z}_s^{(e)}$ is useful for domain-level routing, but they do not imply statistical independence from labels.

A.2.3 iWildCam-WILDS

Table 7: iWildCam-WILDS results over five random seeds. Test accuracy, macro F1, and filtered average-group accuracy are reported in %.

Method	Test ↑	Macro F1 ↑	Filt. Avg. ↑
ERM	42.26 ± 5.48	11.23 ± 1.35	33.29 ± 7.61
IRM	34.59 ± 10.11	6.20 ± 1.08	23.42 ± 9.71
EQRM	40.75 ± 6.62	10.45 ± 1.37	32.22 ± 8.07
DSNR	47.36 ± 7.96	12.34 ± 1.12	39.44 ± 2.07
RDM	39.36 ± 7.67	8.97 ± 1.78	31.09 ± 5.37
ARM-CML	54.08 ± 1.03	$\mathbf{14.74 \pm 0.58}$	49.52 ± 0.84
LADDER (uniform)	44.42 ± 1.03	4.53 ± 0.25	34.13 ± 0.78
LADDER ($K = 1$)	57.96 ± 2.73	11.62 ± 0.90	48.41 ± 0.44
LADDER (sample gate)	58.00 ± 3.05	14.40 ± 0.70	48.16 ± 1.23
LADDER (group gate)	59.22 ± 3.10	12.00 ± 0.86	48.29 ± 1.98
LADDER (split summary)	41.19 ± 3.76	2.42 ± 0.55	36.22 ± 1.80
LADDER	$\mathbf{61.71 \pm 1.76}$	11.61 ± 0.65	$\mathbf{50.01 \pm 0.84}$

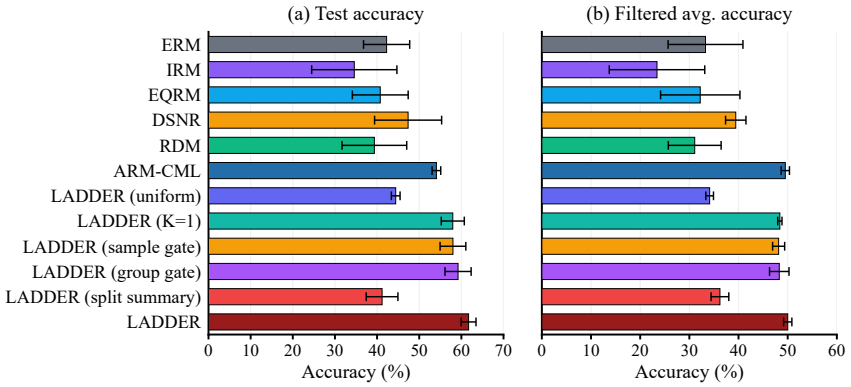

Figure 14: iWildCam-WILDS test accuracy and filtered average-group accuracy.

LADDER forms a separate target summary for each camera location, matching the domain definition in this experiment. Table 7 and Figures 14 and 15 also report LADDER (split summary), which pools all target locations into one summary.

Table 8 evaluates the same fitted ARM-CML model with whole-split, true-group, and shuffled-group contexts, preserving group sizes and summary caps (regions on FMoW; camera locations on iWildCam). Test accuracy changes little across these choices, and true location groups do not improve macro F1 or filtered average-group accuracy over shuffled groups on iWildCam.

Table 8: ARM-CML target-context controls (mean ± sample SD, five seeds; all metrics in %).

Dataset	Context	Test ↑	Worst / Macro F1 ↑	Avg. / Filt. Avg. ↑
FMoW	Split	47.89 ± 0.35	29.53 ± 1.49	46.32 ± 0.80
FMoW	True groups	47.92 ± 0.35	29.70 ± 1.62	46.34 ± 0.81
FMoW	Shuffled groups	47.90 ± 0.34	29.54 ± 1.47	46.32 ± 0.80
iWildCam	Split	54.06 ± 0.86	15.05 ± 0.60	49.98 ± 0.65
iWildCam	True groups	54.08 ± 1.03	14.74 ± 0.58	49.52 ± 0.84
iWildCam	Shuffled groups	54.07 ± 0.86	15.10 ± 0.60	49.99 ± 0.62

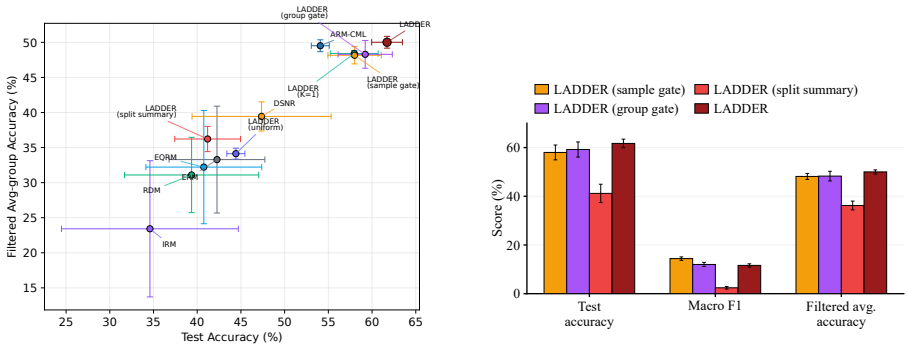

Figure 15: iWildCam-WILDS accuracy/robustness trade-off (left) and routing/target-summary comparison on test accuracy, macro F1, and filtered average-group accuracy (right), including LADDER (sample gate), LADDER (group gate), LADDER (split summary), and LADDER.

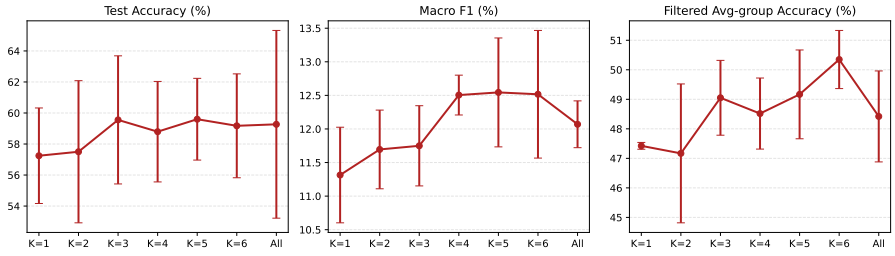

Figure 16: iWildCam-WILDS KNN source-selection ablation over three random seeds. The main setting fixes $K = 6$, while "All" removes the KNN truncation and reweights all retained source domains. Across $K \in \{1, 2, 3, 4, 5, 6\}$, $K = 5$ gives the highest test accuracy and macro F1, whereas $K = 6$ gives the strongest filtered average-group accuracy. The sweep uses up to 8192 covariates per target location; the main table uses five seeds and a cap of 1024.

The iWildCam-WILDS KNN ablation evaluates sensitivity to the neighborhood size. The main protocol fixes $K = 6$, and the ablation shows that local source selection remains competitive across nearby choices while avoiding the robustness degradation caused by using all source classifiers.

iWildCam worst-location diagnostics. This subsection explains why the main text does not emphasize raw worst-location accuracy on iWildCam-WILDS. The target locations are extremely long-

tailed, so the minimum over locations is often determined by locations with very few examples. We therefore report the five-seed diagnostic table below to make this choice explicit.

Table 9: iWildCam-WILDS worst-location diagnostics over five random seeds. Raw worst-location accuracy takes the minimum over all target locations; filtered worst-location accuracy and zero-location counts use locations with at least 50 samples.

Method	Raw worst ↑	Filtered worst ↑	Zero filt. locs. ↓
ERM	0.00 ± 0.00	0.20 ± 0.44	1.4 ± 1.5
IRM	0.00 ± 0.00	0.00 ± 0.00	4.0 ± 5.0
EQRM	0.00 ± 0.00	0.98 ± 1.39	0.6 ± 0.5
DSNR	0.00 ± 0.00	1.96 ± 1.55	0.4 ± 0.9
RDM	0.00 ± 0.00	0.24 ± 0.54	1.4 ± 0.9
ARM-CML	0.00 ± 0.00	1.76 ± 1.75	0.4 ± 0.5
LADDER (uniform)	0.00 ± 0.00	0.00 ± 0.00	2.8 ± 0.8
LADDER ($K = 1$)	0.00 ± 0.00	0.00 ± 0.00	1.0 ± 0.0
LADDER (sample gate)	0.00 ± 0.00	1.96 ± 1.83	0.2 ± 0.4
LADDER (group gate)	0.00 ± 0.00	0.00 ± 0.00	1.8 ± 1.3
LADDER (split summary)	0.00 ± 0.00	0.00 ± 0.00	5.2 ± 0.4
LADDER	0.00 ± 0.00	0.95 ± 1.70	0.6 ± 0.5

These diagnostics explain why the main paper emphasizes filtered average-group accuracy rather than worst-location accuracy. Even after filtering locations with fewer than 50 examples, the worst-location metric remains close to zero for almost all methods, while the average over filtered locations provides a more stable view of target-location robustness. At the 50-sample threshold, the test split retains 35 locations and 42,477 examples. As a threshold sensitivity check for LADDER, using thresholds $m_g \in \{25, 50, 100, 200\}$ gives filtered average-group accuracies of 51.49%, 50.01%, 47.82%, and 51.57%, respectively. These values describe the sensitivity of LADDER to the filtering threshold; the retained location set changes with the threshold.

A.2.4 Paired Routing Comparisons

Table 10: Paired-seed differences for protocol-matched routing controls. Each entry is the first method minus the second in percentage points; intervals are descriptive bootstrap percentile 95% CIs over the five matched seeds. Positive values favor the first method except for ID-to-test drop, where negative values are better.

Dataset	Comparison	Metric	Mean Δ [95% CI]
FMoW	LADDER – LADDER (sample gate)	Test	$+0.21\ [-0.21, 0.55]$
FMoW	LADDER – LADDER (sample gate)	Worst region	$+0.79\ [-0.07, 1.85]$
FMoW	LADDER – LADDER (sample gate)	Average region	$+0.39\ [-0.29, 0.91]$
FMoW	LADDER – LADDER (sample gate)	ID-to-test drop	$-0.14\ [-0.69, 0.41]$
iWildCam	LADDER – LADDER (group gate)	Test	$+2.50\ [0.50, 4.49]$
iWildCam	LADDER – LADDER (group gate)	Macro F1	$-0.38\ [-1.69, 0.54]$
iWildCam	LADDER – LADDER (group gate)	Filtered average	$+1.72\ [-0.16, 4.03]$

A.3 Implementation Details and Hyperparameters

The theoretical analysis uses $\mathcal{W}_1$ on style distributions as a clean population metric. In the main real-data experiments, we approximate the empirical W_1 geometry using the balanced debiased Sinkhorn divergence with Euclidean ground cost $c(\boldsymbol{z}, \boldsymbol{z}') = \|\boldsymbol{z} - \boldsymbol{z}'\|_2$. For each source domain and each target summary unit, we encode at most $m = 1024$ covariates with the fixed style encoder using seed-controlled uniform random subsampling. Style samples are normalized by source-training statistics,

$$\widetilde{\boldsymbol{z}}_s = (\boldsymbol{z}_s - \widehat{\mu}_{\mathrm{src}}) / \max\{\widehat{\sigma}_{\mathrm{src}}, 10^{-6}\},$$

where $\widehat{\mu}_{\mathrm{src}}$ and $\widehat{\sigma}_{\mathrm{src}}$ are computed only from cached source-training style samples and then applied unchanged to target covariates; the maximum and division are coordinate-wise. The implementation uses GeomLoss

```
SamplesLoss(loss="sinkhorn", p=1, blur=0.05,
scaling=0.8, debias=True, backend="tensorized")
```

We use balanced OT because each fingerprint is a probability measure, debiasing to remove entropic self-bias, and $p = 1$ so the ground cost matches the W_1 geometry. The tensorized backend is used for the main runs with $m \leq 1024$. The displayed theory is for exact empirical W_1 and does not include entropic regularization or numerical solver error. The simulation uses a separable empirical quantile proxy. For each style coordinate, it linearly interpolates empirical quantiles at 256 equally spaced probabilities on $[0, 1]$, averages the squared source–target differences over these quantiles, and sums over coordinates. Before source weighting, these distances are divided by the largest source–target distance plus 10^{-8}. The reported runs have 1000 source samples per domain, so the quantile grid has 256 points even when the target summary is smaller. In all cases, the distance estimator is fixed before target evaluation and is computed only from unlabeled target covariates.

Table 11: Implementation settings.

Setting	Simulation	FMoW-WILDS	iWildCam-WILDS
Backbone	Linear latent SCM	ImageNet ResNet-50	ImageNet ResNet-50
Input resolution	–	224×224	224×224
Optimizer / scheduler	Adam / none	AdamW / none	AdamW / none
Learning rates	10^{-2}	representation 10^{-4}, source classifiers 10^{-3}	representation 10^{-4}, source classifiers 10^{-3}
Weight decay	10^{-2}	10^{-4}	10^{-4}
Epochs / steps	150 epochs	25 rep. epochs, 8 head epochs, 800 steps/epoch	20 rep. epochs, 6 head epochs, 700 steps/epoch
Batching	1000 samples/domain	batch 32, eval batch 64, 4 domains/step	batch 32, eval batch 64, 4 domains/step
Dimensions	$d_c = 10, d_s = 8$	$d_c = 512, d_s = 128$	$d_c = 512, d_s = 128$
Loss weights	–	$(\lambda_s, \lambda_a, \lambda_o, \lambda_r) = (0.15, 0.03, 5{\times}10^{-4}, 10^{-5})$	same as FMoW
Source classifiers	source GLMs	frozen-encoder linear classifiers, CE	frozen-encoder linear classifiers, class-balanced CE
Source filtering	$E = 9$ fixed	year domains, min size 512	largest 64 locations, min size 512
Main (K, τ)	$(4, 0.1)$	$(5, 0.5)$	$(6, 1.0)$
Target summary	target latent samples	split-level style summary	location-level style summary
Style distance	coord. quantile W_2^2 proxy	debiased Sinkhorn, $p = 1$, $m \leq 1024$	debiased Sinkhorn, $p = 1$, $m \leq 1024$

FMoW images are resized to 224×224. iWildCam uses random resized cropping and horizontal flipping during training, and resizing to 256 followed by a 224×224 center crop for evaluation; both datasets use ImageNet normalization.

Representation training uses batches containing equal numbers of examples from four sampled source domains. Classification losses use a mean reduction, with the weighted label-loss normalization for iWildCam specified below. The orthogonality loss averages squared entries of the minibatch cross-covariance matrix, and each branch's KL term is averaged over examples and latent coordinates before the two branch terms are added. The loss weights in Table 11 use these averaging conventions.

The reported implementation uses λ_a both as the outer adversarial-loss weight and inside the gradient-reversal layer, so the encoder-side reversed gradient is scaled by $\lambda_a^2 = 9 \times 10^{-4}$ for $\lambda_a = 0.03$; this fixed protocol is used across all reported seeds and datasets. For iWildCam, class-balanced CE uses effective-number class weight $a_y = (1 - \beta_{\mathrm{cb}})/(1 - \beta_{\mathrm{cb}}^{n_y})$ with $\beta_{\mathrm{cb}} = 0.999$, multiplying the ordinary CE for class y (Cui et al., 2019). Here n_y is the class count pooled over the retained source training set; the source classifiers reuse these weights. The weighted CE sums the weighted losses and divides by the sum of class weights in the minibatch.

We manually selected K and τ using validation worst-region accuracy on FMoW and validation macro F1 on iWildCam, following the respective WILDS metrics, and kept them fixed for test evaluation. The reported main settings are $(K, \tau) = (4, 0.1)$ in simulation, $(5, 0.5)$ on FMoW, and $(6, 1.0)$ on iWildCam. Simulation uses $K = 4$ for the main comparison and reports $\tau \in \{0.05, 0.1, 0.5, 1, 5\}$ as a controlled sensitivity analysis over target-summary reliability. The earlier real-data Sinkhorn KNN sweeps use a summary sampling cap of 8192 covariates per domain, compared with 1024 in the main runs. Within each sweep, K varies while the other settings are fixed: FMoW uses $K \in \{1, 2, 3, 5, \mathrm{All}\}$ with $\tau = 0.5$, and iWildCam uses $K \in \{1, 2, 3, 4, 5, 6, \mathrm{All}\}$ with $\tau = 1.0$. For the main real-data runs, the Sinkhorn blur is fixed to 0.05 after source-only style normalization. Training follows the fixed source-data schedules in Table 11, and the resulting final models are evaluated on each split.

The real-data baseline settings use IRM penalty weight 0.1 with 1000 annealing updates, EQRM quantile 0.75 with 200 burn-in updates, and RDM weight 1 with variance weight 0.1 and 200 annealing updates. RDM uses RBF factors $\{0.001, 0.01, 0.1, 1, 10\}$. The DSNR index has dimension 8; its MLPs use hidden dimension 512, ReLU activations, and dropout 0.1.

ARM-CML uses a three-layer 5×5 convolutional context network with hidden width 64 and 12 output channels, concatenated with RGB inputs before the ResNet-50 predictor. The added input-channel kernels copy the pretrained RGB kernels cyclically. Each source episode contains 8 examples from each of 4 sampled domains (batch 32). AdamW, learning rate 10^{-4}, weight decay 10^{-4}, the dataset-specific losses, and the $25 \times 800/20 \times 700$ training schedules match the respective real-data settings above. Target contexts use at most 1024 unlabeled examples per split or location. In simulation, the context and prediction networks are two-hidden-layer MLPs of width 128, with a 16-dimensional context and source-only input standardization; AdamW uses learning rate 10^{-3} and weight decay 10^{-4}. Training uses 150 epochs of 45 episodes, each containing 50 examples from each of 4 source domains. All ARM-CML results use the final source-trained model. The implementation is adapted from the authors' MIT-licensed code (release_2021, commit `e4a0ce2`).

A.4 Runtime and Storage

Table 12: Wall-clock runtime in minutes per seed from available runtime logs. Ordinary baseline rows use the retained original runtime logs; source-classifier routing rows use the Sinkhorn reruns over five seeds.

Method	FMoW-WILDS	iWildCam-WILDS
ERM	59.0 ± 1.4	54.7 ± 1.2
IRM	60.7 ± 1.2	57.0 ± 1.1
EQRM	58.3 ± 1.0	54.1 ± 0.7
DSNR	60.6 ± 1.3	56.9 ± 1.1
RDM	60.6 ± 1.1	57.8 ± 3.8
ARM-CML (training only)	99.4 ± 0.9	79.5 ± 2.6
LADDER (uniform)	85.2 ± 1.4	102.4 ± 5.4
LADDER ($K = 1$)	85.2 ± 1.0	104.2 ± 3.3
LADDER (sample gate)	85.0 ± 1.0	101.3 ± 3.1
LADDER (group gate)	–	101.2 ± 3.1
LADDER	85.2 ± 1.0	104.1 ± 3.2
LADDER (split summary)	–	115.5 ± 5.1

The reported LADDER runtimes include representation training, source-classifier fitting, style-summary construction, and target evaluation. ARM-CML reports the sum of its recorded epoch

training times; its original evaluation times were not saved, so this row is not an end-to-end runtime comparison.

Table 13: Approximate routing storage beyond the shared representation model: source-classifier weight matrices and cached style samples, using float32, $d_c = 512$, $d_s = 128$, and $m = 1024$ samples per source domain. Classifier biases are omitted.

Dataset	Classifier weights	Style fingerprints	Total routing storage
FMoW-WILDS ($E = 11, C = 62$)	1.33 MiB	5.50 MiB	6.83 MiB
iWildCam-WILDS ($E = 64, C = 182$)	22.75 MiB	32.00 MiB	54.75 MiB

For linear source classifiers and m cached style samples per source domain, these weight matrices and fingerprints occupy $4ECd_c + 4Emd_s$ bytes. The target-domain style summary is a temporary prediction buffer and is not part of persistent model storage. Source covariates are not stored after the cached style samples are constructed.

B Theoretical Assumptions and Auxiliary Statements

This section records the formal assumptions and auxiliary statements used in the theoretical analysis. The first subsection expands the compact assumptions from the main paper; the next subsection collects the proposition and lemmas used in the proofs.

B.1 Formal Assumptions and Conditions

For the theoretical analysis, conditional on the domain distributions, observations within each domain are sampled independently and identically from that domain, and the samples from distinct source domains and the target summary sample are mutually independent.

