Beyond Structural Symmetries: Linear Mode Connectivity via Neuron Identifiability

Vincent Bürgin [*1] **Daniel Herbst** [*1] **Ya-Wei Eileen Lin** [1] **Stefanie Jegelka** [1 2]

Abstract

Many striking phenomena in deep learning, such as linear mode connectivity and the structured behavior of training dynamics, are closely tied to parameter symmetries: transformations that leave the realized function unchanged. Despite growing attention to parameter symmetries, the exact interplay between parameters, data, and representations remains underexplored. To investigate this, we develop a theoretical framework of effective function classes, i.e., the set of functions a neuron can realize on its input support, and the norm cost of realizing them. We then formalize *effective symmetry breaking* via neuron identifiability across independent training runs. Our analysis shows that neural networks can admit large families of approximately equivalent solutions even in *structurally asymmetric* models. We further show that neuron identifiability enables representation merging *without prior alignment*, and characterize when such merging admits a linear low-loss path. These findings highlight the role of effective function classes in affecting the loss landscape.

1. Introduction

Modern neural networks are overparameterized and have highly nonconvex loss landscapes, yet independently trained models often achieve similar predictions and generalization (Goodfellow et al., 2015; Nguyen et al., 2019; Zhang et al., 2017; Li et al., 2018). One reason is that architectures admit large parameter symmetry groups (Hecht-Nielsen, 1990; Zhao et al., 2026), which affect optimization geometry and model merging. The most ubiquitous are permutations of hidden units (Simsek et al., 2021; Zhao et al., 2022; 2024), which are central to Linear Mode Connectivity (LMC) (Frankle et al., 2020): independently trained networks can often be linearly interpolated with little loss degradation only after accounting for hidden-unit permutations (Entezari et al., 2022; Singh & Jaggi, 2020; Ainsworth et al., 2023).

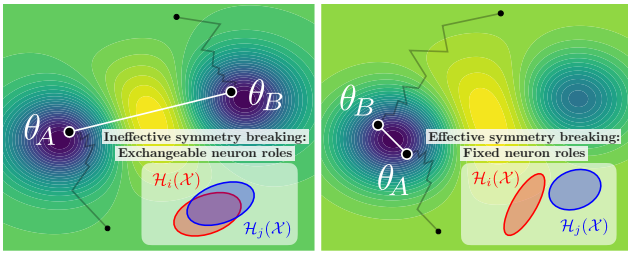


Figure 1. Effect of neuron identifiability on the loss landscape.

Recent work has shown that breaking structural symmetries can yield unaligned LMC (Lim et al., 2024b; Ziyin et al., 2025). However, this outcome is not universal. Since both data and representations are often close to low-rank (Ansuini et al., 2019), transformations may remain approximate, data-dependent symmetries on actual inputs to a layer while differing substantially elsewhere. In this work, we ask:

> When does a given parameter symmetry breaking mechanism select a consistent assignment of features to neurons across runs, and how do the data and representation distribution control its effectiveness?

Our Contributions. We develop a theoretical framework for *neuron identifiability*, i.e., the consistent assignment of features to neurons across random training seeds. In each layer, we view neurons functionally on their input support. Given a symmetry-breaking mechanism and a regularization budget, we characterize their corresponding *effective function classes*. This perspective shows that effective symmetry breaking is controlled by which architectural perturbations are observable on the input, and explains when hidden-layer representations can be merged without alignment, leading to unaligned LMC. Overall, we show that symmetry breaking is not binary, but governed by the interaction between architecture, data geometry, and effective function classes.

[*]Equal contribution [1]Technical University of Munich; School of CIT, MCML, MDSI [2]MIT; Dept. of EECS, CSAIL. Correspondence to: Vincent Bürgin <vincent.buergin@tum.de>, Daniel Herbst <daniel.herbst@tum.de>.
Code: `github.com/vuenc/neuron-identifiability`.
This is a short version of the work with the same title at the 43rd International Conference on Machine Learning (ICML 2026).