The main paper states compact versions of the assumptions to keep the theoretical presentation readable. We restate the full formal assumptions and conditions used by the proofs below. Throughout, $n_{\min} := \min_{1 \le e \le E} n_e$. Positive constants $C_\star, c_\star$ and their indexed versions may vary between bounds and depend on fixed regularity parameters (and on fixed E in the generic representation bound); C always denotes the number of classes.

Assumption 6 (Oracle latent structure)**.** *For each domain e, the centered observation $\boldsymbol{X}^{(e)} \in \mathbb{R}^p$ follows*

$$\boldsymbol{X}^{(e)} = \boldsymbol{A}_c \boldsymbol{Z}_c^{(e)} + \boldsymbol{A}_s \boldsymbol{Z}_s^{(e)} + \boldsymbol{\epsilon}^{(e)},$$

where $\boldsymbol{A} = [\boldsymbol{A}_c, \boldsymbol{A}_s]$ has full column rank. The causal and style components satisfy

$$\mathrm{Cov}(\boldsymbol{Z}_c^{(e)}, \boldsymbol{Z}_s^{(e)}) = \boldsymbol{0}, \quad \mathrm{Cov}(\boldsymbol{Z}_c^{(e)}) = \boldsymbol{\Sigma}_c, \quad \mathrm{Cov}(\boldsymbol{Z}_s^{(e)}) = \boldsymbol{\Lambda}_e.$$

Here $\boldsymbol{\Sigma}_c$ is domain-invariant and $\boldsymbol{\Lambda}_e = \mathrm{diag}(\lambda_{e1}, \ldots, \lambda_{ed_s})$ is diagonal and domain-dependent. The noise satisfies $\mathbb{E}\boldsymbol{\epsilon}^{(e)} = \boldsymbol{0}$, is independent of the latents and labels, and has domain-invariant covariance $\mathrm{Cov}(\boldsymbol{\epsilon}^{(e)}) = \boldsymbol{\Sigma}_\epsilon$. Assume $\boldsymbol{\Lambda}_1 \succ 0$. For each style coordinate m, define its normalized cross-domain variance profile

$$\boldsymbol{\rho}_m = (\lambda_{2m}/\lambda_{1m}, \ldots, \lambda_{Em}/\lambda_{1m}).$$

The profiles are jointly separating: $\boldsymbol{\rho}_m \neq \boldsymbol{\rho}_{m'}$ for all $m \neq m'$.

Remark 2 (Repeated variance profiles)**.** *If several style coordinates share the same normalized variance profile, they are not separately identifiable from second-order domain variation. The corresponding joint eigenspace is identifiable, but coordinates within that eigenspace are recoverable only up to an orthogonal transformation after whitening. Pairwise profile separation is the special case in which every block is one-dimensional.*

Condition 1 (Sufficient condition for leakage removal)**.** *For leakage removal, the candidate extractor $\boldsymbol{W}_s$ additionally satisfies label-invariant conditional means for the extracted style representation, while the true style representation has label-invariant conditional means within each source*

domain: $\mathbb{E}[\boldsymbol{Z}_s^{(e)} \mid Y^{(e)} = y] = \boldsymbol{m}_s^{(e)}$ *and* $\mathbb{E}[\widehat{\boldsymbol{Z}}_s^{(e)} \mid Y^{(e)} = y] = \widehat{\boldsymbol{m}}_s^{(e)}$ *for all* $y \in \mathrm{supp}(Y^{(e)})$. *Let* $\boldsymbol{\mu}_{e,y} = \mathbb{E}[\boldsymbol{Z}_c^{(e)} \mid Y^{(e)} = y]$ *and choose one reference label* $y_e^0 \in \mathrm{supp}(Y^{(e)})$ *per source domain. The causal conditional means span the causal space:* $\mathrm{span}\{\boldsymbol{\mu}_{e,y} - \boldsymbol{\mu}_{e,y_e^0} : e = 1, \dots, E,\ y \in \mathrm{supp}(Y^{(e)})\} = \mathbb{R}^{d_c}$.

Assumption 7 (Smooth domain geometry)**.** *There is a map* Ψ^* *with* $\boldsymbol{\beta}_*^{(e)} = \Psi^*(\nu_e)$ *and* $\|\Psi^*(\nu) - \Psi^*(\nu')\|_F \le L_\beta \mathcal{W}_1(\nu, \nu')$. *All style distributions are supported on a compact* $\mathcal{Z}_s \subset \mathbb{R}^{d_s}$ *of diameter at most* D_s. *Moreover,* $\Pi_{\mathcal{M}}$ *is* $d_{\mathcal{M}}$*-dimensional Ahlfors-regular on* $\mathcal{M}$, *with* $d_{\mathcal{M}} \in \mathbb{N}$. *The source style laws* $\nu_1, \dots, \nu_E$ *are independently sampled from* $\Pi_{\mathcal{M}}$; ν_0 *is either an independent draw or fixed independently of the source sample, with the Ahlfors-regular bounds holding uniformly around* ν_0. *For the fixed target law* ν_0, *the distance distribution function*

$$F_{\nu_0}(r) = \Pi_{\mathcal{M}}\big(B_{\mathcal{W}_1}(\nu_0, r)\big)$$

is continuous in r *on* $[0, \mathrm{diam}(\mathcal{M})]$.

Assumption 8 (GLM regularity)**.** *Let*

$$\mathcal{S}_\beta = \{\boldsymbol{\beta} \in \mathbb{R}^{C \times d_c} : \mathbf{1}^\top \boldsymbol{\beta} = \mathbf{0}\},$$

and let $\mathcal{B} \subset \mathcal{S}_\beta$ *be a compact convex parameter set with* $\|\boldsymbol{\beta}\|_F \le R_\beta$. *For all samples,* $\|\boldsymbol{Z}_c\|_2 \le R_c$. *For* $\ell_{\mathrm{CE}}(\boldsymbol{u}, y) = -\log(\mathrm{Softmax}(\boldsymbol{u})_y)$, *define*

$$\mathcal{R}_e(\boldsymbol{\beta}) = \mathbb{E}[\ell_{\mathrm{CE}}(\boldsymbol{\beta}\boldsymbol{Z}_c, Y) \mid e], \qquad \widehat{\mathcal{R}}_e(\boldsymbol{\beta}) = \frac{1}{n_e}\sum_{i=1}^{n_e} \ell_{\mathrm{CE}}(\boldsymbol{\beta}\boldsymbol{z}_{c,i}^{(e)}, y_i^{(e)}).$$

For every domain considered in the oracle analysis, $\boldsymbol{\beta}_*^{(e)} \in \mathrm{relint}(\mathcal{B})$, $\nabla_{\mathcal{S}_\beta}\mathcal{R}_e(\boldsymbol{\beta}_*^{(e)}) = \mathbf{0}$, *and* $\mathcal{R}_e$ *is uniformly* μ*-strongly convex on* $\mathcal{B}$ *with respect to the Frobenius norm restricted to* $\mathcal{S}_\beta$. *For every source domain and sample, the single-sample Hessian is uniformly bounded and* L_H*-Lipschitz in* $\boldsymbol{\beta}$ *over* $\mathcal{B}$ *as an operator on* $\mathcal{S}_\beta$. *With* $\boldsymbol{\Pi}_C = \boldsymbol{I}_C - C^{-1}\mathbf{1}\mathbf{1}^\top$, *let*

$$\mathcal{S}_C = \{\boldsymbol{u} \in \mathbb{R}^C : \mathbf{1}^\top \boldsymbol{u} = 0\}, \qquad \mathcal{U}_R = \{\boldsymbol{u} \in \mathcal{S}_C : \|\boldsymbol{u}\|_2 \le R_{\mathrm{logit}}\}.$$

The oracle and comparison centered logits used in the lower-bound results lie in $\mathcal{U}_R$ *almost surely. On* $\mathcal{U}_R$, *the conditional cross-entropy risk is* μ_ℓ*-strongly convex; on* $\mathcal{S}_C$, *it is globally* L_ℓ*-smooth. Bounded centered logits imply that the Softmax probabilities are uniformly bounded away from zero, yielding a uniform positive-curvature lower bound on* $\mathcal{U}_R$ *along the centered-logit subspace* $\mathcal{S}_C$.

Comparison notation. We compare LADDER, PooledJoint, and the fixed prediction rule using a scalar coordinate c for latent domain position and define

$$f^*(\boldsymbol{Z}_c, c) = (\boldsymbol{\beta}_{\mathrm{base}} + c\boldsymbol{\beta}_{\mathrm{shift}})\boldsymbol{Z}_c.$$

For any predictor f, write

$$\mathcal{R}_c(f) = \mathbb{E}[\ell_{\mathrm{CE}}(f(\boldsymbol{Z}_c, \boldsymbol{Z}_s), Y_c)], \qquad Y_c \sim \mathrm{Softmax}(f^*(\boldsymbol{Z}_c, c)),$$

and define the bounded additive pooled class

$$\mathcal{H}_{\mathrm{pool}}(R_W) = \{f = \boldsymbol{W}_c\boldsymbol{Z}_c + \boldsymbol{W}_s\boldsymbol{Z}_s : \|\boldsymbol{W}_c\|_F, \|\boldsymbol{W}_s\|_F \le R_W\}.$$

For target coordinate c_0, define

$$f_{\mathrm{L}} = f^*(\cdot, \widehat{c}_{\mathrm{L}}), \qquad \widehat{c}_{\mathrm{L}} = \sum_e w_e c_e, \qquad f_{\mathrm{inv}} = f^*(\cdot, c_{\mathrm{inv}}),$$

where c_e is source e's coordinate and c_{inv} does not depend on the target domain.

Assumption 9 (Comparison shift model)**.** *The coordinate satisfies* $c \sim \pi_c$, *is bounded, and has* $\mathrm{Var}_{\pi_c}(c) > 0$. *Moreover,* $\mathbb{E}\boldsymbol{Z}_c = \mathbf{0}$, $\mathbb{E}\boldsymbol{Z}_c\boldsymbol{Z}_c^\top = \boldsymbol{\Sigma}_c \succ 0$, *and* $\boldsymbol{Z}_c \perp (c, \boldsymbol{Z}_s)$. *For the target coordinate,* $c_{\mathrm{inv}} \ne c_0$. *The radius* R_W *is large enough to contain one optimizer of the unrestricted additive projection used in Lemma 8. For the comparison lower bounds,* $\boldsymbol{\Pi}_C f^*(\boldsymbol{Z}_c, c)$ *and the centered logits of the fixed predictor lie in* $\mathcal{U}_R$ *almost surely; moreover,* $\boldsymbol{\Pi}_C f(\boldsymbol{Z}_c, \boldsymbol{Z}_s) \in \mathcal{U}_R$ *almost surely for every* $f \in \mathcal{H}_{\mathrm{pool}}(R_W)$.

Assumption 10 (Second-moment representation error). *For $e = 0, \dots, E$ and $i = 1, \dots, n_e$, let $\boldsymbol{z}_{c,i}^{(e)} \in \mathbb{R}^{d_c}$ and $\boldsymbol{z}_{s,i}^{(e)} \in \mathbb{R}^{d_s}$ be the oracle causal and style features of observation i in domain e. After a fixed coordinate alignment, the learned features satisfy*

$$\widetilde{\boldsymbol{z}}_{c,i}^{(e)} = \boldsymbol{z}_{c,i}^{(e)} + \boldsymbol{\xi}_{c,i}^{(e)}, \qquad \widetilde{\boldsymbol{z}}_{s,i}^{(e)} = \boldsymbol{z}_{s,i}^{(e)} + \boldsymbol{\xi}_{s,i}^{(e)}.$$

The representation errors have a common scale $\epsilon_{\mathrm{rep}} \geq 0$:

$$\mathbb{E}\|\boldsymbol{\xi}_{c,i}^{(e)}\|_2^2 \vee \mathbb{E}\|\boldsymbol{\xi}_{s,i}^{(e)}\|_2^2 \leq \epsilon_{\mathrm{rep}}^2 \quad \textit{for all } e, i.$$

For some $\mu_{\mathrm{rep}} > 0$, each source empirical risk

$$\widetilde{\mathcal{R}}_e^{\mathrm{emp}}(\boldsymbol{\beta}) = \frac{1}{n_e} \sum_{i=1}^{n_e} \ell_{\mathrm{CE}}(\boldsymbol{\beta} \widetilde{\boldsymbol{z}}_{c,i}^{(e)}, y_i^{(e)}), \qquad e = 1, \dots, E,$$

is μ_{rep}-strongly convex on the same compact convex set $\mathcal{B}$. To define target population risk, condition on the fitted encoders, source classifiers, and the sample used to construct the target summary. Draw a fresh target observation independently of the fitting and summary samples. Let $(\boldsymbol{Z}_c^{(0)}, \boldsymbol{Z}_s^{(0)}, Y^{(0)})$ be its oracle quantities and $(\boldsymbol{\xi}_c^{(0)}, \boldsymbol{\xi}_s^{(0)})$ its representation errors. Under this conditioning,

$$\Pr\left(Y^{(0)} = y \mid \boldsymbol{Z}_c^{(0)}, \boldsymbol{Z}_s^{(0)}, \boldsymbol{\xi}_c^{(0)}, \boldsymbol{\xi}_s^{(0)}\right) = \left[\mathrm{Softmax}(\boldsymbol{\beta}_*^{(0)} \boldsymbol{Z}_c^{(0)})\right]_y.$$

The fresh-observation errors satisfy

$$\mathbb{E}\|\boldsymbol{\xi}_c^{(0)}\|_2^2 \vee \mathbb{E}\|\boldsymbol{\xi}_s^{(0)}\|_2^2 \leq \epsilon_{\mathrm{rep}}^2,$$

where expectation is over the fitted objects and the fresh observation.

B.2 Auxiliary Propositions and Lemmas

The proposition and lemmas below are the technical tools invoked by the proof section.

Proposition 1 (Unique Structure of the Global Linear Transform). *Let $\boldsymbol{W}_s \in \mathbb{R}^{d_s \times p}$ be a candidate population style-extraction matrix, and suppose the induced style transform $\boldsymbol{M} = \boldsymbol{W}_s \boldsymbol{A}_s \in \mathbb{R}^{d_s \times d_s}$ is invertible. Under Assumption 6, if $\boldsymbol{M} \boldsymbol{\Lambda}_e \boldsymbol{M}^\top$ is diagonal for every source domain $e = 1, \dots, E$, then $\boldsymbol{M} = \boldsymbol{P}\boldsymbol{D}$, where $\boldsymbol{P}$ is a permutation matrix and $\boldsymbol{D}$ is a non-singular diagonal scaling matrix.*

Lemma 1 (Uniform Concentration of Empirical Hessian). *Let $d_\beta = (C-1)d_c$. Under Assumption 8, the empirical Hessian converges uniformly to the population Hessian over $\mathcal{B}$ as an operator on $\mathcal{S}_\beta$. In particular, there exists a constant $C_\star > 0$, depending only on the uniform single-sample Hessian bound and hence on R_c, such that for any $\delta > 0$, if*

$$n_e \geq \frac{C_\star}{\mu^2}\left[d_\beta \log\left(1 + \frac{R_\beta L_H}{\mu}\right) + \log\frac{d_\beta}{\delta}\right],$$

then

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}} \|\nabla^2 \widehat{\mathcal{R}}_e(\boldsymbol{\beta}) - \nabla^2 \mathcal{R}_e(\boldsymbol{\beta})\|_2 \leq \mu/2$$

with probability at least $1 - \delta$.

Lemma 2 (Finite-Sample Source Estimation). *Under Assumption 8, for any source domain e and sufficiently large n_e,*

$$\mathbb{E}\|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F \leq \mathcal{O}\left(\frac{R_c}{\mu\sqrt{n_e}}\right).$$

Lemma 3 (Uniform Source-Parameter Bound). *Let $d_\beta = (C-1)d_c$. Under Assumption 8, for every $\delta \in (0,1)$, if*

$$n_{\min} \geq \frac{C_\star}{\mu^2}\left[d_\beta \log\left(1 + \frac{R_\beta L_H}{\mu}\right) + \log\frac{d_\beta E}{\delta}\right],$$

then, with probability at least $1 - \delta$,

$$\max_{1 \leq e \leq E} \|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F \leq C_\star \frac{R_c}{\mu}\sqrt{\frac{d_\beta + \log(E/\delta)}{n_{\min}}}.$$

Consequently, under the sample-size requirement of Theorem 2 with a sufficiently large constant,

$$\mathbb{E} \max_{1\le e\le E} \|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F \le \mathcal{O}\left(\frac{R_c}{\mu}\sqrt{\frac{d_\beta + \log E}{n_{\min}}}\right),$$

and the corresponding squared maximum is bounded by $\mathcal{O}(R_c^2(d_\beta + \log E)/(\mu^2 n_{\min}))$.

For Wasserstein rates, write

$$\gamma(n,d) = n^{-1/(d\vee 2)}\ell_d(n), \qquad \Gamma(n,E,d) = ((1+\log E)/n)^{1/(d\vee 2)}\widetilde{\ell}_d(n),$$

where $\ell_2(n) = \widetilde{\ell}_2(n) = \log(1+n)$ and $\ell_d(n) = \widetilde{\ell}_d(n) = 1$ for $d \neq 2$.

Lemma 4 (Empirical Wasserstein Concentration)**.** *Under Assumption 7, for any domain style distribution ν supported on $\mathcal{Z}_s$ and empirical measure $\widehat{\nu}$ from n samples,*

$$\mathbb{E}\mathcal{W}_1(\nu, \widehat{\nu}) \le C_\star \gamma(n, d_s).$$

Moreover, with $\Delta_W = \gamma(n_{\min}, d_s) + \Gamma(n_{\min}, E, d_s)$,

$$\mathbb{E} \max_{1\le e\le E} \mathcal{W}_1(\nu_e, \widehat{\nu}_e) \le \mathcal{O}(\Delta_W), \qquad \mathbb{E} \max_{1\le e\le E} \mathcal{W}_1^2(\nu_e, \widehat{\nu}_e) \le \mathcal{O}(\Delta_W^2).$$

Lemma 5 (Weighted Wasserstein Approximation)**.** *With these empirical Wasserstein rates, under Assumption 7 and Lemma 4,*

$$\mathbb{E}\left[L_\beta \sum_{e\in\widehat{\mathcal{N}}_K(0)} w_e\{\mathcal{W}_1(\nu_e, \widehat{\nu}_e) + \mathcal{W}_1(\widehat{\nu}_0, \nu_0)\}\right] \le \mathcal{O}\{\gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s)\}.$$

Lemma 6 (Manifold Interpolation via Empirical KNN)**.** *Under Assumption 7,*

$$\mathbb{E}\left[L_\beta \sum_{e\in\widehat{\mathcal{N}}_K(0)} w_e \mathcal{W}_1(\widehat{\nu}_e, \widehat{\nu}_0)\right] \le \mathcal{O}\Big((K/E)^{1/d_{\mathcal{M}}}\Big) + \mathcal{O}\{\gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s)\}.$$

Lemma 7 (Centered Logit Risk Bounds)**.** *Under Assumption 8, write* $\boldsymbol{Z} = (\boldsymbol{Z}_c, \boldsymbol{Z}_s)$ *and define*

$$\mathcal{D}_\ell(f, f^*) = \mathbb{E}\,\|\boldsymbol{\Pi}_C\{f(\boldsymbol{Z}) - f^*(\boldsymbol{Z}_c, c)\}\|_2^2.$$

Let $\Delta_\ell(f, f^*)$ *denote the corresponding expected cross-entropy excess risk. If the centered logits of* f *and* f^* *lie in* $\mathcal{U}_R$ *almost surely, then*

$$\Delta_\ell(f, f^*) \ge \frac{\mu_\ell}{2}\mathcal{D}_\ell(f, f^*).$$

For arbitrary finite centered logits,

$$\Delta_\ell(f, f^*) \le \frac{L_\ell}{2}\mathcal{D}_\ell(f, f^*).$$

Lemma 8 (Best Additive Approximation to a Multiplicative Shift)**.** *Under Assumptions 9 and 8, the optimal expected centered-logit squared error of the additive pooled class relative to* $f^*(\boldsymbol{Z}_c, c)$ *satisfies*

$$\inf_{f\in\mathcal{H}_{\text{pool}}(R_W)} \mathbb{E}\|\boldsymbol{\Pi}_C[f(\boldsymbol{Z}_c, \boldsymbol{Z}_s) - f^*(\boldsymbol{Z}_c, c)]\|_2^2 = \mathrm{Var}(c)\,\|\boldsymbol{\Pi}_C\boldsymbol{\beta}_{\text{shift}}\boldsymbol{\Sigma}_c^{1/2}\|_F^2.$$

C Proofs of Theoretical Statements

This section proves the statements from the main paper and the auxiliary statements stated above. We first establish representation identifiability, then source-parameter and Wasserstein navigation bounds, target parameter and risk bounds, baseline comparisons, and finally the generic-representation perturbation bound. The arguments use standard tools from GLMs, M-estimation, concentration, optimal transport, and nearest-neighbor theory (McCullagh & Nelder, 1989; van der Vaart, 1998; Wainwright, 2019; Tropp, 2012; Boucheron et al., 2013; Vershynin, 2018; Fournier & Guillin, 2015; Devroye et al., 1996; Biau & Devroye, 2015).