Workshop on Weight-Space Symmetries, held in conjunction with the 43rd International Conference on Machine Learning, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

2. Preliminaries

Permutation Symmetries. Consider a 2-layer MLP $f_{\boldsymbol{\theta}}(\boldsymbol{x}) := \boldsymbol{W}_2\eta(\boldsymbol{W}_1\boldsymbol{x})$, $\boldsymbol{\theta} = (\boldsymbol{W}_2, \boldsymbol{W}_1)$, where η is an elementwise nonlinearity. For any permutation $\pi \in S_m$ with permutation matrix $\boldsymbol{P}_\pi \in \{0,1\}^{m\times m}$, the mapping $(\boldsymbol{W}_2, \boldsymbol{W}_1) \mapsto (\boldsymbol{W}_2\boldsymbol{P}_\pi^\top, \boldsymbol{P}_\pi\boldsymbol{W}_1)$ is a parameter symmetry. In other words, hidden units are interchangeable, i.e., permuting them and compensating in the adjacent layer does not change the computed function. Depending on η, this group can also be significantly larger (e.g., $\mathrm{GL}(m)$ for linear η). Similar symmetries appear in modern models, e.g., channels in CNNs, attention heads, MoEs, or through any computation graph automorphism (Lim et al., 2024a;b).

Symmetry Breaking. Symmetry breaking refers to any modification of a network's forward pass that reduces the effective parameter symmetry group so that fewer distinct parameters represent the same function. In this work, we study an architectural intervention that replaces each trainable weight matrix $\boldsymbol{W}$ with an effective matrix

$$\boldsymbol{W}_{\text{eff}} = \mathbf{F} + \mathbf{D} \odot \boldsymbol{W}, \tag{1}$$

where $\mathbf{F}, \mathbf{D} \in \mathbb{R}^{m\times d}$ are considered fixed parts of the architecture, $\odot$ denotes elementwise multiplication, and $\boldsymbol{W} \in \mathbb{R}^{m\times d}$ contains the trainable parameters that are randomly initialized independently across runs and then optimized. $\mathbf{F}, \mathbf{D} \in \mathbb{R}^{m\times d}$ can be fixed upfront or sampled once per architecture. As summarized in Table 1, this subsumes several existing parameter symmetry breaking schemes.

Table 1. Symmetry breaking schemes covered by (1).

	$\mathbf{F}$	$\mathbf{D}$
***W**-asym.* (Lim et al., 2024b)	$\mathbf{M} \odot \mathbf{B}$ where $\mathbf{B} \sim \mathcal{N}(0, \sigma_{\mathbf{F}}^2)$ i.i.d., $\mathbf{M} \in \{0,1\}^{m\times d}$	$\mathbf{1}\mathbf{1}^\top - \mathbf{M}$
syre (Ziyin et al., 2025)	$\mathcal{N}(0, \sigma_{\mathbf{F}}^2)$ i.i.d.	$\mathrm{Unif}(1-\varepsilon, 1+\varepsilon)^{-1/2}$ i.i.d.
Linear residual	$\boldsymbol{I}_d$	$\mathbf{1}\mathbf{1}^\top$
Sparse network	0	$\mathbf{M} \in \{0,1\}^{m\times d}$

For ***W**-asymmetric networks* (Lim et al., 2024b), a binary mask $\mathbf{M}$ selects a subset of coordinates that $\mathbf{F} := \mathbf{M} \odot \mathbf{B}$ fixes to constant random values, while unmasked coordinates remain trainable. *syre* (Ziyin et al., 2025) directly draws $\mathbf{F}$ and $\mathbf{D}$ i.i.d. Linear residual connections and sparse networks, both of which are known to break hidden-unit permutation symmetries (Lim et al., 2024b; Zhao et al., 2025), can also be interpreted in this framework.

Setup and Assumptions. Throughout most of the work, we will consider an asymmetric layer (1) in isolation. Concretely, we study the layer map $\boldsymbol{H} : \mathbb{R}^{m\times d} \times \mathbb{R}^d \to \mathbb{R}^m$,

$$\boldsymbol{H}(\boldsymbol{W}; \boldsymbol{x}) := \eta((\mathbf{F} + \mathbf{D} \odot \boldsymbol{W})\boldsymbol{x}). \tag{2}$$