C.1 Proof of Proposition 1

Proof. Let $\boldsymbol{D}_e = \boldsymbol{M}\boldsymbol{\Lambda}_e\boldsymbol{M}^\top$. By assumption, each $\boldsymbol{D}_e$ is diagonal. Since $\boldsymbol{\Lambda}_1 \succ 0$ and $\boldsymbol{M}$ is invertible, $\boldsymbol{D}_1 \succ 0$. Define

$$\boldsymbol{Q} = \boldsymbol{D}_1^{-1/2}\boldsymbol{M}\boldsymbol{\Lambda}_1^{1/2}.$$

The whitening identity gives

$$\begin{aligned}\boldsymbol{Q}\boldsymbol{Q}^\top &= \boldsymbol{D}_1^{-1/2}\boldsymbol{M}\boldsymbol{\Lambda}_1^{1/2}\boldsymbol{\Lambda}_1^{1/2}\boldsymbol{M}^\top\boldsymbol{D}_1^{-1/2}\\ &= \boldsymbol{D}_1^{-1/2}(\boldsymbol{M}\boldsymbol{\Lambda}_1\boldsymbol{M}^\top)\boldsymbol{D}_1^{-1/2}\\ &= \boldsymbol{D}_1^{-1/2}\boldsymbol{D}_1\boldsymbol{D}_1^{-1/2} = \boldsymbol{I}.\end{aligned}$$

Since $\boldsymbol{Q}$ is square, $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^\top$ and $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}$. For $e = 2, \dots, E$, let

$$\boldsymbol{R}_e = \boldsymbol{\Lambda}_1^{-1/2}\boldsymbol{\Lambda}_e\boldsymbol{\Lambda}_1^{-1/2}.$$

Each $\boldsymbol{R}_e$ is diagonal, and

$$\boldsymbol{Q}\boldsymbol{R}_e\boldsymbol{Q}^\top = \boldsymbol{D}_1^{-1/2}\boldsymbol{D}_e\boldsymbol{D}_1^{-1/2}$$

is diagonal. Equivalently, the rows of $\boldsymbol{Q}$ form a common orthonormal eigenbasis of the family $\{\boldsymbol{R}_e\}_{e=2}^{E}$.

Write $\boldsymbol{D}_e = \mathrm{diag}(d_{e1}, \dots, d_{ed_s})$. Right-multiplication of the preceding diagonalization identity by $\boldsymbol{Q}$ gives

$$\boldsymbol{Q}\boldsymbol{R}_e = \boldsymbol{D}_1^{-1/2}\boldsymbol{D}_e\boldsymbol{D}_1^{-1/2}\boldsymbol{Q}, \qquad e = 2, \dots, E.$$

Its (j, m) entry is

$$Q_{jm}\frac{\lambda_{em}}{\lambda_{1m}} = \frac{d_{ej}}{d_{1j}}Q_{jm}, \qquad \left(\frac{\lambda_{em}}{\lambda_{1m}} - \frac{d_{ej}}{d_{1j}}\right)Q_{jm} = 0.$$

The joint profile is $\boldsymbol{\rho}_m = (\lambda_{2m}/\lambda_{1m}, \dots, \lambda_{Em}/\lambda_{1m})$. If $Q_{jm} \neq 0$ and $Q_{jn} \neq 0$, then

$$\frac{\lambda_{em}}{\lambda_{1m}} = \frac{d_{ej}}{d_{1j}} = \frac{\lambda_{en}}{\lambda_{1n}} \quad (e = 2, \dots, E), \qquad \boldsymbol{\rho}_m = \boldsymbol{\rho}_n.$$

The joint profiles are pairwise distinct by Assumption 6, so each row has at most one nonzero entry. Orthogonality gives

$$\sum_m Q_{jm}^2 = 1, \qquad \sum_m Q_{jm}Q_{km} = 0 \quad (j \neq k).$$

Each row therefore has one entry equal to $+1$ or -1, and different rows use different columns. Thus $\boldsymbol{Q} = \boldsymbol{P}\boldsymbol{S}$ for a permutation matrix $\boldsymbol{P}$ and a diagonal sign matrix $\boldsymbol{S}$. Substitution gives

$$\boldsymbol{M} = \boldsymbol{D}_1^{1/2}\boldsymbol{P}\boldsymbol{S}\boldsymbol{\Lambda}_1^{-1/2} = \boldsymbol{P}\widetilde{\boldsymbol{D}},$$

where $\widetilde{\boldsymbol{D}} = (\boldsymbol{P}^\top\boldsymbol{D}_1^{1/2}\boldsymbol{P})\boldsymbol{S}\boldsymbol{\Lambda}_1^{-1/2}$ is nonsingular and diagonal. Renaming $\widetilde{\boldsymbol{D}}$ as $\boldsymbol{D}$ gives the result. □

C.2 Proof of Theorem 1

Proof. From Assumption 6, the covariance of $\boldsymbol{X}^{(e)}$ decomposes as

$$\boldsymbol{\Sigma}_{\boldsymbol{X}}^{(e)} = \boldsymbol{A}_c\boldsymbol{\Sigma}_c\boldsymbol{A}_c^\top + \boldsymbol{A}_s\boldsymbol{\Lambda}_e\boldsymbol{A}_s^\top + \boldsymbol{\Sigma}_\epsilon.$$

The causal covariance and noise covariance are domain-invariant, while the domain-varying style covariance is $\boldsymbol{A}_s\boldsymbol{\Lambda}_e\boldsymbol{A}_s^\top$. By the oracle diagonalization condition in Theorem 1, $\boldsymbol{M}\boldsymbol{\Lambda}_e\boldsymbol{M}^\top$ is diagonal for every source domain, where $\boldsymbol{M} = \boldsymbol{W}_s\boldsymbol{A}_s$. Proposition 1 gives $\boldsymbol{M} = \boldsymbol{P}\boldsymbol{D}$ for a permutation $\boldsymbol{P}$ and nonsingular diagonal $\boldsymbol{D}$.

Expanding the extracted representation gives

$$\widehat{\boldsymbol{Z}}_s^{(e)} = \boldsymbol{W}_s\boldsymbol{X}^{(e)} = \boldsymbol{W}_s\boldsymbol{A}_c\boldsymbol{Z}_c^{(e)} + \boldsymbol{W}_s\boldsymbol{A}_s\boldsymbol{Z}_s^{(e)} + \boldsymbol{W}_s\boldsymbol{\epsilon}^{(e)} = \boldsymbol{B}\boldsymbol{Z}_c^{(e)} + \boldsymbol{P}\boldsymbol{D}\boldsymbol{Z}_s^{(e)} + \widetilde{\boldsymbol{\epsilon}}^{(e)},$$

where $\boldsymbol{B} = \boldsymbol{W}_s \boldsymbol{A}_c$ and $\widetilde{\boldsymbol{\epsilon}}^{(e)} = \boldsymbol{W}_s \boldsymbol{\epsilon}^{(e)}$. This identifies the style mixing component up to permutation and scaling while allowing the extracted representation to contain a causal leakage term $\boldsymbol{B}\boldsymbol{Z}_c^{(e)}$.

For the leakage-removal statement, assume Condition 1. Taking conditional expectations and using the zero-mean independent noise gives, for any $y \in \mathrm{supp}(Y^{(e)})$,

$$\widehat{\boldsymbol{m}}_s^{(e)} = \boldsymbol{B}\boldsymbol{\mu}_{e,y} + \boldsymbol{P}\boldsymbol{D}\boldsymbol{m}_s^{(e)}.$$

Subtracting the conditional-mean identity at the reference label y_e^0 gives

$$\begin{aligned}\mathbf{0} &= \widehat{\boldsymbol{m}}_s^{(e)} - \widehat{\boldsymbol{m}}_s^{(e)} \\ &= (\boldsymbol{B}\boldsymbol{\mu}_{e,y} + \boldsymbol{P}\boldsymbol{D}\boldsymbol{m}_s^{(e)}) - (\boldsymbol{B}\boldsymbol{\mu}_{e,y_e^0} + \boldsymbol{P}\boldsymbol{D}\boldsymbol{m}_s^{(e)}) \\ &= \boldsymbol{B}(\boldsymbol{\mu}_{e,y} - \boldsymbol{\mu}_{e,y_e^0}).\end{aligned}$$

By the spanning condition, every $\boldsymbol{v} \in \mathbb{R}^{d_c}$ has coefficients $a_{e,y}$ such that

$$\boldsymbol{v} = \sum_{e=1}^{E} \sum_{y \in \mathrm{supp}(Y^{(e)})} a_{e,y}(\boldsymbol{\mu}_{e,y} - \boldsymbol{\mu}_{e,y_e^0}).$$

Applying $\boldsymbol{B}$ gives

$$\boldsymbol{B}\boldsymbol{v} = \sum_{e=1}^{E} \sum_{y \in \mathrm{supp}(Y^{(e)})} a_{e,y}\boldsymbol{B}(\boldsymbol{\mu}_{e,y} - \boldsymbol{\mu}_{e,y_e^0}) = \mathbf{0} \quad \text{for every } \boldsymbol{v} \in \mathbb{R}^{d_c}.$$

Hence $\boldsymbol{B} = \mathbf{0}$ and

$$\widehat{\boldsymbol{Z}}_s^{(e)} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{Z}_s^{(e)} + \widetilde{\boldsymbol{\epsilon}}^{(e)},$$

which is recovery up to permutation, coordinate-wise scaling, and projected noise. □

Remark 3 (Effect of permutation and scaling on style geometry)**.** *Let $T = \boldsymbol{P}\boldsymbol{D}$. In the leakage-free and noise-free case, or when considering only the deterministic coordinate transform of the oracle style laws, for any two style distributions ν and ν',*

$$\sigma_{\min}(\boldsymbol{D})\mathcal{W}_1(\nu,\nu') \le \mathcal{W}_1(T_\#\nu, T_\#\nu') \le \|\boldsymbol{D}\|_{\mathrm{op}}\mathcal{W}_1(\nu,\nu').$$

Hence the oracle and transformed-oracle style geometries are bi-Lipschitz equivalent. For transformed laws $\bar{\nu} = T_\#\nu$, the map $\bar{\Psi}^*(\bar{\nu}) = \Psi^*(T_\#^{-1}\bar{\nu})$ remains Lipschitz with constant $L_\beta/\sigma_{\min}(\boldsymbol{D})$. This statement concerns the deterministic transform T alone. In the presence of projected representation noise, exact bi-Lipschitz equivalence need not hold; such residual error is instead handled by the perturbation interface in Assumption 10.

C.3 Proof of Lemma 1

Proof. We aim to bound the maximum spectral norm of the empirical Hessian $\nabla^2\widehat{\mathcal{R}}_e(\boldsymbol{\beta})$ deviating from the population Hessian $\nabla^2\mathcal{R}_e(\boldsymbol{\beta})$ over the entire compact parameter space $\mathcal{B}$. The argument combines a finite covering of $\mathcal{B}$, pointwise matrix concentration on the cover, and Lipschitz extension to the full parameter space (Tropp, 2012; Vershynin, 2018).

Covering argument. The parameter space $\mathcal{B} \subset \mathbb{R}^{C\times d_c}$ is a bounded compact convex set, and for any $\boldsymbol{\beta} \in \mathcal{B}$, $\|\boldsymbol{\beta}\|_F \le R_\beta$. The effective dimension of this centered-gauge space is $d_\beta = (C-1)d_c$. Standard covering number bounds in high-dimensional geometry (Vershynin, 2018) imply that, for any precision $\epsilon > 0$, there exists a finite subset $\mathcal{B}_\epsilon \subset \mathcal{B}$ (an ϵ-net) such that for any $\boldsymbol{\beta} \in \mathcal{B}$, there exists at least one $\boldsymbol{\beta}' \in \mathcal{B}_\epsilon$ satisfying $\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_F \le \epsilon$. The cardinality of this ϵ-net satisfies:

$$|\mathcal{B}_\epsilon| \le \left(1 + \frac{2R_\beta}{\epsilon}\right)^{d_\beta}$$

Pointwise concentration and union bound. Fix an arbitrary point $\boldsymbol{\beta}' \in \mathcal{B}_\epsilon$. The empirical Hessian is the average of n_e independent random matrices. For the multi-class cross-entropy loss, its single-sample Hessian has the Kronecker-product form: $\nabla^2\ell_{\mathrm{CE}} = (\boldsymbol{Z}_c\boldsymbol{Z}_c^\top) \otimes (\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top)$, where $\boldsymbol{p} = \mathrm{Softmax}(\boldsymbol{\beta}\boldsymbol{Z}_c)$ is the predicted probability vector. Here vec stacks the columns of a matrix.

By the spectral norm property of the Kronecker product, $\|\nabla^2 \ell_{\mathrm{CE}}\|_2 = \|\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\|_2 \cdot \|\boldsymbol{Z}_c \boldsymbol{Z}_c^\top\|_2$. The simplex covariance has maximum eigenvalue at most $\frac{1}{2}$ (achieved in binary classification with uniform probabilities), and $\|\boldsymbol{Z}_c\|_2 \le R_c$; hence the single-sample Hessian has spectral norm at most $M = \frac{1}{2}R_c^2$, independent of the number of classes C.

Work in an orthonormal coordinate system of $\mathcal{S}_\beta$, of dimension d_β. Let $\boldsymbol{H}_i$ be the single-sample Hessian restricted to this space at $\boldsymbol{\beta}'$, and set $\boldsymbol{X}_i = \boldsymbol{H}_i - \mathbb{E}\boldsymbol{H}_i$. The positive-semidefinite Hessian bound implies

$$\begin{aligned} \boldsymbol{0} \preceq \boldsymbol{H}_i \preceq M\boldsymbol{I}_{d_\beta}, \quad \boldsymbol{0} \preceq \mathbb{E}\boldsymbol{H}_i \preceq M\boldsymbol{I}_{d_\beta} \implies -M\boldsymbol{I}_{d_\beta} \preceq \boldsymbol{X}_i \preceq M\boldsymbol{I}_{d_\beta}, \\ \mathbb{E}\boldsymbol{X}_i = \boldsymbol{0}, \qquad \boldsymbol{X}_i^2 \preceq M^2\boldsymbol{I}_{d_\beta}, \\ \left\| \sum_{i=1}^{n_e} M^2 \boldsymbol{I}_{d_\beta} \right\|_2 = n_e M^2. \end{aligned}$$

The deterministic bound $M\boldsymbol{I}_{d_\beta}$ commutes with $\boldsymbol{X}_i$. The corresponding matrix Hoeffding bound (Tropp, 2012, Remark 7.4), applied to both $\boldsymbol{X}_i$ and $-\boldsymbol{X}_i$, gives

$$\begin{aligned} \Pr\left(\left\| \frac{1}{n_e} \sum_i \boldsymbol{X}_i \right\|_2 \ge t_H \right) &\le \Pr\left(\lambda_{\max}\left(\sum_i \boldsymbol{X}_i \right) \ge n_e t_H \right) \\ &\quad + \Pr\left(\lambda_{\max}\left(-\sum_i \boldsymbol{X}_i \right) \ge n_e t_H \right) \\ &\le 2d_\beta \exp\left(-\frac{(n_e t_H)^2}{2n_e M^2} \right). \end{aligned}$$

Identifying the centered sum with the empirical Hessian deviation yields

$$\mathbb{P}\left(\left\| \nabla^2 \widehat{\mathcal{R}}_e(\boldsymbol{\beta}') - \nabla^2 \mathcal{R}_e(\boldsymbol{\beta}') \right\|_2 \ge t_H \right) \le 2d_\beta \exp\left(-\frac{n_e t_H^2}{2M^2} \right)$$

Applying the union bound scales this probability across the entire ϵ-net $\mathcal{B}_\epsilon$:

$$\mathbb{P}\left(\sup_{\boldsymbol{\beta}' \in \mathcal{B}_\epsilon} \left\| \nabla^2 \widehat{\mathcal{R}}_e(\boldsymbol{\beta}') - \nabla^2 \mathcal{R}_e(\boldsymbol{\beta}') \right\|_2 \ge t_H \right) \le 2d_\beta \left(1 + \frac{2R_\beta}{\epsilon} \right)^{d_\beta} \exp\left(-\frac{n_e t_H^2}{2M^2} \right)$$

Extension to the full parameter space. For any point $\boldsymbol{\beta} \in \mathcal{B}$ in the full space, since $\mathcal{B}_\epsilon$ is an ϵ-net, there must exist a $\boldsymbol{\beta}' \in \mathcal{B}_\epsilon$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_F \le \epsilon$. Using the triangle inequality to decompose the deviation:

$$\begin{aligned} \left\| \nabla^2 \widehat{\mathcal{R}}_e(\boldsymbol{\beta}) - \nabla^2 \mathcal{R}_e(\boldsymbol{\beta}) \right\|_2 &\le \left\| \nabla^2 \widehat{\mathcal{R}}_e(\boldsymbol{\beta}) - \nabla^2 \widehat{\mathcal{R}}_e(\boldsymbol{\beta}') \right\|_2 \\ &\quad + \left\| \nabla^2 \widehat{\mathcal{R}}_e(\boldsymbol{\beta}') - \nabla^2 \mathcal{R}_e(\boldsymbol{\beta}') \right\|_2 + \left\| \nabla^2 \mathcal{R}_e(\boldsymbol{\beta}') - \nabla^2 \mathcal{R}_e(\boldsymbol{\beta}) \right\|_2. \end{aligned}$$

By Assumption 8, the Hessian matrix of the multi-class cross-entropy is L_H-Lipschitz continuous. Thus, both the empirical and population deviations are controlled:

$$\begin{aligned} \left\| \nabla^2 \widehat{\mathcal{R}}_e(\boldsymbol{\beta}) - \nabla^2 \widehat{\mathcal{R}}_e(\boldsymbol{\beta}') \right\|_2 &\le L_H \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_F \le L_H \epsilon \\ \left\| \nabla^2 \mathcal{R}_e(\boldsymbol{\beta}) - \nabla^2 \mathcal{R}_e(\boldsymbol{\beta}') \right\|_2 &\le L_H \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_F \le L_H \epsilon \end{aligned}$$

The triangle inequality and the preceding bounds give:

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}} \left\| \nabla^2 \widehat{\mathcal{R}}_e(\boldsymbol{\beta}) - \nabla^2 \mathcal{R}_e(\boldsymbol{\beta}) \right\|_2 \le \sup_{\boldsymbol{\beta}' \in \mathcal{B}_\epsilon} \left\| \nabla^2 \widehat{\mathcal{R}}_e(\boldsymbol{\beta}') - \nabla^2 \mathcal{R}_e(\boldsymbol{\beta}') \right\|_2 + 2L_H \epsilon$$

Choice of parameters. To bound the error over the entire space by $\frac{\mu}{2}$, we set $2L_H\epsilon = \frac{\mu}{4}$, yielding the grid precision $\epsilon = \frac{\mu}{8L_H}$. Accordingly, the deviation threshold at the grid points is required to be $t_H = \frac{\mu}{4}$. Substituting this into the pointwise concentration bound and requiring the failure probability to be at most δ gives:

$$2d_\beta \left(1 + \frac{16R_\beta L_H}{\mu} \right)^{d_\beta} \exp\left(-\frac{n_e(\mu/4)^2}{2M^2} \right) \le \delta$$

Taking the natural logarithm on both sides and rearranging to solve for the lower bound of the sample size n_e:

$$n_e \geq \frac{32M^2}{\mu^2}\left(d_\beta \log\left(1+\frac{16R_\beta L_H}{\mu}\right)+\log(2d_\beta)+\log\left(\frac{1}{\delta}\right)\right)$$

Equivalently, up to a universal constant depending only on the uniform Hessian bound,

$$n_e \geq \frac{C_\star}{\mu^2}\left[d_\beta \log\left(1+\frac{R_\beta L_H}{\mu}\right)+\log\frac{d_\beta}{\delta}\right].$$

Under this sample size condition, the displayed uniform Hessian bound holds with probability at least $1-\delta$. □

C.4 Proof of Lemma 2

Proof. **Constrained quadratic control.** Set

$$\Delta_e = \widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}.$$

We set the probability decay level to $\delta = n_e^{-k}$ for some $k \geq 2$. Under the stated coverage-complexity condition, Lemma 1 holds for sufficiently large n_e. Let $\mathcal{E}$ denote the uniform Hessian event in Lemma 1. On $\mathcal{E}$, the empirical risk is $\mu/2$-strongly convex on $\mathcal{B}$. Therefore,

$$\widehat{\mathcal{R}}_e(\widehat{\boldsymbol{\beta}}^{(e)}) \geq \widehat{\mathcal{R}}_e(\boldsymbol{\beta}_*^{(e)}) + \left\langle \nabla\widehat{\mathcal{R}}_e(\boldsymbol{\beta}_*^{(e)}), \Delta_e \right\rangle + \frac{\mu}{4}\|\Delta_e\|_F^2.$$

Because $\widehat{\boldsymbol{\beta}}^{(e)}$ minimizes $\widehat{\mathcal{R}}_e$ over the convex set $\mathcal{B}$ and $\boldsymbol{\beta}_*^{(e)} \in \mathcal{B}$,

$$\widehat{\mathcal{R}}_e(\widehat{\boldsymbol{\beta}}^{(e)}) \leq \widehat{\mathcal{R}}_e(\boldsymbol{\beta}_*^{(e)}).$$

On $\mathcal{E}$,

$$\frac{\mu}{4}\|\Delta_e\|_F^2 \leq -\left\langle \nabla\widehat{\mathcal{R}}_e(\boldsymbol{\beta}_*^{(e)}), \Delta_e \right\rangle \leq \|\nabla\widehat{\mathcal{R}}_e(\boldsymbol{\beta}_*^{(e)})\|_F\|\Delta_e\|_F,$$

and hence

$$\|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F \leq \frac{4}{\mu}\|\nabla\widehat{\mathcal{R}}_e(\boldsymbol{\beta}_*^{(e)})\|_F.$$

This argument uses only constrained optimality and does not require the empirical minimizer to be an interior point.

Expected empirical gradient size. Let

$$\boldsymbol{G}_i = \nabla_{\boldsymbol{\beta}}\ell_{\mathrm{CE}}(\boldsymbol{\beta}_*^{(e)}\boldsymbol{Z}_{c,i}^{(e)}, Y_i^{(e)}), \qquad \bar{\boldsymbol{G}}_e = \frac{1}{n_e}\sum_{i=1}^{n_e}\boldsymbol{G}_i = \nabla\widehat{\mathcal{R}}_e(\boldsymbol{\beta}_*^{(e)}).$$

Assumption 8 gives $\mathbb{E}\boldsymbol{G}_i = \nabla\mathcal{R}_e(\boldsymbol{\beta}_*^{(e)}) = \mathbf{0}$. Independence and the zero-mean gradients give, for $i \neq j$,

$$\mathbb{E}\langle \boldsymbol{G}_i, \boldsymbol{G}_j\rangle_F = \langle \mathbb{E}\boldsymbol{G}_i, \mathbb{E}\boldsymbol{G}_j\rangle_F = 0.$$

Expanding the squared empirical gradient yields

$$\begin{aligned}
\mathbb{E}\|\bar{\boldsymbol{G}}_e\|_F^2 &= \frac{1}{n_e^2}\mathbb{E}\left\langle \sum_i \boldsymbol{G}_i, \sum_j \boldsymbol{G}_j \right\rangle_F \\
&= \frac{1}{n_e^2}\sum_{i=1}^{n_e}\mathbb{E}\|\boldsymbol{G}_i\|_F^2 + \frac{1}{n_e^2}\sum_{i\neq j}\mathbb{E}\langle \boldsymbol{G}_i, \boldsymbol{G}_j\rangle_F \\
&= \frac{1}{n_e^2}\sum_{i=1}^{n_e}\mathbb{E}\|\boldsymbol{G}_i\|_F^2 \\
&= \frac{1}{n_e}\mathbb{E}\left\|\nabla_{\boldsymbol{\beta}}\ell_{\mathrm{CE}}(\boldsymbol{\beta}_*^{(e)}\boldsymbol{Z}_c, Y)\right\|_F^2.
\end{aligned}$$

For multi-class cross-entropy, with $\boldsymbol{p}_{\boldsymbol{\beta}}(\boldsymbol{Z}_c) = \mathrm{Softmax}(\boldsymbol{\beta}\boldsymbol{Z}_c)$ and one-hot label vector $\boldsymbol{e}_Y$,

$$\nabla_{\boldsymbol{\beta}}\ell_{\mathrm{CE}}(\boldsymbol{\beta}\boldsymbol{Z}_c, Y) = (\boldsymbol{p}_{\boldsymbol{\beta}}(\boldsymbol{Z}_c) - \boldsymbol{e}_Y)\boldsymbol{Z}_c^\top,$$

The probability-vector identity and the outer-product norm give

$$\|\boldsymbol{p}_{\boldsymbol{\beta}}(\boldsymbol{Z}_c) - \boldsymbol{e}_Y\|_2^2 = \|\boldsymbol{p}_{\boldsymbol{\beta}}(\boldsymbol{Z}_c)\|_2^2 + 1 - 2[\boldsymbol{p}_{\boldsymbol{\beta}}(\boldsymbol{Z}_c)]_Y \le 2,$$
$$\|\nabla_{\boldsymbol{\beta}} \ell_{\mathrm{CE}}(\boldsymbol{\beta}\boldsymbol{Z}_c, Y)\|_F = \|\boldsymbol{p}_{\boldsymbol{\beta}}(\boldsymbol{Z}_c) - \boldsymbol{e}_Y\|_2 \|\boldsymbol{Z}_c\|_2 \le \sqrt{2} R_c.$$

The preceding second-moment identity and Cauchy–Schwarz yield

$$\mathbb{E}\|\nabla \widehat{\mathcal{R}}_e(\boldsymbol{\beta}_*^{(e)})\|_F^2 \le \frac{2R_c^2}{n_e},$$
$$\mathbb{E}\|\nabla \widehat{\mathcal{R}}_e(\boldsymbol{\beta}_*^{(e)})\|_F \le \left(\mathbb{E}\|\nabla \widehat{\mathcal{R}}_e(\boldsymbol{\beta}_*^{(e)})\|_F^2\right)^{1/2} \le \frac{\sqrt{2}R_c}{\sqrt{n_e}}.$$

Expectation over the high-probability event. Let $X_e = \|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F$. On $\mathcal{E}$,

$$\mathbb{E}[X_e \mathbf{1}_{\mathcal{E}}] \le \frac{4}{\mu}\mathbb{E}\|\nabla \widehat{\mathcal{R}}_e(\boldsymbol{\beta}_*^{(e)})\|_F \le \frac{4\sqrt{2}R_c}{\mu\sqrt{n_e}}.$$

On $\mathcal{E}^c$, compactness gives $X_e \le 2R_\beta$ and $\mathbb{P}(\mathcal{E}^c) \le n_e^{-k}$, so

$$\mathbb{E}[X_e \mathbf{1}_{\mathcal{E}^c}] \le 2R_\beta n_e^{-k}.$$

The polynomial tail is lower order for $k \ge 2$, which gives

$$\mathbb{E}\|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F \le \mathcal{O}\left(\frac{R_c}{\mu\sqrt{n_e}}\right).$$

□

C.5 Proof of Lemma 3

Proof. Let $\mathcal{E}_e$ be the uniform Hessian event of Lemma 1 for source e. On this event, the empirical risk in source domain e is $\mu/2$-strongly convex on $\mathcal{B}$. The constrained-optimality argument in Lemma 2 gives

$$\|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F \le \frac{4}{\mu}\|\nabla \widehat{\mathcal{R}}_e(\boldsymbol{\beta}_*^{(e)})\|_F.$$

Choose the Hessian failure level $\delta/(2E)$ in Lemma 1 and union bound over $e = 1, \dots, E$; the stated lower bound on $n_{\min}$ ensures $\mathbb{P}(\cap_e \mathcal{E}_e) \ge 1 - \delta/2$.

It remains to control the empirical gradients. Let $\boldsymbol{g}_i^{(e)} \in \mathbb{R}^{d_\beta}$ be the coordinates of $\nabla_{\boldsymbol{\beta}} \ell_{\mathrm{CE}}(\boldsymbol{\beta}_*^{(e)} \boldsymbol{Z}_{c,i}^{(e)}, Y_i^{(e)})$ in an orthonormal basis of $\mathcal{S}_\beta$. Assumption 8 gives $\mathbb{E}\boldsymbol{g}_i^{(e)} = \mathbf{0}$, and the bounded-feature assumption gives $\|\boldsymbol{g}_i^{(e)}\|_2 \le \sqrt{2}R_c$ for multi-class cross-entropy. Write $\bar{\boldsymbol{g}}^{(e)} = n_e^{-1}\sum_i \boldsymbol{g}_i^{(e)}$. Let $\mathcal{V}$ be a $1/2$-net of the unit sphere in $\mathbb{R}^{d_\beta}$, with $|\mathcal{V}| \le 5^{d_\beta}$. For any $\boldsymbol{v} \ne \mathbf{0}$, choose $\boldsymbol{u} \in \mathcal{V}$ within $1/2$ of $\boldsymbol{v}/\|\boldsymbol{v}\|_2$. Then

$$\begin{aligned}
\|\boldsymbol{v}\|_2 &= \left(\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}\right)^\top \boldsymbol{v} \\
&\le |\boldsymbol{u}^\top \boldsymbol{v}| + \left\|\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2} - \boldsymbol{u}\right\|_2 \|\boldsymbol{v}\|_2 \qquad \|\boldsymbol{v}\|_2 \le 2\max_{\boldsymbol{u}\in\mathcal{V}} |\boldsymbol{u}^\top \boldsymbol{v}|. \\
&\le \max_{\boldsymbol{u}\in\mathcal{V}} |\boldsymbol{u}^\top \boldsymbol{v}| + \tfrac{1}{2}\|\boldsymbol{v}\|_2,
\end{aligned}$$

For each $\boldsymbol{u} \in \mathcal{V}$, the scalar summands have mean zero and $|\boldsymbol{u}^\top \boldsymbol{g}_i^{(e)}| \le \sqrt{2}R_c$. Scalar Hoeffding and a union bound give

$$\Pr(|\boldsymbol{u}^\top \bar{\boldsymbol{g}}^{(e)}| > t) \le 2\exp\left(-\frac{n_e t^2}{4R_c^2}\right),$$
$$\Pr(\|\bar{\boldsymbol{g}}^{(e)}\|_2 > s) \le \sum_{\boldsymbol{u}\in\mathcal{V}} \Pr(|\boldsymbol{u}^\top \bar{\boldsymbol{g}}^{(e)}| > s/2)$$
$$\le 2 \cdot 5^{d_\beta} \exp\left(-\frac{n_e s^2}{16R_c^2}\right).$$

At $s = 4R_c\sqrt{\{d_\beta \log 5 + \log(4E/\delta)\}/n_e}$,

$$2 \cdot 5^{d_\beta} \exp\{-d_\beta \log 5 - \log(4E/\delta)\} = \frac{\delta}{2E}.$$

Absorbing numerical constants gives

$$\Pr\left(\left\|\frac{1}{n_e}\sum_i \boldsymbol{g}_i^{(e)}\right\|_2 > C_\star R_c\sqrt{\frac{d_\beta + \log(2E/\delta)}{n_e}}\right) \le \frac{\delta}{2E}.$$

A second union bound shows that, with probability at least $1-\delta$, both the Hessian event and the gradient event hold for all sources. Since $n_e \ge n_{\min}$, on this event

$$Y := \max_{1\le e\le E} \|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F \le C_\star \frac{R_c}{\mu}\sqrt{\frac{d_\beta + \log(E/\delta)}{n_{\min}}},$$

which proves the high-probability statement after absorbing constants.

For the expectation bounds, set $\delta = e^{-t}$. The high-probability sample-size condition then has the form

$$n_{\min} \ge \frac{C_\star}{\mu^2}\left[d_\beta \log\left(1 + \frac{R_\beta L_H}{\mu}\right) + \log(d_\beta E) + t\right].$$

Equivalently, after adjusting constants, this condition is satisfied whenever

$$0 \le t \le t_{\max}, \qquad t_{\max} = c_\star \mu^2 n_{\min} - d_\beta \log\left(1 + \frac{R_\beta L_H}{\mu}\right) - \log(c_{\star,1} d_\beta E).$$

Set

$$q(t) = C_\star \frac{R_c}{\mu}\sqrt{\frac{d_\beta + \log E + t}{n_{\min}}}, \qquad q'(t) = \frac{C_\star R_c}{2\mu\sqrt{n_{\min}}\sqrt{d_\beta + \log E + t}}.$$

The tail bound and compactness give

$$\Pr(Y > q(t)) \le e^{-t} \quad (0 \le t \le t_{\max}), \qquad 0 \le Y \le 2R_\beta.$$

Splitting the first-moment integral at $q(0)$ and $q(t_{\max})$ yields

$$\begin{aligned}
\mathbb{E}Y &= \int_0^{2R_\beta} \Pr(Y > u)\, du \\
&\le q(0) + \int_{q(0)}^{q(t_{\max})} \Pr(Y > u)\, du + 2R_\beta e^{-t_{\max}} \\
&= q(0) + \int_0^{t_{\max}} \Pr(Y > q(t)) q'(t)\, dt + 2R_\beta e^{-t_{\max}} \\
&\le q(0) + \frac{C_\star R_c}{2\mu\sqrt{n_{\min}}} \int_0^{t_{\max}} \frac{e^{-t}}{\sqrt{d_\beta + \log E + t}}\, dt + 2R_\beta e^{-t_{\max}} \\
&\le q(0) + \frac{C_\star R_c}{2\mu\sqrt{n_{\min}(d_\beta + \log E)}} \int_0^{\infty} e^{-t}\, dt + 2R_\beta e^{-t_{\max}} \\
&\le \frac{3C_\star R_c}{2\mu}\sqrt{\frac{d_\beta + \log E}{n_{\min}}} + 2R_\beta e^{-t_{\max}}.
\end{aligned}$$

The last inequality uses $d_\beta + \log E \geq 1$. For the second moment,

$$\begin{aligned}
\mathbb{E}Y^2 &= \int_0^{2R_\beta} 2u \, \Pr(Y > u)\, du \\
&\leq q(0)^2 + \int_{q(0)}^{q(t_{\max})} 2u \, \Pr(Y > u)\, du + 4R_\beta^2 e^{-t_{\max}} \\
&= q(0)^2 + \int_0^{t_{\max}} 2q(t)q'(t) \Pr(Y > q(t))\, dt + 4R_\beta^2 e^{-t_{\max}} \\
&\leq q(0)^2 + \frac{C_\star^2 R_c^2}{\mu^2 n_{\min}} \int_0^{t_{\max}} e^{-t}\, dt + 4R_\beta^2 e^{-t_{\max}} \\
&\leq \frac{C_\star^2 R_c^2 (d_\beta + \log E + 1)}{\mu^2 n_{\min}} + 4R_\beta^2 e^{-t_{\max}} \\
&\leq \frac{2C_\star^2 R_c^2 (d_\beta + \log E)}{\mu^2 n_{\min}} + 4R_\beta^2 e^{-t_{\max}}.
\end{aligned}$$

The exponential remainder is

$$e^{-t_{\max}} = c_{\star,1} d_\beta E \left(1 + \frac{R_\beta L_H}{\mu}\right)^{d_\beta} e^{-c_\star \mu^2 n_{\min}}.$$

A sufficiently large constant in the stated sample-size condition ensures

$$\begin{aligned}
d_\beta \log\left(1 + \frac{R_\beta L_H}{\mu}\right) + \log(c_{\star,1} d_\beta E) &\leq \tfrac{c_\star}{2} \mu^2 n_{\min}, \\
t_{\max} &\geq \tfrac{c_\star}{2} \mu^2 n_{\min}, \\
e^{-t_{\max}} &\leq e^{-c_\star \mu^2 n_{\min}/2}.
\end{aligned}$$

For $x > 0$, both $x^{1/2} e^{-c_\star x/2}$ and $x e^{-c_\star x/2}$ are bounded. With $x = \mu^2 n_{\min}$,

$$e^{-t_{\max}} \leq \frac{C_\star}{\mu \sqrt{n_{\min}}}, \qquad e^{-t_{\max}} \leq \frac{C_\star}{\mu^2 n_{\min}}.$$

Absorbing fixed regularity constants gives

$$\mathbb{E}Y \leq \mathcal{O}\left(\frac{R_c}{\mu}\sqrt{\frac{d_\beta + \log E}{n_{\min}}}\right), \qquad \mathbb{E}Y^2 \leq \mathcal{O}\left(\frac{R_c^2 (d_\beta + \log E)}{\mu^2 n_{\min}}\right).$$

These bounds also cover growing E subject to the stated sample-size condition. □

C.6 Proof of Lemma 4

Proof. Let

$$X_e = \mathcal{W}_1(\nu_e, \widehat{\nu}_e).$$

For compactly supported measures in $\mathbb{R}^{d_s}$, empirical optimal-transport theory gives constants $C_\star, c_\star > 0$, independent of n_e, such that

$$\mathbb{E}X_e \leq C_\star \gamma(n_e, d_s), \tag{6}$$

where $\gamma(\cdot,\cdot)$ is the rate shorthand defined above (Fournier & Guillin, 2015; Villani, 2009). The same result gives the following centered tails:

$$\mathbb{P}(X_e > C_\star \gamma(n_e, d_s) + t) \leq \begin{cases} C_\star \exp(-c_\star n_e t^2), & d_s < 2, \\ C_\star \exp\left(-c_\star n_e t^2/(\log(1+n_e))^2\right), & d_s = 2, \\ C_\star \exp(-c_\star n_e t^{d_s}), & d_s > 2. \end{cases} \tag{7}$$

For $n_e \geq n_{\min}$, $\gamma(n_e, d_s) \leq C_\star \gamma(n_{\min}, d_s)$. Enlarging constants to cover small sample sizes, for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{1 \leq e \leq E} X_e > C_\star \gamma(n_{\min}, d_s) + t\right) \\
&\qquad \leq \sum_{e=1}^{E} \mathbb{P}\left(X_e > C_\star \gamma(n_e, d_s) + t\right).
\end{aligned} \tag{8}$$

Substituting (7) into (8), choose a sufficiently large constant in the threshold

$$T_E(n_{\min}, d_s) = \begin{cases} C_\star \sqrt{\frac{1+\log E}{n_{\min}}}, & d_s < 2, \\ C_\star \log(1+n_{\min}) \sqrt{\frac{1+\log E}{n_{\min}}}, & d_s = 2, \\ C_\star \left(\frac{1+\log E}{n_{\min}}\right)^{1/d_s}, & d_s > 2. \end{cases} \tag{9}$$

This threshold is $\Gamma(n_{\min}, E, d_s)$ up to constants. For $d_s > 2$, choose its constant so that $c_\star n_{\min} T_E^{d_s} \geq 2(1+\log E)$; the corresponding transformed thresholds below satisfy the same lower bound in the other two dimension regimes.