Let $\mathbf{f}_i, \mathbf{d}_i$, and $\boldsymbol{w}_i$ be the i-th rows of $\mathbf{F}, \mathbf{D}$, and $\boldsymbol{W}$. The layer's i-th neuron in (2) can be written as $\boldsymbol{h}_i(\boldsymbol{w}_i; \boldsymbol{x}) := \eta((\mathbf{f}_i + \mathbf{d}_i \odot \boldsymbol{w}_i)^\top \boldsymbol{x})$. Let the input $\boldsymbol{x}$ be supported on $\mathcal{X} \subset \mathbb{R}^d$ and drawn from a distribution $\mathbb{P}$ on $\mathcal{X}$. Define $\mathcal{H}(\mathcal{X}) := \{\boldsymbol{H}(\boldsymbol{W}; \cdot) \mid \boldsymbol{W} \in \mathbb{R}^{m\times d}\}$ and analogously $\mathcal{H}_i(\mathcal{X})$ as the set of functions a layer (or neuron) can implement. $\mathcal{X}$ is typically far from filling the ambient space $\mathbb{R}^d$. For our analysis, we assume that $\mathbb{P}$ is supported on a k-dim. subspace $\mathcal{U} \subset \mathbb{R}^d$ and fix an orthonormal basis $\boldsymbol{U} \in \mathbb{R}^{d\times k}$ of $\mathcal{U}$ (§ C). While this is idealized (§ E.1), it enables exact projections onto $\mathcal{U}$ and provides explicit coordinates via $\boldsymbol{U}$.

3. Neuron Identifiability

We now relate the effectiveness of a symmetry breaking intervention according to (1) to its ability of equipping neurons with distinguishable identities that lead to consistent assignments across training runs. Within one layer, each neuron can be viewed as implementing a particular feature. Across training runs, a layer often realizes essentially the same *set* of features, and runs differ mainly by a permutation that matches neurons. In a symmetric layer, this assignment is determined by the initialization. By *effective symmetry breaking* we mean that, instead, architectural asymmetries bias training towards a consistent assignment. We analyze this through each neuron's *effective function class*, i.e., the functions it can realize on the input support together with their *realization costs*, measured by the minimum weight norm required to implement them. This is well-motivated under explicit weight decay and the implicit norm bias of gradient-based optimization. The underlying premise is that uniformly large reassignment costs over nontrivial permutations are equivalent to there being a unique minimum-complexity assignment of the feature set to neurons.

We additionally assume $\mathrm{span}(\mathrm{supp}(\mathbb{P})) = \mathcal{U}$ if η is injective (e.g., tanh, LeakyReLU). If $s \mapsto (\eta(s), \eta(-s))$ is injective (e.g., ReLU, GELU), assume there is $T \subseteq \mathrm{supp}(\mathbb{P})$ s.t. $T = -T$ and $\mathrm{span}(T) = \mathcal{U}$. This lets us write realizability of a feature directly in subspace coordinates: Setting $\mathbf{v}_i := \boldsymbol{U}^\top \mathbf{f}_i \in \mathbb{R}^k$, $\mathbf{M}_i := \boldsymbol{U}^\top \mathrm{Diag}(\mathbf{d}_i) \in \mathbb{R}^{k\times d}$, $\boldsymbol{h} \in \mathcal{H}_i(\mathcal{X})$ can be written as $\boldsymbol{h}(\boldsymbol{x}) = \eta((\boldsymbol{U}\boldsymbol{a})^\top \boldsymbol{x})$, where

$$\boldsymbol{a} := \boldsymbol{U}^\top\big(\mathbf{f}_i + \mathrm{Diag}(\mathbf{d}_i)\boldsymbol{w}\big) = \mathbf{v}_i + \mathbf{M}_i\boldsymbol{w} \in \mathbb{R}^k. \tag{3}$$

The map $\boldsymbol{h} \mapsto \boldsymbol{a}$ is a bijection that identifies $\mathcal{H}_i(\mathcal{X})$ with $\mathbf{v}_i + \mathrm{im}(\mathbf{M}_i) \subseteq \mathbb{R}^k$ (cf. Prop. C.4). In particular, $\mathcal{H}_i(\mathcal{X})$ is an affine subspace of $\mathbb{R}^k$ of dimension $\mathrm{rk}(\mathbf{M}_i) \le k$, which allows for a geometric comparison of neuron function classes. This implies that if all $\mathbf{M}_i$ have full rank, all neurons can represent exactly the same functions, i.e., $\mathcal{H}_1(\mathcal{X}) = \cdots = \mathcal{H}_m(\mathcal{X}) \cong \mathbb{R}^k$. We mainly focus on this case.