We now integrate the tail explicitly. Set

$$M_E = \max_{1 \leq e \leq E} X_e, \qquad a_E = C_\star \gamma(n_{\min}, d_s) + T_E(n_{\min}, d_s).$$

By the layer-cake identity,

$$\begin{aligned} \mathbb{E}M_E &= \int_0^\infty \mathbb{P}(M_E > s)\, ds \\ &\leq a_E + \int_{a_E}^\infty \mathbb{P}(M_E > s)\, ds. \end{aligned}$$

For example, in the case $d_s > 2$, write $s = C_\star \gamma(n_{\min}, d_s) + r$ and use (8):

$$\begin{aligned} \int_{a_E}^\infty \mathbb{P}(M_E > s)\, ds &\leq C_\star E \int_{T_E}^\infty \exp(-c_\star n_{\min} r^{d_s})\, dr \\ &= \frac{C_\star E (c_\star n_{\min})^{-1/d_s}}{d_s} \int_{c_\star n_{\min} T_E^{d_s}}^\infty e^{-u} u^{1/d_s - 1}\, du \\ &\leq C_\star \left(\frac{1+\log E}{n_{\min}}\right)^{1/d_s}. \end{aligned}$$

For $d_s > 2$, set $u_0 = c_\star n_{\min} T_E^{d_s}$. For $j \in \{1, 2\}$, $j/d_s - 1 \leq 0$ and $u_0 \geq 2(1+\log E) > 1$, so

$$E \int_{u_0}^\infty e^{-u} u^{j/d_s - 1}\, du \leq E \int_{u_0}^\infty e^{-u}\, du = E e^{-u_0} \leq e^{-2} E^{-1} \leq e^{-2}.$$

For $d_s \leq 2$, use $\ell_{d_s}(n_{\min})$ from Theorem 2 and the substitution

$$u = \frac{c_\star n_{\min} r^2}{\ell_{d_s}(n_{\min})^2}, \qquad r = \frac{\ell_{d_s}(n_{\min})}{\sqrt{c_\star n_{\min}}} u^{1/2}, \qquad dr = \frac{\ell_{d_s}(n_{\min})}{2\sqrt{c_\star n_{\min}}} u^{-1/2}\, du.$$

The transformed threshold is $u_0 = c_\star n_{\min} T_E^2 / \ell_{d_s}(n_{\min})^2 \geq 2(1+\log E)$. For $j \in \{1, 2\}$,

$$\begin{aligned} & E \int_{T_E}^\infty r^{j-1} \exp\left(-\frac{c_\star n_{\min} r^2}{\ell_{d_s}(n_{\min})^2}\right) dr \\ &= \frac{E}{2} \left(\frac{\ell_{d_s}(n_{\min})}{\sqrt{c_\star n_{\min}}}\right)^j \int_{u_0}^\infty e^{-u} u^{j/2-1}\, du \\ &\leq \frac{E}{2} \left(\frac{\ell_{d_s}(n_{\min})}{\sqrt{c_\star n_{\min}}}\right)^j e^{-u_0} \leq C_\star \gamma(n_{\min}, d_s)^j \leq C_\star \Gamma(n_{\min}, E, d_s)^j. \end{aligned}$$

The cases $d_s < 2$ and $d_s = 2$ correspond to $\ell_{d_s}(n_{\min}) = 1$ and $\log(1+n_{\min})$, respectively.

$$\mathbb{E}M_E \leq \mathcal{O}(\gamma(n_{\min}, d_s) + \Gamma(n_{\min}, E, d_s)). \tag{10}$$

For the second moment, the layer-cake identity gives

$$\begin{aligned} \mathbb{E}M_E^2 &= \int_0^\infty 2s\, \Pr(M_E > s)\, ds \\ &\leq a_E^2 + \int_{T_E}^\infty 2\{C_\star \gamma(n_{\min}, d_s) + r\} \Pr\{M_E > C_\star \gamma(n_{\min}, d_s) + r\}\, dr. \end{aligned}$$

For $d_s > 2$, the tail contribution is bounded by

$$2C_\star E \int_{T_E}^{\infty} \{C_\star \gamma(n_{\min}, d_s) + r\} e^{-c_\star n_{\min} r^{d_s}}\, dr$$
$$= 2C_\star^2 \gamma(n_{\min}, d_s) E \int_{T_E}^{\infty} e^{-c_\star n_{\min} r^{d_s}}\, dr + 2C_\star E \int_{T_E}^{\infty} r e^{-c_\star n_{\min} r^{d_s}}\, dr.$$

The first integral has already been bounded. For the second,

$$E \int_{T_E}^{\infty} r e^{-c_\star n_{\min} r^{d_s}}\, dr = \frac{E(c_\star n_{\min})^{-2/d_s}}{d_s} \int_{u_0}^{\infty} e^{-u} u^{2/d_s - 1}\, du$$
$$\le C_\star n_{\min}^{-2/d_s} \le C_\star \Gamma(n_{\min}, E, d_s)^2.$$

For $d_s \le 2$, the preceding substitutions with $j = 1, 2$ give the same bounds. In all three dimension regimes,

$$\mathbb{E} M_E^2 \le a_E^2 + C_\star \gamma(n_{\min}, d_s)\Gamma(n_{\min}, E, d_s) + C_\star \Gamma(n_{\min}, E, d_s)^2$$
$$\le 2C_\star^2 \gamma(n_{\min}, d_s)^2 + 2T_E^2 + \frac{C_\star}{2}\{\gamma(n_{\min}, d_s)^2 + \Gamma(n_{\min}, E, d_s)^2\} + C_\star \Gamma(n_{\min}, E, d_s)^2$$
$$\le C_\star \{\gamma(n_{\min}, d_s)^2 + \Gamma(n_{\min}, E, d_s)^2\}.$$
$$\mathbb{E} M_E^2 \le \mathcal{O}\big(\gamma(n_{\min}, d_s)^2 + \Gamma(n_{\min}, E, d_s)^2\big). \tag{11}$$

Equations (6), (10), and (11) prove the lemma. □

C.7 Proof of Lemma 5

Proof. We decouple the quantity in the lemma into target-domain empirical error and source-domain empirical errors:

$$L_\beta \sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e \{\mathcal{W}_1(\nu_e, \widehat{\nu}_e) + \mathcal{W}_1(\widehat{\nu}_0, \nu_0)\}$$
$$= L_\beta \sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e\, \mathcal{W}_1(\widehat{\nu}_0, \nu_0) + L_\beta \sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e\, \mathcal{W}_1(\nu_e, \widehat{\nu}_e).$$

Since $\mathcal{W}_1(\widehat{\nu}_0, \nu_0)$ does not depend on the source index and $\sum_e w_e = 1$,

$$\sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e \mathcal{W}_1(\widehat{\nu}_0, \nu_0) = \mathcal{W}_1(\widehat{\nu}_0, \nu_0) \sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e$$

$$\mathbb{E}\left[\sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e \mathcal{W}_1(\widehat{\nu}_0, \nu_0)\right] = \mathbb{E}\left[\mathcal{W}_1(\widehat{\nu}_0, \nu_0)\right] \le \mathcal{O}\left(\gamma(n_0, d_s)\right),$$

where the last inequality follows from Lemma 4.

For the source part, non-negativity and normalization of the weights imply the pointwise bound

$$\sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e \mathcal{W}_1(\nu_e, \widehat{\nu}_e) \le \max_{e \in \widehat{\mathcal{N}}_K(0)} \mathcal{W}_1(\nu_e, \widehat{\nu}_e) \le \max_{1 \le e \le E} \mathcal{W}_1(\nu_e, \widehat{\nu}_e)$$

and Lemma 4 gives

$$\mathbb{E}\left[\sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e \mathcal{W}_1(\nu_e, \widehat{\nu}_e)\right] \le \mathcal{O}\left(\gamma(n_{\min}, d_s) + \Gamma(n_{\min}, E, d_s)\right).$$

Using $\gamma(n_{\min}, d_s) \le \Gamma(n_{\min}, E, d_s)$ and adding the two parts gives

$$\mathbb{E}\left[L_\beta \sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e \{\mathcal{W}_1(\nu_e, \widehat{\nu}_e) + \mathcal{W}_1(\widehat{\nu}_0, \nu_0)\}\right] \le \mathcal{O}\Big(\gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s)\Big).$$

□

C.8 Proof of Lemma 6

Proof. **Weighted radius reduction.** Since $w_e \geq 0$ and $\sum w_e = 1$, the weighted convex combination is bounded by the maximum distance within the selected set. Let $\widehat{R}_K$ be the distance of the K-th nearest neighbor found by the algorithm in the empirical space (i.e., the empirical KNN radius):

$$L_\beta \sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e \mathcal{W}_1(\widehat{\nu}_e, \widehat{\nu}_0) \leq L_\beta \max_{e \in \widehat{\mathcal{N}}_K(0)} \mathcal{W}_1(\widehat{\nu}_e, \widehat{\nu}_0) := L_\beta \widehat{R}_K.$$

From empirical neighbors to true neighbors. We define the set $\mathcal{N}_K^*(0)$ as the K true nearest neighbors of the target domain ν_0 on the true measure manifold $\mathcal{M}$. By construction, $\widehat{\mathcal{N}}_K(0)$ consists of the K domains with the smallest empirical distances among all E source domains. Thus, the maximum empirical distance among these K domains, $\widehat{R}_K$, is the K-th order statistic of empirical distances across the global set. Conversely, under the empirical metric, the true nearest-neighbor set $\mathcal{N}_K^*(0)$ is another subset of size K among the E domains. The maximum over any subset of size K is at least the maximum over the K empirically closest domains. Hence, the following inequality holds deterministically:

$$\widehat{R}_K = \max_{e \in \widehat{\mathcal{N}}_K(0)} \mathcal{W}_1(\widehat{\nu}_e, \widehat{\nu}_0) \leq \max_{j \in \mathcal{N}_K^*(0)} \mathcal{W}_1(\widehat{\nu}_j, \widehat{\nu}_0)$$

Applying the Wasserstein distance triangle inequality to the true neighbor $j \in \mathcal{N}_K^*(0)$ and leveraging the sub-additivity of the maximum operator:

$$\mathcal{W}_1(\widehat{\nu}_j, \widehat{\nu}_0) \leq \mathcal{W}_1(\widehat{\nu}_j, \nu_j) + \mathcal{W}_1(\nu_j, \nu_0) + \mathcal{W}_1(\nu_0, \widehat{\nu}_0)$$

$$\widehat{R}_K \leq R_K^* + \max_{j \in \mathcal{N}_K^*(0)} \mathcal{W}_1(\widehat{\nu}_j, \nu_j) + \mathcal{W}_1(\nu_0, \widehat{\nu}_0),$$

where $R_K^* := \max_{j \in \mathcal{N}_K^*(0)} \mathcal{W}_1(\nu_j, \nu_0)$. Taking the mathematical expectation on both sides, the latter two terms are bounded by $\mathcal{O}\left(\gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s)\right)$ as established in Lemma 5. The remaining task is to evaluate $\mathbb{E}[R_K^*]$.

Order statistic of the true KNN radius. Condition on the target law ν_0. For $R_e = \mathcal{W}_1(\nu_e, \nu_0)$,

$$F_{\nu_0}(r) = \Pr(R_e \leq r \mid \nu_0) = \Pi_{\mathcal{M}}(B_{\mathcal{W}_1}(\nu_0, r)).$$

All distances are at most D_s. Uniform Ahlfors regularity provides $a > 0$ and $0 < r_0 \leq D_s$ with $F_{\nu_0}(r) \geq a r^{d_{\mathcal{M}}}$ for $0 \leq r \leq r_0$. For larger radii,

$$F_{\nu_0}(r) \geq F_{\nu_0}(r_0) \geq a r_0^{d_{\mathcal{M}}} \geq a (r_0/D_s)^{d_{\mathcal{M}}} r^{d_{\mathcal{M}}}, \quad r_0 < r \leq D_s.$$

With $a_0 = a(r_0/D_s)^{d_{\mathcal{M}}}$, this gives a global bound

$$F_{\nu_0}(r) \geq a_0 r^{d_{\mathcal{M}}} \quad (0 \leq r \leq D_s), \qquad R_K^* \leq a_0^{-1/d_{\mathcal{M}}} F_{\nu_0}(R_K^*)^{1/d_{\mathcal{M}}}.$$

The assumed continuity of F_{ν_0} implies that $U_e = F_{\nu_0}(R_e)$ are conditionally i.i.d. Uniform$(0, 1)$. Monotonicity of F_{ν_0} gives $U_{(K)} = F_{\nu_0}(R_K^*)$. For $0 < u < 1$,

$$\begin{aligned}
\Pr(U_{(K)} \leq u \mid \nu_0) &= \Pr\left(\sum_{e=1}^{E} \mathbf{1}_{\{U_e \leq u\}} \geq K \,\middle|\, \nu_0 \right) \\
&= \sum_{j=K}^{E} \binom{E}{j} u^j (1-u)^{E-j}, \\
f_{U_{(K)}|\nu_0}(u) &= \frac{d}{du} \sum_{j=K}^{E} \binom{E}{j} u^j (1-u)^{E-j} \\
&= \frac{E!}{(K-1)!(E-K)!} u^{K-1} (1-u)^{E-K}.
\end{aligned}$$

Its first moment is

$$\mathbb{E}[U_{(K)} \mid \nu_0] = \frac{E!}{(K-1)!(E-K)!} \int_0^1 u^K (1-u)^{E-K}\, du$$
$$= \frac{E!}{(K-1)!(E-K)!} \frac{K!(E-K)!}{(E+1)!} = \frac{K}{E+1}.$$

Since $d_{\mathcal{M}} \in \mathbb{N}$, $u^{1/d_{\mathcal{M}}}$ is concave. The radius bound and Jensen's inequality yield

$$\mathbb{E}[R_K^* \mid \nu_0] \le a_0^{-1/d_{\mathcal{M}}} \mathbb{E}[U_{(K)}^{1/d_{\mathcal{M}}} \mid \nu_0]$$
$$\le a_0^{-1/d_{\mathcal{M}}} \{\mathbb{E}[U_{(K)} \mid \nu_0]\}^{1/d_{\mathcal{M}}}$$
$$= a_0^{-1/d_{\mathcal{M}}} \left(\frac{K}{E+1}\right)^{1/d_{\mathcal{M}}} \le C_\star (K/E)^{1/d_{\mathcal{M}}}.$$

The constants are uniform in ν_0. For an independently sampled target,

$$\mathbb{E}R_K^* = \mathbb{E}_{\nu_0}\mathbb{E}[R_K^* \mid \nu_0] \le C_\star (K/E)^{1/d_{\mathcal{M}}}.$$

The same bound holds for a fixed target satisfying the stated regularity condition.

The empirical-measure and nearest-neighbor radius bounds give:

$$\mathbb{E}\left[L_\beta \sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e \mathcal{W}_1(\widehat{\nu}_e, \widehat{\nu}_0)\right] \le L_\beta \mathcal{O}\left(\left(\frac{K}{E}\right)^{\frac{1}{d_{\mathcal{M}}}}\right) + L_\beta \mathcal{O}\left(\gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s)\right).$$

External Lipschitz constants are absorbed, completing the proof. □

C.9 Proof of Theorem 2

Proof. Write the truncated reweighted estimator as

$$\widehat{\boldsymbol{\beta}}^{(0)} = \sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e \widehat{\boldsymbol{\beta}}^{(e)}, \qquad w_e \ge 0, \qquad \sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e = 1.$$

The normalization of the weights allows us to write

$$\boldsymbol{\beta}_*^{(0)} = \sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e \boldsymbol{\beta}_*^{(0)}.$$

Adding and subtracting $\boldsymbol{\beta}_*^{(e)}$ inside each summand gives the exact identity

$$\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)} = \sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e \left(\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(0)}\right)$$
$$= \sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e \left[\left(\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\right) + \left(\boldsymbol{\beta}_*^{(e)} - \boldsymbol{\beta}_*^{(0)}\right)\right].$$

Taking Frobenius norms, applying the triangle inequality, and using $w_e \ge 0$ yields

$$\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\|_F \le \sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e \|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F + \sum_{e \in \widehat{\mathcal{N}}_K(0)} w_e \|\boldsymbol{\beta}_*^{(e)} - \boldsymbol{\beta}_*^{(0)}\|_F.$$

Assumption 7 gives $\boldsymbol{\beta}_*^{(e)} = \Psi^*(\nu_e)$ and $\boldsymbol{\beta}_*^{(0)} = \Psi^*(\nu_0)$, so

$$\|\boldsymbol{\beta}_*^{(e)} - \boldsymbol{\beta}_*^{(0)}\|_F = \|\Psi^*(\nu_e) - \Psi^*(\nu_0)\|_F \le L_\beta \mathcal{W}_1(\nu_e, \nu_0).$$

For each selected source domain, the Wasserstein triangle inequality gives

$$\mathcal{W}_1(\nu_e, \nu_0) \le \mathcal{W}_1(\nu_e, \widehat{\nu}_e) + \mathcal{W}_1(\widehat{\nu}_e, \widehat{\nu}_0) + \mathcal{W}_1(\widehat{\nu}_0, \nu_0).$$

The resulting three-term decomposition is:

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\|_F \le & \underbrace{\sum_{e\in\widehat{\mathcal{N}}_K(0)} w_e \|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F}_{\text{Term A}} \\ & \underbrace{+ L_\beta \sum_{e\in\widehat{\mathcal{N}}_K(0)} w_e \Big(\mathcal{W}_1(\nu_e, \widehat{\nu}_e) + \mathcal{W}_1(\widehat{\nu}_0, \nu_0)\Big)}_{\text{Term B}} \\ & \underbrace{+ L_\beta \sum_{e\in\widehat{\mathcal{N}}_K(0)} w_e \mathcal{W}_1(\widehat{\nu}_e, \widehat{\nu}_0)}_{\text{Term C}}.\end{aligned}$$

Taking the unconditional mathematical expectation $\mathbb{E}[\cdot]$ on both sides, we bound the three terms separately.

For **Term A**, the possible dependence between the empirically computed neighbor set $\widehat{\mathcal{N}}_K(0)$ and the parameter estimators is handled by the deterministic maximum bound

$$\sum_{e\in\widehat{\mathcal{N}}_K(0)} w_e \|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F \le \max_{e\in\widehat{\mathcal{N}}_K(0)} \|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F \le \max_{1\le e\le E} \|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F$$

Lemma 3 then directly implies

$$\mathbb{E}[\text{Term A}] \le \mathcal{O}\left(\frac{R_c}{\mu}\sqrt{\frac{d_\beta + \log E}{n_{\min}}}\right).$$

For **Term B**, Lemma 5 gives the explicit empirical-measure calculation

$$\begin{aligned}\mathbb{E}[\text{Term B}] &= L_\beta\, \mathbb{E}\left[\sum_{e\in\widehat{\mathcal{N}}_K(0)} w_e \left\{\mathcal{W}_1(\nu_e, \widehat{\nu}_e) + \mathcal{W}_1(\widehat{\nu}_0, \nu_0)\right\}\right] \\ &\le L_\beta \left[\mathbb{E}\max_{1\le e\le E} \mathcal{W}_1(\nu_e, \widehat{\nu}_e) + \mathbb{E}\mathcal{W}_1(\widehat{\nu}_0, \nu_0)\right] \\ &\le \mathcal{O}\left(\Gamma(n_{\min}, E, d_s) + \gamma(n_0, d_s)\right).\end{aligned}$$

The first inequality uses $w_e \ge 0$ and $\sum_e w_e = 1$; the second uses Lemma 4.

For **Term C**, Lemma 6 gives

$$\begin{aligned}\mathbb{E}[\text{Term C}] &= L_\beta\, \mathbb{E}\left[\sum_{e\in\widehat{\mathcal{N}}_K(0)} w_e \mathcal{W}_1(\widehat{\nu}_e, \widehat{\nu}_0)\right] \\ &\le L_\beta\, \mathbb{E}\widehat{R}_K \\ &\le L_\beta \mathbb{E} R_K^* + L_\beta \mathbb{E}\max_{1\le e\le E} \mathcal{W}_1(\widehat{\nu}_e, \nu_e) + L_\beta \mathbb{E}\mathcal{W}_1(\widehat{\nu}_0, \nu_0) \\ &\le \mathcal{O}\left(\left(\frac{K}{E}\right)^{1/d_{\mathcal{M}}} + \Gamma(n_{\min}, E, d_s) + \gamma(n_0, d_s)\right).\end{aligned}$$

Here $\widehat{R}_K$ is the empirical top-K radius and R_K^* is the oracle top-K radius on the population space of style distributions; the last line uses the order-statistic calculation in Lemma 6.