Realization Costs. Having characterized *which* functions a neuron can represent, we now quantify *at which cost*. We define the *realization cost* of $\boldsymbol{h} : \mathcal{X} \to \mathbb{R}$ w.r.t. $\mathcal{H}_i(\mathcal{X})$ as

$$\|\boldsymbol{h}\|_{\mathcal{H}_i(\mathcal{X})} := \inf\{\|\boldsymbol{w}\|_2 \mid \boldsymbol{h} = \boldsymbol{h}_i(\boldsymbol{w}; \cdot)\}, \tag{4}$$

with the convention $\|\boldsymbol{h}\|_{\mathcal{H}_i(\mathcal{X})} := +\infty$ for $\boldsymbol{h} \notin \mathcal{H}_i(\mathcal{X})$. Define $\|\cdot\|_{\mathcal{H}(\mathcal{X})}$ analogously w.r.t. the Frobenius norm $\|\cdot\|_F$.[1]

[1] $\|\cdot\|_{\mathcal{H}(\mathcal{X})}$ is not a norm ($\boldsymbol{H}(0; \cdot)$ can be $\neq 0$ but has cost 0).

(4) can be computed in closed form as a Mahalanobis seminorm: Defining the Gram matrix $\mathbf{S}_i := \mathbf{M}_i\mathbf{M}_i^\top = \boldsymbol{U}^\top \mathrm{Diag}(\mathbf{d}_i)^2\boldsymbol{U}$, $\boldsymbol{h} = \eta((\boldsymbol{U}\boldsymbol{a})^\top \cdot)$ has realization cost

$$\left\|\eta((\boldsymbol{U}\boldsymbol{a})^\top \cdot)\right\|_{\mathcal{H}_i(\mathcal{X})} = \|\boldsymbol{a} - \mathbf{v}_i\|_{\mathbf{S}_i^\dagger}, \tag{5}$$

see Lem. C.5. Here, $\|\boldsymbol{x}\|_{\boldsymbol{A}} := \sqrt{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}$ and $\dagger$ indicates the pseudoinverse. This shows that the realization costs w.r.t. neuron i depend crucially on the anisotropy of its Gram matrix $\mathbf{S}_i$. If $\mathbf{S}_i \approx c\,\boldsymbol{I}_k$, the cost is near-Euclidean, whereas anisotropy makes some features much more expensive than others. In the case of the diagonal asymmetry $\mathbf{d}_i$ being sampled i.i.d. with mean $\mu_{\mathbf{D}}$, $\mathbb{E}[\mathbf{S}_i] = \mu_{\mathbf{D}}\boldsymbol{I}_k$, and the deviation $\|\mathbf{S}_i - \mu_{\mathbf{D}}\boldsymbol{I}_k\|_{\mathrm{op}}$ can be controlled by the *subspace coherence* (Candes & Tao, 2005; Candes & Recht, 2009)

$$\nu(\mathcal{U}) := \max_{\ell\in[d]} (\boldsymbol{U}\boldsymbol{U}^\top)_{\ell\ell} \in [k/d, 1]. \tag{6}$$

We make this precise in Thm. C.6. Intuitively, $\nu(\mathcal{U})$ quantifies how much of the k-dim. variability of the input can be seen from any ambient coordinate. High coherence (≈ 1) makes the effect of coordinate masking on symmetry breaking pronounced; low coherence ($\approx k/d$) can diffuse it across ambient dimensions. This means that the impact of $\mathbf{D}$ is not severe if the input is sufficiently incoherent, which is often the case in data or hidden representations. Hence, one should not automatically expect $\mathbf{D}$ alone to effectively break symmetries. We now discuss the interplay of $\mathbf{D}$ and $\mathbf{F}$.