Summing the three expected bounds gives

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\right\|_F\right] \le & \mathcal{O}\left(\frac{R_c}{\mu}\sqrt{\frac{d_\beta + \log E}{n_{\min}}}\right) + \mathcal{O}\left(\gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s)\right) \\ & + \mathcal{O}\left(\left(\frac{K}{E}\right)^{1/d_{\mathcal{M}}}\right).\end{aligned}$$

□

C.10 Proof of Theorem 3

Proof. Here $\mathcal{R}_0$ is the population cross-entropy risk of the fixed target domain, evaluated on a fresh observation independent of the samples used to construct the predictor and its target summary. The fitted predictor is held fixed inside this risk; the outer expectation averages the random quantities used to construct it. **Smoothness reduction.** By Assumption 8, $\boldsymbol{\beta}_*^{(0)}$ is a relative-interior population optimum and $\nabla_{\mathcal{S}_\beta}\mathcal{R}_0(\boldsymbol{\beta}_*^{(0)}) = \mathbf{0}$. Set

$$\Delta_0 := \widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}.$$

Since both $\widehat{\boldsymbol{\beta}}^{(0)}$ and $\boldsymbol{\beta}_*^{(0)}$ belong to $\mathcal{B} \subset \mathcal{S}_\beta$, we have $\Delta_0 \in \mathcal{S}_\beta$. Hence the first-order term vanishes in the centered-gauge tangent space:

$$\left\langle \nabla\mathcal{R}_0(\boldsymbol{\beta}_*^{(0)}), \Delta_0 \right\rangle = \left\langle \nabla_{\mathcal{S}_\beta}\mathcal{R}_0(\boldsymbol{\beta}_*^{(0)}), \Delta_0 \right\rangle = 0.$$

A second-order Taylor expansion of the target-domain expected risk at $\boldsymbol{\beta}_*^{(0)}$ gives

$$\mathcal{R}_0(\widehat{\boldsymbol{\beta}}^{(0)}) - \mathcal{R}_0(\boldsymbol{\beta}_*^{(0)}) = \frac{1}{2}\,\mathrm{vec}(\Delta_0)^\top \nabla^2\mathcal{R}_0(\tilde{\boldsymbol{\beta}})\,\mathrm{vec}(\Delta_0)$$

where $\tilde{\boldsymbol{\beta}}$ lies between $\widehat{\boldsymbol{\beta}}^{(0)}$ and $\boldsymbol{\beta}_*^{(0)}$. Based on standard properties of matrix operator quadratic forms, the second-order expansion term is bounded by the product of the population Hessian matrix's spectral norm and the squared L_2 norm of the vector:

$$\frac{1}{2}\mathrm{vec}(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)})^\top \nabla^2\mathcal{R}_0(\tilde{\boldsymbol{\beta}})\mathrm{vec}(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}) \le \frac{1}{2}\|\nabla^2\mathcal{R}_0(\tilde{\boldsymbol{\beta}})\|_2 \cdot \|\mathrm{vec}(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)})\|_2^2$$

We note that the single-sample Hessian matrix structure for the multi-class cross-entropy loss is:

$$\nabla^2_{\boldsymbol{\beta}}\ell_{\mathrm{CE}}(\boldsymbol{\beta}\boldsymbol{Z}_c, Y) = (\boldsymbol{Z}_c\boldsymbol{Z}_c^\top) \otimes (\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top)$$

where $\otimes$ denotes the Kronecker product. Because the maximum eigenvalue of the covariance matrix on the simplex $\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top$ is bounded by $\frac{1}{2}$, and $\|\boldsymbol{Z}_c\|_2 \le R_c$, the spectral norm satisfies:

$$\|\nabla^2_{\boldsymbol{\beta}}\ell_{\mathrm{CE}}(\boldsymbol{\beta}\boldsymbol{Z}_c, Y)\|_2 = \|\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\|_2 \cdot \|\boldsymbol{Z}_c\boldsymbol{Z}_c^\top\|_2 \le \frac{1}{2}\|\boldsymbol{Z}_c\|_2^2 \le \frac{1}{2}R_c^2$$

By the sub-additivity and monotonicity of the expectation operator, the spectral norm of the population Hessian matrix is bounded over the compact convex set:

$$\|\nabla^2\mathcal{R}_0(\tilde{\boldsymbol{\beta}})\|_2 \le \mathbb{E}\left[\|\nabla^2_{\boldsymbol{\beta}}\ell_{\mathrm{CE}}(\tilde{\boldsymbol{\beta}}\boldsymbol{Z}_c, Y)\|_2\right] \le \frac{1}{2}R_c^2$$

Also, the L_2 norm of the vectorized matrix exactly equals the Frobenius norm of the original matrix:

$$\|\mathrm{vec}(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)})\|_2^2 = \left\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\right\|_F^2$$

The expected risk is $R_c^2/2$-smooth over the compact domain:

$$\mathcal{R}_0(\widehat{\boldsymbol{\beta}}^{(0)}) - \mathcal{R}_0(\boldsymbol{\beta}_*^{(0)}) \le \frac{R_c^2}{4}\left\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\right\|_F^2$$

Taking the external expectation gives the dominant inequality:

$$\mathbb{E}\left[\mathcal{R}_0(\widehat{\boldsymbol{\beta}}^{(0)}) - \mathcal{R}_0(\boldsymbol{\beta}_*^{(0)})\right] \le \frac{R_c^2}{4}\mathbb{E}\left[\left\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\right\|_F^2\right]$$

Second moment of the decomposed parameter error. We recall the triangle inequality of the error decoupling from Theorem 2:

$$\left\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\right\|_F \le \text{Term A} + \text{Term B} + \text{Term C}$$

Using the algebraic inequality $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, we scale the square:

$$\left\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\right\|_F^2 \le 3\left((\text{Term A})^2 + (\text{Term B})^2 + (\text{Term C})^2\right)$$

For $(\textbf{Term A})^2$, Lemma 3 gives the squared maximal estimator error:

$$\mathbb{E}\left[(\text{Term A})^2\right] \le \mathbb{E}\left[\max_{1\le e\le E} \|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F^2\right] \le \mathcal{O}\left(\frac{R_c^2(d_\beta + \log E)}{\mu^2 n_{\min}}\right)$$

For $(\textbf{Term B})^2$, use the same deterministic decomposition as in Lemma 5:

$$\text{Term B} \le L_\beta \left(\mathcal{W}_1(\widehat{\nu}_0, \nu_0) + \max_{1\le e\le E} \mathcal{W}_1(\nu_e, \widehat{\nu}_e)\right).$$

Using $(a+b)^2 \le 2(a^2+b^2)$ and Lemma 4,

$$\begin{aligned}
\mathbb{E}[(\text{Term B})^2] &\le 2L_\beta^2 \left[\mathbb{E}\mathcal{W}_1(\widehat{\nu}_0, \nu_0)^2 + \mathbb{E}\max_{1\le e\le E} \mathcal{W}_1(\nu_e, \widehat{\nu}_e)^2\right] \\
&\le C_\star L_\beta^2 \left[\gamma(n_0, d_s)^2 + \gamma(n_{\min}, d_s)^2 + \Gamma(n_{\min}, E, d_s)^2\right] \\
&\le \mathcal{O}\big(\gamma(n_0, d_s)^2 + \Gamma(n_{\min}, E, d_s)^2\big).
\end{aligned}$$

The last step uses $\gamma(n_{\min}, d_s) \le \Gamma(n_{\min}, E, d_s)$.

For $(\textbf{Term C})^2$, the proof of Lemma 6 gives

$$\text{Term C} \le L_\beta \left(R_K^* + \max_{1\le e\le E} \mathcal{W}_1(\widehat{\nu}_e, \nu_e) + \mathcal{W}_1(\widehat{\nu}_0, \nu_0)\right).$$

For the radius term, use $U_{(K)} = F_{\nu_0}(R_K^*)$ and $R_K^* \le a_0^{-1/d_\mathcal{M}} U_{(K)}^{1/d_\mathcal{M}}$ from the proof of Lemma 6. For $d_\mathcal{M} = 1$, the displayed order-statistic density gives

$$\begin{aligned}
\mathbb{E}[U_{(K)}^2 \mid \nu_0] &= \frac{E!}{(K-1)!(E-K)!} \int_0^1 u^{K+1}(1-u)^{E-K}\, du \\
&= \frac{E!}{(K-1)!(E-K)!} \frac{(K+1)!(E-K)!}{(E+2)!} \\
&= \frac{K(K+1)}{(E+1)(E+2)} \le 2(K/E)^2, \\
\mathbb{E}[(R_K^*)^2 \mid \nu_0] &\le a_0^{-2}\mathbb{E}[U_{(K)}^2 \mid \nu_0] \le 2a_0^{-2}(K/E)^2.
\end{aligned}$$

For $d_\mathcal{M} \ge 2$, $u^{2/d_\mathcal{M}}$ is concave, so

$$\begin{aligned}
\mathbb{E}[(R_K^*)^2 \mid \nu_0] &\le a_0^{-2/d_\mathcal{M}} \mathbb{E}[U_{(K)}^{2/d_\mathcal{M}} \mid \nu_0] \\
&\le a_0^{-2/d_\mathcal{M}} \{\mathbb{E}[U_{(K)} \mid \nu_0]\}^{2/d_\mathcal{M}} \\
&= a_0^{-2/d_\mathcal{M}} \left(\frac{K}{E+1}\right)^{2/d_\mathcal{M}} \le C_\star (K/E)^{2/d_\mathcal{M}}.
\end{aligned}$$

The uniform bounds cover a fixed target and, by iterated expectation, an independent random target:

$$\mathbb{E}(R_K^*)^2 \le C_\star (K/E)^{2/d_\mathcal{M}}.$$

Applying $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ to Term C gives

$$\begin{aligned}
\mathbb{E}[(\text{Term C})^2] &\le 3L_\beta^2 \left[\mathbb{E}(R_K^*)^2 + \mathbb{E}\max_{1\le e\le E} \mathcal{W}_1(\widehat{\nu}_e, \nu_e)^2 + \mathbb{E}\mathcal{W}_1(\widehat{\nu}_0, \nu_0)^2\right] \\
&\le C_\star L_\beta^2 \left[(K/E)^{2/d_\mathcal{M}} + \gamma(n_{\min}, d_s)^2 + \Gamma(n_{\min}, E, d_s)^2 + \gamma(n_0, d_s)^2\right] \\
&\le \mathcal{O}\Big((K/E)^{2/d_\mathcal{M}} + \Gamma(n_{\min}, E, d_s)^2 + \gamma(n_0, d_s)^2\Big).
\end{aligned}$$

The three bounds yield the parameter second moment

$$\begin{aligned}
\mathbb{E}\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\|_F^2 &\le 3\left\{\mathbb{E}[(\text{Term A})^2] + \mathbb{E}[(\text{Term B})^2] + \mathbb{E}[(\text{Term C})^2]\right\} \\
&\le C_\star \left[\frac{R_c^2(d_\beta + \log E)}{\mu^2 n_{\min}} + \gamma(n_0, d_s)^2 + \Gamma(n_{\min}, E, d_s)^2 + (K/E)^{2/d_\mathcal{M}}\right].
\end{aligned}$$

Substitution into the smoothness bound gives

$$\begin{aligned}
&\mathbb{E}\{\mathcal{R}_0(\widehat{\boldsymbol{\beta}}^{(0)}) - \mathcal{R}_0(\boldsymbol{\beta}_*^{(0)})\} \\
&\le \frac{R_c^2}{4}\mathbb{E}\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\|_F^2 \\
&\le \frac{C_\star R_c^2}{4}\left[\frac{R_c^2(d_\beta + \log E)}{\mu^2 n_{\min}} + \gamma(n_0, d_s)^2 + \Gamma(n_{\min}, E, d_s)^2 + (K/E)^{2/d_{\mathcal{M}}}\right] \\
&\le \frac{C_\star R_c^2}{4}\left[\frac{R_c}{\mu}\sqrt{\frac{d_\beta + \log E}{n_{\min}}} + \gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s) + (K/E)^{1/d_{\mathcal{M}}}\right]^2 \\
&\le \mathcal{O}\left(R_c^2\left[\frac{R_c}{\mu}\sqrt{\frac{d_\beta + \log E}{n_{\min}}} + \gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s) + (K/E)^{1/d_{\mathcal{M}}}\right]^2\right).
\end{aligned}$$

The penultimate step uses $\sum_j a_j^2 \le (\sum_j a_j)^2$ for nonnegative a_j. □

C.11 Proof of Lemma 7

Proof. For a fixed pair $(\boldsymbol{Z}_c, c)$, define the conditional cross-entropy risk as a function of the logit vector $\boldsymbol{u} \in \mathbb{R}^C$:

$$\rho_{\boldsymbol{Z}_c,c}(\boldsymbol{u}) = \mathbb{E}[\ell_{\mathrm{CE}}(\boldsymbol{u}, Y) \mid \boldsymbol{Z}_c, c]\,.$$

Because the softmax probabilities are invariant under common logit translations,

$$\mathrm{Softmax}(\boldsymbol{u} + a\mathbf{1}) = \mathrm{Softmax}(\boldsymbol{u}) \qquad \text{for every } a \in \mathbb{R},$$

the conditional risk depends only on the centered component of the logits. Hence

$$\rho_{\boldsymbol{Z}_c,c}(\boldsymbol{u}) = \rho_{\boldsymbol{Z}_c,c}(\boldsymbol{\Pi}_C\boldsymbol{u}).$$

Let $\boldsymbol{u} = \boldsymbol{\Pi}_C f(\boldsymbol{Z})$ and $\boldsymbol{u}^* = \boldsymbol{\Pi}_C f^*(\boldsymbol{Z}_c, c)$. For the lower bound, assume $\boldsymbol{u}, \boldsymbol{u}^* \in \mathcal{U}_R$ almost surely. By Assumption 8, $\rho_{\boldsymbol{Z}_c,c}$ is μ_ℓ-strongly convex on $\mathcal{U}_R$. Therefore,

$$\rho_{\boldsymbol{Z}_c,c}(\boldsymbol{u}) \ge \rho_{\boldsymbol{Z}_c,c}(\boldsymbol{u}^*) + \langle \nabla_{\mathcal{S}_C}\rho_{\boldsymbol{Z}_c,c}(\boldsymbol{u}^*), \boldsymbol{u} - \boldsymbol{u}^*\rangle + \frac{\mu_\ell}{2}\|\boldsymbol{u} - \boldsymbol{u}^*\|_2^2,$$

where $\nabla_{\mathcal{S}_C}$ is the gradient restricted to $\mathcal{S}_C$. Set $\boldsymbol{p}_* = \mathrm{Softmax}(f^*(\boldsymbol{Z}_c, c))$. The conditional label law gives

$$\begin{aligned}
\rho_{\boldsymbol{Z}_c,c}(\boldsymbol{u}) &= -\sum_{j=1}^{C} p_{*,j}\log\left(\frac{e^{u_j}}{\sum_k e^{u_k}}\right) \\
&= \log\left(\sum_k e^{u_k}\right) - \boldsymbol{p}_*^\top\boldsymbol{u}, \\
\nabla\rho_{\boldsymbol{Z}_c,c}(\boldsymbol{u}) &= \mathrm{Softmax}(\boldsymbol{u}) - \boldsymbol{p}_*, \\
\mathrm{Softmax}(\boldsymbol{u}^*) &= \mathrm{Softmax}(\boldsymbol{\Pi}_C f^*(\boldsymbol{Z}_c, c)) = \mathrm{Softmax}(f^*(\boldsymbol{Z}_c, c)) = \boldsymbol{p}_*, \\
\nabla_{\mathcal{S}_C}\rho_{\boldsymbol{Z}_c,c}(\boldsymbol{u}^*) &= \boldsymbol{\Pi}_C(\mathrm{Softmax}(\boldsymbol{u}^*) - \boldsymbol{p}_*) = \mathbf{0}.
\end{aligned}$$

The first-order term is therefore

$$\langle \nabla_{\mathcal{S}_C}\rho_{\boldsymbol{Z}_c,c}(\boldsymbol{u}^*), \boldsymbol{u} - \boldsymbol{u}^*\rangle = 0.$$

Substituting $\boldsymbol{u} = \boldsymbol{\Pi}_C f(\boldsymbol{Z})$ and $\boldsymbol{u}^* = \boldsymbol{\Pi}_C f^*(\boldsymbol{Z}_c, c)$ gives the pointwise inequality

$$\rho_{\boldsymbol{Z}_c,c}(f(\boldsymbol{Z})) - \rho_{\boldsymbol{Z}_c,c}(f^*(\boldsymbol{Z}_c, c)) \ge \frac{\mu_\ell}{2}\,\|\boldsymbol{\Pi}_C(f(\boldsymbol{Z}) - f^*(\boldsymbol{Z}_c, c))\|_2^2\,.$$

This proves the lower bound after taking expectation over $(\boldsymbol{Z}_c, \boldsymbol{Z}_s, c)$.

For the upper bound, no bounded-logit condition is needed. Assumption 8 gives global L_ℓ-smoothness on $\mathcal{S}_C$, so for arbitrary finite centered logits the standard smoothness inequality at $\boldsymbol{u}^*$ yields

$$\rho_{\boldsymbol{Z}_c,c}(\boldsymbol{u}) \le \rho_{\boldsymbol{Z}_c,c}(\boldsymbol{u}^*) + \langle \nabla_{\mathcal{S}_C}\rho_{\boldsymbol{Z}_c,c}(\boldsymbol{u}^*), \boldsymbol{u} - \boldsymbol{u}^*\rangle + \frac{L_\ell}{2}\|\boldsymbol{u} - \boldsymbol{u}^*\|_2^2.$$

The first-order term again vanishes because $\boldsymbol{u}^*$ is the conditional Bayes optimum on $\mathcal{S}_C$. Hence

$$\rho_{\boldsymbol{Z}_c,c}(f(\boldsymbol{Z})) - \rho_{\boldsymbol{Z}_c,c}(f^*(\boldsymbol{Z}_c,c)) \le \frac{L_\ell}{2} \left\|\boldsymbol{\Pi}_C(f(\boldsymbol{Z}) - f^*(\boldsymbol{Z}_c,c))\right\|_2^2 .$$

Taking expectation over $(\boldsymbol{Z}_c, \boldsymbol{Z}_s, c)$ completes the proof. □

C.12 Proof of Lemma 8

Proof. For $\boldsymbol{W}_c$ and $\boldsymbol{W}_s$, define the centered approximation objective

$$J(\boldsymbol{W}_c, \boldsymbol{W}_s) = \mathbb{E}\left[\left\|\boldsymbol{\Pi}_C\left(\boldsymbol{W}_c\boldsymbol{Z}_c + \boldsymbol{W}_s\boldsymbol{Z}_s - (\boldsymbol{\beta}_{\text{base}} + c\boldsymbol{\beta}_{\text{shift}})\boldsymbol{Z}_c\right)\right\|_2^2\right].$$

Set

$$\boldsymbol{U} = \boldsymbol{\Pi}_C(\boldsymbol{W}_c - \boldsymbol{\beta}_{\text{base}}), \qquad \boldsymbol{V} = \boldsymbol{\Pi}_C\boldsymbol{W}_s, \qquad \boldsymbol{B} = \boldsymbol{\Pi}_C\boldsymbol{\beta}_{\text{shift}}.$$

Using the linearity of $\boldsymbol{\Pi}_C$, the objective becomes

$$J(\boldsymbol{W}_c, \boldsymbol{W}_s) = \mathbb{E}\left[\left\|(\boldsymbol{U} - c\boldsymbol{B})\boldsymbol{Z}_c + \boldsymbol{V}\boldsymbol{Z}_s\right\|_2^2\right].$$

Expanding the squared norm gives

$$\begin{aligned} J(\boldsymbol{W}_c, \boldsymbol{W}_s) = \mathbb{E}\left[\left\|(\boldsymbol{U} - c\boldsymbol{B})\boldsymbol{Z}_c\right\|_2^2\right] + \mathbb{E}\left[\left\|\boldsymbol{V}\boldsymbol{Z}_s\right\|_2^2\right] \\ + 2\,\mathbb{E}\left[\left((\boldsymbol{U} - c\boldsymbol{B})\boldsymbol{Z}_c\right)^\top \boldsymbol{V}\boldsymbol{Z}_s\right]. \end{aligned}$$

The cross term is zero. Indeed, conditioning on $(c, \boldsymbol{Z}_s)$ and using Assumption 9,

$$\begin{aligned} &\mathbb{E}\left[\left((\boldsymbol{U} - c\boldsymbol{B})\boldsymbol{Z}_c\right)^\top \boldsymbol{V}\boldsymbol{Z}_s \mid c, \boldsymbol{Z}_s\right] \\ &= \mathbb{E}\left[\boldsymbol{Z}_c^\top(\boldsymbol{U} - c\boldsymbol{B})^\top \boldsymbol{V}\boldsymbol{Z}_s \mid c, \boldsymbol{Z}_s\right] \\ &= \mathbb{E}[\boldsymbol{Z}_c]^\top(\boldsymbol{U} - c\boldsymbol{B})^\top \boldsymbol{V}\boldsymbol{Z}_s = 0. \end{aligned}$$

The objective decomposes as

$$J(\boldsymbol{W}_c, \boldsymbol{W}_s) = J_1(\boldsymbol{U}) + J_2(\boldsymbol{V}),$$

where

$$J_1(\boldsymbol{U}) = \mathbb{E}_{c,\boldsymbol{Z}_c}\left[\left\|(\boldsymbol{U} - c\boldsymbol{B})\boldsymbol{Z}_c\right\|_2^2\right], \qquad J_2(\boldsymbol{V}) = \mathbb{E}_{\boldsymbol{Z}_s}\left[\left\|\boldsymbol{V}\boldsymbol{Z}_s\right\|_2^2\right].$$

Since $J_2(\boldsymbol{V}) \ge 0$ and $J_2(\boldsymbol{0}) = 0$, the style-additive part cannot reduce the multiplicative approximation error. Hence it remains to minimize $J_1(\boldsymbol{U})$.

For any fixed matrix $\boldsymbol{M} \in \mathbb{R}^{C\times d_c}$, the second moment identity for $\boldsymbol{Z}_c$ gives

$$\begin{aligned} \mathbb{E}_{\boldsymbol{Z}_c}\left[\left\|\boldsymbol{M}\boldsymbol{Z}_c\right\|_2^2\right] &= \mathbb{E}_{\boldsymbol{Z}_c}\left[\operatorname{Tr}\left(\boldsymbol{M}\boldsymbol{Z}_c\boldsymbol{Z}_c^\top\boldsymbol{M}^\top\right)\right] \\ &= \operatorname{Tr}\left(\boldsymbol{M}\,\mathbb{E}[\boldsymbol{Z}_c\boldsymbol{Z}_c^\top]\,\boldsymbol{M}^\top\right) \\ &= \operatorname{Tr}\left(\boldsymbol{M}\boldsymbol{\Sigma}_c\boldsymbol{M}^\top\right) = \left\|\boldsymbol{M}\boldsymbol{\Sigma}_c^{1/2}\right\|_F^2. \end{aligned}$$

Applying this identity with $\boldsymbol{M} = \boldsymbol{U} - c\boldsymbol{B}$ yields

$$J_1(\boldsymbol{U}) = \mathbb{E}_c\left[\left\|(\boldsymbol{U} - c\boldsymbol{B})\boldsymbol{\Sigma}_c^{1/2}\right\|_F^2\right].$$

Let $\widetilde{\boldsymbol{U}} = \boldsymbol{U}\boldsymbol{\Sigma}_c^{1/2}$ and $\widetilde{\boldsymbol{B}} = \boldsymbol{B}\boldsymbol{\Sigma}_c^{1/2}$. Then

$$J_1(\boldsymbol{U}) = \mathbb{E}_c\left[\left\|\widetilde{\boldsymbol{U}} - c\widetilde{\boldsymbol{B}}\right\|_F^2\right].$$

Writing $\bar{c} = \mathbb{E}[c]$ and adding/subtracting $\bar{c}\,\widetilde{\boldsymbol{B}}$ gives the bias-variance decomposition

$$\begin{aligned} \mathbb{E}_c\left[\left\|\widetilde{\boldsymbol{U}} - c\widetilde{\boldsymbol{B}}\right\|_F^2\right] &= \mathbb{E}_c\left[\left\|(\widetilde{\boldsymbol{U}} - \bar{c}\,\widetilde{\boldsymbol{B}}) - (c - \bar{c})\widetilde{\boldsymbol{B}}\right\|_F^2\right] \\ &= \left\|\widetilde{\boldsymbol{U}} - \bar{c}\,\widetilde{\boldsymbol{B}}\right\|_F^2 + \mathbb{E}\left[(c - \bar{c})^2\right]\left\|\widetilde{\boldsymbol{B}}\right\|_F^2 \\ &\quad - 2\,\mathbb{E}[c - \bar{c}]\left\langle\widetilde{\boldsymbol{U}} - \bar{c}\,\widetilde{\boldsymbol{B}}, \widetilde{\boldsymbol{B}}\right\rangle_F \\ &= \left\|\widetilde{\boldsymbol{U}} - \bar{c}\,\widetilde{\boldsymbol{B}}\right\|_F^2 + \operatorname{Var}(c)\left\|\widetilde{\boldsymbol{B}}\right\|_F^2. \end{aligned}$$

The first term is minimized at $\widetilde{\boldsymbol{U}} = \bar{c}\,\widetilde{\boldsymbol{B}}$. Because $\boldsymbol{\Sigma}_c \succ \mathbf{0}$, this choice is attainable by $\boldsymbol{U} = \bar{c}\,\boldsymbol{B}$, equivalently by any $\boldsymbol{W}_c$ satisfying

$$\boldsymbol{\Pi}_C \boldsymbol{W}_c = \boldsymbol{\Pi}_C \boldsymbol{\beta}_{\text{base}} + \bar{c}\,\boldsymbol{\Pi}_C \boldsymbol{\beta}_{\text{shift}}.$$

Combining the minimized value of J_1 with the minimized value of J_2 gives

$$\min_{\boldsymbol{W}_c, \boldsymbol{W}_s} J(\boldsymbol{W}_c, \boldsymbol{W}_s) = \text{Var}(c) \left\| \boldsymbol{\Pi}_C \boldsymbol{\beta}_{\text{shift}} \boldsymbol{\Sigma}_c^{1/2} \right\|_F^2.$$

By Assumption 9, R_W is chosen large enough that one minimizer described above belongs to $\mathcal{H}_{\text{pool}}(R_W)$. Therefore, restricting the infimum to $\mathcal{H}_{\text{pool}}(R_W)$ does not change the optimal value. □

C.13 Proof of Corollary 1

Proof. For any $f_{\text{pool}} \in \mathcal{H}_{\text{pool}}(R_W)$, the lower-bound half of Lemma 7 implies

$$\mathbb{E}_{c\sim\pi_c}\left[\mathcal{R}_c(f_{\text{pool}}) - \mathcal{R}_c(f^*)\right] \geq \frac{\mu_\ell}{2}\,\mathbb{E}\left[\left\|\boldsymbol{\Pi}_C\left(f_{\text{pool}}(\boldsymbol{Z}_c, \boldsymbol{Z}_s) - f^*(\boldsymbol{Z}_c, c)\right)\right\|_2^2\right].$$

Taking the infimum over $f_{\text{pool}} \in \mathcal{H}_{\text{pool}}(R_W)$ on both sides and substituting the exact additive approximation error from Lemma 8 yields

$$\inf_{f_{\text{pool}}\in\mathcal{H}_{\text{pool}}(R_W)} \mathbb{E}_{c\sim\pi_c}\left[\mathcal{R}_c(f_{\text{pool}}) - \mathcal{R}_c(f^*)\right] \geq \frac{\mu_\ell}{2}\,\text{Var}(c)\left\|\boldsymbol{\Pi}_C \boldsymbol{\beta}_{\text{shift}} \boldsymbol{\Sigma}_c^{1/2}\right\|_F^2.$$

The right-hand side is positive whenever $\text{Var}(c) > 0$ and $\|\boldsymbol{\Pi}_C \boldsymbol{\beta}_{\text{shift}} \boldsymbol{\Sigma}_c^{1/2}\|_F > 0$. The first condition states that the domain population contains distinct domains, and the second states that the domain shift changes the identifiable centered logits. Under these non-degeneracy conditions, the pooled-joint additive class has a positive structural misspecification error. □

C.14 Proof of Corollary 2

Proof. Fix the target domain coordinate c_0. Under Assumption 9, the target-domain Bayes-optimal parameter is

$$\boldsymbol{\beta}_*^{(0)} = \boldsymbol{\beta}_{\text{base}} + c_0 \boldsymbol{\beta}_{\text{shift}}.$$

By Assumption 9, write $\boldsymbol{\beta}_{\text{L}} = \boldsymbol{\beta}_{\text{base}} + \widehat{c}_{\text{L}} \boldsymbol{\beta}_{\text{shift}}$ and $\boldsymbol{\beta}_{\text{inv}} = \boldsymbol{\beta}_{\text{base}} + c_{\text{inv}} \boldsymbol{\beta}_{\text{shift}}$. Their identifiable centered differences are

$$\boldsymbol{\Pi}_C(\boldsymbol{\beta}_{\text{L}} - \boldsymbol{\beta}_*^{(0)}) = (\widehat{c}_{\text{L}} - c_0)\boldsymbol{\Pi}_C \boldsymbol{\beta}_{\text{shift}},$$
$$\boldsymbol{\Pi}_C(\boldsymbol{\beta}_{\text{inv}} - \boldsymbol{\beta}_*^{(0)}) = (c_{\text{inv}} - c_0)\boldsymbol{\Pi}_C \boldsymbol{\beta}_{\text{shift}}.$$

Let

$$\boldsymbol{B} = \boldsymbol{\Pi}_C \boldsymbol{\beta}_{\text{shift}} \boldsymbol{\Sigma}_c^{1/2}.$$

Using the second-moment identity from the proof of Lemma 8, for any scalar a,

$$\mathbb{E}_{\boldsymbol{Z}_c}\left[\|a\,\boldsymbol{\Pi}_C \boldsymbol{\beta}_{\text{shift}} \boldsymbol{Z}_c\|_2^2\right] = a^2 \|\boldsymbol{B}\|_F^2.$$

Applying the global upper-smoothness half of Lemma 7 to f_{L} on the fixed target domain c_0 gives

$$\mathcal{R}_{c_0}(f_{\text{L}}) - \mathcal{R}_{c_0}(f^*) \leq \frac{L_\ell}{2}(c_0 - \widehat{c}_{\text{L}})^2 \|\boldsymbol{B}\|_F^2.$$

Applying the lower half of the same lemma to f_{inv} gives

$$\mathcal{R}_{c_0}(f_{\text{inv}}) - \mathcal{R}_{c_0}(f^*) \geq \frac{\mu_\ell}{2}(c_0 - c_{\text{inv}})^2 \|\boldsymbol{B}\|_F^2.$$

By the non-degeneracy assumption in the corollary, $\|\boldsymbol{B}\|_F^2 > 0$. Since $c_{\text{inv}} \neq c_0$ by Assumption 9, the denominator is strictly positive. Dividing the two inequalities yields

$$\frac{\mathcal{R}_{c_0}(f_{\text{L}}) - \mathcal{R}_{c_0}(f^*)}{\mathcal{R}_{c_0}(f_{\text{inv}}) - \mathcal{R}_{c_0}(f^*)} \leq \frac{L_\ell}{\mu_\ell}\,\frac{(c_0 - \widehat{c}_{\text{L}})^2}{(c_0 - c_{\text{inv}})^2}.$$

The strict improvement condition follows immediately: if

$$|c_0 - \widehat{c}_{\text{L}}| < \sqrt{\frac{\mu_\ell}{L_\ell}}\,|c_0 - c_{\text{inv}}|,$$

then the right-hand side is strictly smaller than 1, and hence

$$\mathcal{R}_{c_0}(f_{\text{L}}) - \mathcal{R}_{c_0}(f^*) < \mathcal{R}_{c_0}(f_{\text{inv}}) - \mathcal{R}_{c_0}(f^*).$$

□

C.15 Proof of Corollary 3

Proof. The source estimators and empirical style measures are

$$\widehat{\boldsymbol{\beta}}^{(e)} \in \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \widehat{\mathcal{R}}_e(\boldsymbol{\beta}), \quad \widetilde{\boldsymbol{\beta}}^{(e)} \in \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \widetilde{\mathcal{R}}_e^{\mathrm{emp}}(\boldsymbol{\beta}), \quad e = 1, \dots, E,$$

$$\widehat{\nu}_e = \frac{1}{n_e}\sum_{i=1}^{n_e} \delta_{\boldsymbol{z}_{s,i}^{(e)}}, \qquad \widetilde{\nu}_e = \frac{1}{n_e}\sum_{i=1}^{n_e} \delta_{\widetilde{\boldsymbol{z}}_{s,i}^{(e)}}, \qquad e = 0, \dots, E.$$

The weights satisfy

$$\widetilde{w}_e \geq 0, \qquad \sum_{e\in\widetilde{\mathcal{N}}_K(0)} \widetilde{w}_e = 1, \qquad \widetilde{\boldsymbol{\beta}}^{(0)} = \sum_{e\in\widetilde{\mathcal{N}}_K(0)} \widetilde{w}_e \widetilde{\boldsymbol{\beta}}^{(e)} \in \mathcal{B}.$$

Constants denoted by C_E may depend on the fixed number of sources E.

Empirical-measure perturbations. Pairing each learned style point with its oracle counterpart gives

$$\begin{aligned}\mathcal{W}_1(\widetilde{\nu}_e, \widehat{\nu}_e) &\leq \frac{1}{n_e}\sum_{i=1}^{n_e} \|\widetilde{\boldsymbol{z}}_{s,i}^{(e)} - \boldsymbol{z}_{s,i}^{(e)}\|_2 \\ &= \frac{1}{n_e}\sum_{i=1}^{n_e} \|\boldsymbol{\xi}_{s,i}^{(e)}\|_2, \qquad e = 0, \dots, E.\end{aligned}$$

For $q \in \{c, s\}$, Jensen's inequality and Assumption 10 give

$$\begin{aligned}\mathbb{E}\left(\frac{1}{n_e}\sum_{i=1}^{n_e} \|\boldsymbol{\xi}_{q,i}^{(e)}\|_2\right)^2 &\leq \frac{1}{n_e}\sum_{i=1}^{n_e} \mathbb{E}\|\boldsymbol{\xi}_{q,i}^{(e)}\|_2^2 \\ &\leq \frac{1}{n_e}\sum_{i=1}^{n_e} \epsilon_{\mathrm{rep}}^2 = \epsilon_{\mathrm{rep}}^2,\end{aligned}$$

$$\mathbb{E}\left(\frac{1}{n_e}\sum_{i=1}^{n_e} \|\boldsymbol{\xi}_{q,i}^{(e)}\|_2\right) \leq \left[\mathbb{E}\left(\frac{1}{n_e}\sum_{i=1}^{n_e} \|\boldsymbol{\xi}_{q,i}^{(e)}\|_2\right)^2\right]^{1/2} \leq \epsilon_{\mathrm{rep}}.$$

For the source maximum,

$$\mathbb{E}\left\{\left[\max_{1\leq e\leq E} \frac{1}{n_e}\sum_i \|\boldsymbol{\xi}_{q,i}^{(e)}\|_2\right]^2\right\} \leq \sum_{e=1}^{E} \mathbb{E}\left(\frac{1}{n_e}\sum_i \|\boldsymbol{\xi}_{q,i}^{(e)}\|_2\right)^2 \leq E\epsilon_{\mathrm{rep}}^2,$$

$$\mathbb{E}\max_{1\leq e\leq E} \frac{1}{n_e}\sum_i \|\boldsymbol{\xi}_{q,i}^{(e)}\|_2 \leq \left\{\mathbb{E}\left\{\left[\max_{1\leq e\leq E} \frac{1}{n_e}\sum_i \|\boldsymbol{\xi}_{q,i}^{(e)}\|_2\right]^2\right\}\right\}^{1/2} \leq \sqrt{E}\,\epsilon_{\mathrm{rep}}.$$

In particular, the target and source style perturbations satisfy

$$\mathbb{E}\mathcal{W}_1(\widetilde{\nu}_0, \widehat{\nu}_0) \leq \epsilon_{\mathrm{rep}}, \qquad \mathbb{E}\mathcal{W}_1(\widetilde{\nu}_0, \widehat{\nu}_0)^2 \leq \epsilon_{\mathrm{rep}}^2,$$

$$\mathbb{E}\max_{1\leq e\leq E} \mathcal{W}_1(\widetilde{\nu}_e, \widehat{\nu}_e) \leq \sqrt{E}\,\epsilon_{\mathrm{rep}}, \quad \mathbb{E}\max_{1\leq e\leq E} \mathcal{W}_1(\widetilde{\nu}_e, \widehat{\nu}_e)^2 \leq E\epsilon_{\mathrm{rep}}^2.$$

Source-classifier perturbations. Write $\boldsymbol{e}_y$ for the one-hot label vector and set

$$\boldsymbol{p} = \mathrm{Softmax}(\boldsymbol{\beta}\boldsymbol{z}), \qquad \boldsymbol{p}' = \mathrm{Softmax}(\boldsymbol{\beta}(\boldsymbol{z} + \boldsymbol{\xi})).$$

For any probability vector $\boldsymbol{p}$, the symmetric Softmax Jacobian obeys

$$\begin{aligned}\boldsymbol{H}(\boldsymbol{p}) &= \mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top, \\ \|\boldsymbol{H}(\boldsymbol{p})\|_{\mathrm{op}} &\leq \max_j \sum_k |H_{jk}(\boldsymbol{p})| = \max_j 2p_j(1-p_j) \leq \tfrac{1}{2}.\end{aligned}$$

The fundamental theorem of calculus gives

$$\boldsymbol{p}' - \boldsymbol{p} = \int_0^1 \boldsymbol{H}(\mathrm{Softmax}(\boldsymbol{\beta}(\boldsymbol{z} + t\boldsymbol{\xi}))\,\boldsymbol{\beta}\boldsymbol{\xi}\,dt,$$

$$\|\boldsymbol{p}' - \boldsymbol{p}\|_2 \le \int_0^1 \tfrac{1}{2}\|\boldsymbol{\beta}\boldsymbol{\xi}\|_2\,dt \le \tfrac{1}{2}R_\beta\|\boldsymbol{\xi}\|_2,$$

$$\|\boldsymbol{p}' - \boldsymbol{e}_y\|_2^2 = \|\boldsymbol{p}'\|_2^2 + 1 - 2p'_y \le 2.$$

The single-sample gradient difference is

$$\begin{aligned}
&\nabla_{\boldsymbol{\beta}}\ell_{\mathrm{CE}}(\boldsymbol{\beta}(\boldsymbol{z}+\boldsymbol{\xi}), y) - \nabla_{\boldsymbol{\beta}}\ell_{\mathrm{CE}}(\boldsymbol{\beta}\boldsymbol{z}, y) \\
&\quad = (\boldsymbol{p}' - \boldsymbol{e}_y)(\boldsymbol{z}+\boldsymbol{\xi})^\top - (\boldsymbol{p} - \boldsymbol{e}_y)\boldsymbol{z}^\top \\
&\quad = (\boldsymbol{p}' - \boldsymbol{e}_y)\boldsymbol{\xi}^\top + (\boldsymbol{p}' - \boldsymbol{p})\boldsymbol{z}^\top.
\end{aligned}$$

For $\boldsymbol{\beta} \in \mathcal{B}$ and $\|\boldsymbol{z}\|_2 \le R_c$,

$$\begin{aligned}
&\|\nabla_{\boldsymbol{\beta}}\ell_{\mathrm{CE}}(\boldsymbol{\beta}(\boldsymbol{z}+\boldsymbol{\xi}), y) - \nabla_{\boldsymbol{\beta}}\ell_{\mathrm{CE}}(\boldsymbol{\beta}\boldsymbol{z}, y)\|_F \\
&\quad \le \|\boldsymbol{p}' - \boldsymbol{e}_y\|_2\|\boldsymbol{\xi}\|_2 + \|\boldsymbol{p}' - \boldsymbol{p}\|_2\|\boldsymbol{z}\|_2 \\
&\quad \le \sqrt{2}\|\boldsymbol{\xi}\|_2 + \tfrac{1}{2}R_\beta R_c\|\boldsymbol{\xi}\|_2 = L_{\nabla z}\|\boldsymbol{\xi}\|_2.
\end{aligned}$$

Averaging over source e yields

$$\begin{aligned}
&\|\nabla\widetilde{\mathcal{R}}_e^{\mathrm{emp}}(\boldsymbol{\beta}) - \nabla\widehat{\mathcal{R}}_e(\boldsymbol{\beta})\|_F \\
&\quad \le \frac{1}{n_e}\sum_{i=1}^{n_e}\|\nabla_{\boldsymbol{\beta}}\ell_{\mathrm{CE}}(\boldsymbol{\beta}\widetilde{\boldsymbol{z}}_{c,i}^{(e)}, y_i^{(e)}) - \nabla_{\boldsymbol{\beta}}\ell_{\mathrm{CE}}(\boldsymbol{\beta}\boldsymbol{z}_{c,i}^{(e)}, y_i^{(e)})\|_F \\
&\quad \le \frac{L_{\nabla z}}{n_e}\sum_{i=1}^{n_e}\|\boldsymbol{\xi}_{c,i}^{(e)}\|_2.
\end{aligned}$$

Set $\Delta_e = \widetilde{\boldsymbol{\beta}}^{(e)} - \widehat{\boldsymbol{\beta}}^{(e)}$. The constrained first-order conditions are

$$\left\langle \nabla\widehat{\mathcal{R}}_e(\widehat{\boldsymbol{\beta}}^{(e)}), \Delta_e \right\rangle_F \ge 0, \qquad \left\langle \nabla\widetilde{\mathcal{R}}_e^{\mathrm{emp}}(\widetilde{\boldsymbol{\beta}}^{(e)}), \Delta_e \right\rangle_F \le 0.$$

Strong monotonicity on $\mathcal{B}$ and these conditions imply

$$\begin{aligned}
\mu_{\mathrm{rep}}\|\Delta_e\|_F^2 &\le \left\langle \nabla\widetilde{\mathcal{R}}_e^{\mathrm{emp}}(\widetilde{\boldsymbol{\beta}}^{(e)}) - \nabla\widetilde{\mathcal{R}}_e^{\mathrm{emp}}(\widehat{\boldsymbol{\beta}}^{(e)}), \Delta_e \right\rangle_F \\
&= \left\langle \nabla\widetilde{\mathcal{R}}_e^{\mathrm{emp}}(\widetilde{\boldsymbol{\beta}}^{(e)}), \Delta_e \right\rangle_F - \left\langle \nabla\widehat{\mathcal{R}}_e(\widehat{\boldsymbol{\beta}}^{(e)}), \Delta_e \right\rangle_F \\
&\quad + \left\langle \nabla\widehat{\mathcal{R}}_e(\widehat{\boldsymbol{\beta}}^{(e)}) - \nabla\widetilde{\mathcal{R}}_e^{\mathrm{emp}}(\widehat{\boldsymbol{\beta}}^{(e)}), \Delta_e \right\rangle_F \\
&\le \left\langle \nabla\widehat{\mathcal{R}}_e(\widehat{\boldsymbol{\beta}}^{(e)}) - \nabla\widetilde{\mathcal{R}}_e^{\mathrm{emp}}(\widehat{\boldsymbol{\beta}}^{(e)}), \Delta_e \right\rangle_F \\
&\le \|\nabla\widehat{\mathcal{R}}_e(\widehat{\boldsymbol{\beta}}^{(e)}) - \nabla\widetilde{\mathcal{R}}_e^{\mathrm{emp}}(\widehat{\boldsymbol{\beta}}^{(e)})\|_F\,\|\Delta_e\|_F \\
&\le \frac{L_{\nabla z}}{n_e}\sum_i \|\boldsymbol{\xi}_{c,i}^{(e)}\|_2\,\|\Delta_e\|_F.
\end{aligned}$$

For $\Delta_e \neq 0$, division by $\mu_{\mathrm{rep}}\|\Delta_e\|_F$ gives the following bound; it also holds for $\Delta_e = 0$:

$$\|\widetilde{\boldsymbol{\beta}}^{(e)} - \widehat{\boldsymbol{\beta}}^{(e)}\|_F \le \frac{L_{\nabla z}}{\mu_{\mathrm{rep}}}\frac{1}{n_e}\sum_i \|\boldsymbol{\xi}_{c,i}^{(e)}\|_2.$$

The source-maximum moments are

$$\begin{aligned}
\mathbb{E}\max_e \|\widetilde{\boldsymbol{\beta}}^{(e)} - \widehat{\boldsymbol{\beta}}^{(e)}\|_F &\le \frac{L_{\nabla z}}{\mu_{\mathrm{rep}}}\mathbb{E}\max_e \frac{1}{n_e}\sum_i \|\boldsymbol{\xi}_{c,i}^{(e)}\|_2 \le \frac{\sqrt{E}L_{\nabla z}}{\mu_{\mathrm{rep}}}\epsilon_{\mathrm{rep}}, \\
\mathbb{E}\max_e \|\widetilde{\boldsymbol{\beta}}^{(e)} - \widehat{\boldsymbol{\beta}}^{(e)}\|_F^2 &\le \frac{L_{\nabla z}^2}{\mu_{\mathrm{rep}}^2}\mathbb{E}\left\{\left[\max_e \frac{1}{n_e}\sum_i \|\boldsymbol{\xi}_{c,i}^{(e)}\|_2\right]^2\right\} \\
&\le \frac{EL_{\nabla z}^2}{\mu_{\mathrm{rep}}^2}\epsilon_{\mathrm{rep}}^2.
\end{aligned}$$

Perturbed source selection. Let $\mathcal{N}_K^*(0)$ contain the K nearest source laws to ν_0 and set

$$R_K^* = \max_{e \in \mathcal{N}_K^*(0)} \mathcal{W}_1(\nu_e, \nu_0).$$

The empirical K-nearest-neighbor property and the triangle inequality give

$$\begin{aligned}\max_{e \in \widetilde{\mathcal{N}}_K(0)} \mathcal{W}_1(\widetilde{\nu}_e, \widetilde{\nu}_0) &\le \max_{j \in \mathcal{N}_K^*(0)} \mathcal{W}_1(\widetilde{\nu}_j, \widetilde{\nu}_0) \\ &\le \max_{j \in \mathcal{N}_K^*(0)} \{\mathcal{W}_1(\widetilde{\nu}_j, \nu_j) + \mathcal{W}_1(\nu_j, \nu_0) + \mathcal{W}_1(\nu_0, \widetilde{\nu}_0)\} \\ &\le R_K^* + \max_{1 \le e \le E} \mathcal{W}_1(\widetilde{\nu}_e, \nu_e) + \mathcal{W}_1(\widetilde{\nu}_0, \nu_0).\end{aligned}$$

For each selected source,

$$\begin{aligned}\mathcal{W}_1(\nu_e, \nu_0) &\le \mathcal{W}_1(\nu_e, \widetilde{\nu}_e) + \mathcal{W}_1(\widetilde{\nu}_e, \widetilde{\nu}_0) + \mathcal{W}_1(\widetilde{\nu}_0, \nu_0) \\ &\le R_K^* + 2 \max_{1 \le j \le E} \mathcal{W}_1(\widetilde{\nu}_j, \nu_j) + 2\mathcal{W}_1(\widetilde{\nu}_0, \nu_0).\end{aligned}$$

Splitting through the oracle empirical measures yields

$$\begin{aligned}\max_{1 \le e \le E} \mathcal{W}_1(\widetilde{\nu}_e, \nu_e) &\le \max_{1 \le e \le E} \mathcal{W}_1(\widetilde{\nu}_e, \widehat{\nu}_e) + \max_{1 \le e \le E} \mathcal{W}_1(\widehat{\nu}_e, \nu_e), \\ \mathcal{W}_1(\widetilde{\nu}_0, \nu_0) &\le \mathcal{W}_1(\widetilde{\nu}_0, \widehat{\nu}_0) + \mathcal{W}_1(\widehat{\nu}_0, \nu_0).\end{aligned}$$

For the target parameter,

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)} &= \sum_{e \in \widetilde{\mathcal{N}}_K(0)} \widetilde{w}_e (\widetilde{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(0)}) \\ &= \sum_{e \in \widetilde{\mathcal{N}}_K(0)} \widetilde{w}_e \big[(\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}) + (\widetilde{\boldsymbol{\beta}}^{(e)} - \widehat{\boldsymbol{\beta}}^{(e)}) + (\boldsymbol{\beta}_*^{(e)} - \boldsymbol{\beta}_*^{(0)}) \big].\end{aligned}$$

Convex weights and the Lipschitz rule map give

$$\begin{aligned}\|\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\|_F &\le \sum_{e \in \widetilde{\mathcal{N}}_K(0)} \widetilde{w}_e \big[\|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F + \|\widetilde{\boldsymbol{\beta}}^{(e)} - \widehat{\boldsymbol{\beta}}^{(e)}\|_F \\ &\qquad + L_\beta \mathcal{W}_1(\nu_e, \nu_0) \big] \\ &\le \max_e \|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F + \max_e \|\widetilde{\boldsymbol{\beta}}^{(e)} - \widehat{\boldsymbol{\beta}}^{(e)}\|_F \\ &\quad + L_\beta \big[R_K^* + 2\mathcal{W}_1(\widehat{\nu}_0, \nu_0) + 2 \max_e \mathcal{W}_1(\widehat{\nu}_e, \nu_e) \\ &\qquad + 2\mathcal{W}_1(\widetilde{\nu}_0, \widehat{\nu}_0) + 2 \max_e \mathcal{W}_1(\widetilde{\nu}_e, \widehat{\nu}_e) \big].\end{aligned} \tag{*}$$

Here and below, source maxima range over $e = 1, \dots, E$.

Parameter moments. Lemmas 3 and 4 imply

$$\begin{aligned}\mathbb{E} \max_e \|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F &\le C_\star \frac{R_c}{\mu} \sqrt{\frac{d_\beta + \log E}{n_{\min}}}, \\ \mathbb{E} \max_e \|\widehat{\boldsymbol{\beta}}^{(e)} - \boldsymbol{\beta}_*^{(e)}\|_F^2 &\le C_\star \frac{R_c^2 (d_\beta + \log E)}{\mu^2 n_{\min}}, \\ \mathbb{E}\mathcal{W}_1(\widehat{\nu}_0, \nu_0) &\le C_\star \gamma(n_0, d_s), \\ \mathbb{E}\mathcal{W}_1(\widehat{\nu}_0, \nu_0)^2 &\le C_\star \gamma(n_0, d_s)^2, \\ \mathbb{E} \max_e \mathcal{W}_1(\widehat{\nu}_e, \nu_e) &\le C_\star \{\gamma(n_{\min}, d_s) + \Gamma(n_{\min}, E, d_s)\}, \\ \mathbb{E} \max_e \mathcal{W}_1(\widehat{\nu}_e, \nu_e)^2 &\le C_\star \{\gamma(n_{\min}, d_s)^2 + \Gamma(n_{\min}, E, d_s)^2\}.\end{aligned}$$

The nearest-neighbor calculation in Lemma 6 and the proof of Theorem 3 gives

$$\mathbb{E}R_K^* \le C_\star (K/E)^{1/d_{\mathcal{M}}}, \qquad \mathbb{E}(R_K^*)^2 \le C_\star (K/E)^{2/d_{\mathcal{M}}}.$$

Taking expectations in $(*)$ and using the displayed moment bounds gives

$$\begin{aligned}\mathbb{E}\|\widetilde{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}_*^{(0)}\|_F &\le C_\star\frac{R_c}{\mu}\sqrt{\frac{d_\beta+\log E}{n_{\min}}}+\frac{\sqrt{E}L_{\nabla z}}{\mu_{\mathrm{rep}}}\epsilon_{\mathrm{rep}}\\ &\quad+C_\star L_\beta\big[(K/E)^{1/d_{\mathcal{M}}}+2\gamma(n_0,d_s)\\ &\qquad+2\gamma(n_{\min},d_s)+2\Gamma(n_{\min},E,d_s)\big]\\ &\quad+2L_\beta(1+\sqrt{E})\epsilon_{\mathrm{rep}}\\ &\le\mathcal{O}\Bigg(\frac{R_c}{\mu}\sqrt{\frac{d_\beta+\log E}{n_{\min}}}+\gamma(n_0,d_s)+\Gamma(n_{\min},E,d_s)\\ &\qquad+(K/E)^{1/d_{\mathcal{M}}}+\left(\frac{L_{\nabla z}}{\mu_{\mathrm{rep}}}+L_\beta\right)\epsilon_{\mathrm{rep}}\Bigg).\end{aligned}$$

The last inequality uses $\gamma(n_{\min},d_s)\le\Gamma(n_{\min},E,d_s)$ and fixed E. Applying $(\sum_{j=1}^7 a_j)^2\le 7\sum_{j=1}^7 a_j^2$ to the seven nonnegative terms in $(*)$ gives

$$\begin{aligned}\mathbb{E}\|\widetilde{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}_*^{(0)}\|_F^2 &\le 7\mathbb{E}\max_e\|\widehat{\boldsymbol{\beta}}^{(e)}-\boldsymbol{\beta}_*^{(e)}\|_F^2+7\mathbb{E}\max_e\|\widetilde{\boldsymbol{\beta}}^{(e)}-\widehat{\boldsymbol{\beta}}^{(e)}\|_F^2\\ &\quad+7L_\beta^2\mathbb{E}(R_K^*)^2\\ &\quad+28L_\beta^2\big[\mathbb{E}\mathcal{W}_1(\widehat{\nu}_0,\nu_0)^2+\mathbb{E}\max_e\mathcal{W}_1(\widehat{\nu}_e,\nu_e)^2\big]\\ &\quad+28L_\beta^2\big[\mathbb{E}\mathcal{W}_1(\widetilde{\nu}_0,\widehat{\nu}_0)^2+\mathbb{E}\max_e\mathcal{W}_1(\widetilde{\nu}_e,\widehat{\nu}_e)^2\big]\\ &\le 7C_\star\frac{R_c^2(d_\beta+\log E)}{\mu^2 n_{\min}}+\frac{7EL_{\nabla z}^2}{\mu_{\mathrm{rep}}^2}\epsilon_{\mathrm{rep}}^2\\ &\quad+7C_\star L_\beta^2(K/E)^{2/d_{\mathcal{M}}}\\ &\quad+28C_\star L_\beta^2\big[\gamma(n_0,d_s)^2+\gamma(n_{\min},d_s)^2+\Gamma(n_{\min},E,d_s)^2\big]\\ &\quad+28L_\beta^2(1+E)\epsilon_{\mathrm{rep}}^2\\ &\le C_E\Bigg(\left[\frac{R_c}{\mu}\sqrt{\frac{d_\beta+\log E}{n_{\min}}}+\gamma(n_0,d_s)+\Gamma(n_{\min},E,d_s)+(K/E)^{1/d_{\mathcal{M}}}\right]^2\\ &\qquad+\left(\frac{L_{\nabla z}}{\mu_{\mathrm{rep}}}+L_\beta\right)^2\epsilon_{\mathrm{rep}}^2\Bigg).\end{aligned}\tag{**}$$

Target excess risk. Let $\mathcal{F}$ be generated by the fitted encoders, source classifiers, and target-summary sample. For a fresh target observation, define

$$\begin{aligned}\boldsymbol{u}_* &= \boldsymbol{\beta}_*^{(0)}\boldsymbol{Z}_c^{(0)}, & \widetilde{\boldsymbol{u}} &= \widetilde{\boldsymbol{\beta}}^{(0)}(\boldsymbol{Z}_c^{(0)}+\boldsymbol{\xi}_c^{(0)}),\\ \boldsymbol{p}_* &= \mathrm{Softmax}(\boldsymbol{u}_*), & \Delta_{\boldsymbol{u}} &= \widetilde{\boldsymbol{u}}-\boldsymbol{u}_*.\end{aligned}$$

Assumption 10 gives

$$\mathbb{E}[\boldsymbol{e}_Y\mid\mathcal{F},\boldsymbol{Z}_c^{(0)},\boldsymbol{Z}_s^{(0)},\boldsymbol{\xi}_c^{(0)},\boldsymbol{\xi}_s^{(0)}]=\boldsymbol{p}_*.$$

The conditional cross-entropy as a function of logits is

$$\psi_{\boldsymbol{p}_*}(\boldsymbol{u})=\log\left(\sum_{j=1}^C e^{u_j}\right)-\boldsymbol{p}_*^\top\boldsymbol{u}.$$

Its derivatives satisfy

$$\begin{aligned}&\nabla\psi_{\boldsymbol{p}_*}(\boldsymbol{u}_*)=\mathrm{Softmax}(\boldsymbol{u}_*)-\boldsymbol{p}_*=\boldsymbol{0},\\ &\nabla^2\psi_{\boldsymbol{p}_*}(\boldsymbol{u})=\boldsymbol{H}(\mathrm{Softmax}(\boldsymbol{u})),\qquad \boldsymbol{0}\preceq\nabla^2\psi_{\boldsymbol{p}_*}(\boldsymbol{u})\preceq\tfrac{1}{2}\boldsymbol{I}_C.\end{aligned}$$

Taylor's formula with integral remainder gives

$$\begin{aligned}0 \le \psi_{\boldsymbol{p}_*}(\widetilde{\boldsymbol{u}}) - \psi_{\boldsymbol{p}_*}(\boldsymbol{u}_*) &= \nabla\psi_{\boldsymbol{p}_*}(\boldsymbol{u}_*)^\top \Delta_{\boldsymbol{u}} \\ &\quad + \int_0^1 (1-t)\, \Delta_{\boldsymbol{u}}^\top \nabla^2 \psi_{\boldsymbol{p}_*}(\boldsymbol{u}_* + t\Delta_{\boldsymbol{u}})\Delta_{\boldsymbol{u}}\, dt \\ &\le \tfrac{1}{2}\|\Delta_{\boldsymbol{u}}\|_2^2 \int_0^1 (1-t)\, dt = \tfrac{1}{4}\|\Delta_{\boldsymbol{u}}\|_2^2.\end{aligned}$$

The logit difference obeys

$$\begin{aligned}\Delta_{\boldsymbol{u}} &= (\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)})\boldsymbol{Z}_c^{(0)} + \widetilde{\boldsymbol{\beta}}^{(0)}\boldsymbol{\xi}_c^{(0)}, \\ \|\Delta_{\boldsymbol{u}}\|_2^2 &\le 2\|(\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)})\boldsymbol{Z}_c^{(0)}\|_2^2 + 2\|\widetilde{\boldsymbol{\beta}}^{(0)}\boldsymbol{\xi}_c^{(0)}\|_2^2 \\ &\le 2R_c^2\|\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\|_F^2 + 2R_\beta^2\|\boldsymbol{\xi}_c^{(0)}\|_2^2.\end{aligned}$$

The risks are evaluated with fitted objects fixed:

$$\begin{aligned}\widetilde{\mathcal{R}}_0(\widetilde{\boldsymbol{\beta}}^{(0)}) &= \mathbb{E}[\ell_{\mathrm{CE}}(\widetilde{\boldsymbol{u}}, Y^{(0)}) \mid \mathcal{F}], \\ \mathcal{R}_0(\boldsymbol{\beta}_*^{(0)}) &= \mathbb{E}[\ell_{\mathrm{CE}}(\boldsymbol{u}_*, Y^{(0)}) \mid \mathcal{F}].\end{aligned}$$

Iterated expectation and the fresh-observation second-moment bound yield

$$\begin{aligned}0 &\le \mathbb{E}\{\widetilde{\mathcal{R}}_0(\widetilde{\boldsymbol{\beta}}^{(0)}) - \mathcal{R}_0(\boldsymbol{\beta}_*^{(0)})\} \\ &= \mathbb{E}\{\psi_{\boldsymbol{p}_*}(\widetilde{\boldsymbol{u}}) - \psi_{\boldsymbol{p}_*}(\boldsymbol{u}_*)\} \\ &\le \tfrac{1}{4}\mathbb{E}\|\Delta_{\boldsymbol{u}}\|_2^2 \\ &\le \frac{R_c^2}{2}\mathbb{E}\|\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\|_F^2 + \frac{R_\beta^2}{2}\mathbb{E}\|\boldsymbol{\xi}_c^{(0)}\|_2^2 \\ &\le \frac{R_c^2}{2}\mathbb{E}\|\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_*^{(0)}\|_F^2 + \frac{R_\beta^2}{2}\epsilon_{\mathrm{rep}}^2.\end{aligned}$$

Substitution of $(**)$ gives the full stated rate:

$$\begin{aligned}&\mathbb{E}\{\widetilde{\mathcal{R}}_0(\widetilde{\boldsymbol{\beta}}^{(0)}) - \mathcal{R}_0(\boldsymbol{\beta}_*^{(0)})\} \\ &\quad\le \frac{C_E R_c^2}{2}\Bigg(\left[\frac{R_c}{\mu}\sqrt{\frac{d_\beta + \log E}{n_{\min}}} + \gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s) + (K/E)^{1/d_{\mathcal{M}}}\right]^2 \\ &\qquad\qquad + \left(\frac{L_{\nabla z}}{\mu_{\mathrm{rep}}} + L_\beta\right)^2 \epsilon_{\mathrm{rep}}^2\Bigg) + \frac{R_\beta^2}{2}\epsilon_{\mathrm{rep}}^2 \\ &\quad= \frac{C_E R_c^2}{2}\left[\frac{R_c}{\mu}\sqrt{\frac{d_\beta + \log E}{n_{\min}}} + \gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s) + (K/E)^{1/d_{\mathcal{M}}}\right]^2 \\ &\qquad + \left[\frac{C_E R_c^2}{2}\left(\frac{L_{\nabla z}}{\mu_{\mathrm{rep}}} + L_\beta\right)^2 + \frac{R_\beta^2}{2}\right]\epsilon_{\mathrm{rep}}^2 \\ &\quad\le \mathcal{O}\Bigg(R_c^2\left[\frac{R_c}{\mu}\sqrt{\frac{d_\beta + \log E}{n_{\min}}} + \gamma(n_0, d_s) + \Gamma(n_{\min}, E, d_s) + (K/E)^{1/d_{\mathcal{M}}}\right]^2 \\ &\qquad + \left[R_c^2\left(\frac{L_{\nabla z}}{\mu_{\mathrm{rep}}} + L_\beta\right)^2 + R_\beta^2\right]\epsilon_{\mathrm{rep}}^2\Bigg).\end{aligned}$$

□