Realization Cost Sensitivity. For a layer $\boldsymbol{H} : \mathbb{R}^d \to \mathbb{R}^m$, define a group action of $(\pi, \tau) \in S_m \times S_d$:

$$\big((\pi,\tau)\cdot \boldsymbol{H}\big)_i := \boldsymbol{h}_{\pi(i)}(\boldsymbol{P}_\tau^\top \cdot), \tag{7}$$

and define the *(square) realization cost sensitivity*

$$\Delta_{\pi,\tau}(\boldsymbol{H};\mathcal{X}) := \|(\pi,\tau)\cdot \boldsymbol{H}\|^2_{\mathcal{H}(\boldsymbol{P}_\tau\mathcal{X})} - \|\boldsymbol{H}\|^2_{\mathcal{H}(\mathcal{X})} \tag{8}$$

(we will often omit $\mathcal{X}$ for brevity). It decomposes as

$$\Delta_{\pi,\tau}(\boldsymbol{H};\mathcal{X}) = \underbrace{\Delta_{\pi,\mathrm{id}}}_{=:\Delta^{\mathrm{out}}_\pi}((\mathrm{id},\tau)\cdot\boldsymbol{H};\boldsymbol{P}_\tau\mathcal{X}) + \underbrace{\Delta_{\mathrm{id},\tau}}_{=:\Delta^{\mathrm{in}}_\tau}(\boldsymbol{H};\mathcal{X}). \tag{9}$$

This allows us to separately consider the effects of permuting the inputs (via τ) and outputs (via π) of a layer, for which we will write $\Delta^{\mathrm{in}}_\tau, \Delta^{\mathrm{out}}_\pi$ respectively. In both cases, the sensitivity consists of a displacement term of $\mathbf{F}\boldsymbol{U}$ plus terms from the trainable weights and anisotropy $\mathbf{S}_i$ (cf. Lem. C.8). We focus on the *center-dominated regime*, where the former dominates and $\mathbf{S}_i \approx \mu_{\mathbf{D}}\boldsymbol{I}_k$ are close to isotropic.

Neuron Reassignments ($\Delta^{\mathrm{out}}_\pi$). Assume that all $\mathbf{S}_i$ have full rank, so all neurons can realize the same functions on $\mathcal{X}$. Then, $\pi \in S_m$ only changes which neuron pays the realization cost for which feature. In the center-dominated regime, the leading effect is the projected center displacement

$$\boldsymbol{\delta}^{\mathrm{out}}_\pi := (\boldsymbol{P}_\pi - \boldsymbol{I}_m)\mathbf{F}\boldsymbol{U} \in \mathbb{R}^{m\times k}. \tag{10}$$

When the Gram matrices are approximately isotropic, $\mathbf{S}_i \approx \mu_{\mathbf{D}}\boldsymbol{I}_k$, the realization cost sensitivity satisfies $\Delta^{\mathrm{out}}_\pi(\boldsymbol{H}) \sim \mu_{\mathbf{D}}^{-1}\|\boldsymbol{\delta}^{\mathrm{out}}_\pi\|_F^2$ up to lower-order terms from the trainable part and $\mathbf{S}_i$ (see Thm. C.10). Thus, the smallest neuron reassignment cost is

$$\min_{\pi\neq\mathrm{id}} \Delta^{\mathrm{out}}_\pi(\boldsymbol{H}) \sim \mu_{\mathbf{D}}^{-1}\gamma^2_{\mathrm{out}}, \quad \gamma_{\mathrm{out}} := \min_{i\neq j} \|\boldsymbol{\delta}^{\mathrm{out}}_{(ij)}\|_F. \tag{11}$$

Large γ_{out} gives neurons stable identities, while small γ_{out} leaves approximate symmetries. For sampled $\mathbf{F}$ with centered i.i.d. entries of std. / scale $\sigma_{\mathbf{F}}$,

$$\gamma_{\mathrm{out}} = \Theta(\sigma_{\mathbf{F}}\sqrt{k}\,m^{-2/k}) \tag{12}$$

with high probability (Thm. C.12). Thus, the gap decays slowly with width when the intrinsic dim. k is moderate, but can collapse for low-dim. data or hidden representations.

Input Reindexings ($\Delta^{\mathrm{in}}_\tau$). Input permutations $\tau \in S_d$ change how the fixed weights align with the input subspace. The corresponding projected center displacement is

$$\boldsymbol{\delta}^{\mathrm{in}}_\tau := \mathbf{F}(\boldsymbol{P}_\tau - \boldsymbol{I}_d)\boldsymbol{U} \in \mathbb{R}^{m\times k}. \tag{13}$$

In the same center-dominated and near-isotropic regime, $\Delta^{\mathrm{in}}_\tau(\boldsymbol{H}) \sim \mu_{\mathbf{D}}^{-1}\|\boldsymbol{\delta}^{\mathrm{in}}_\tau\|_F^2$, again up to trainable part and Gram anisotropy effects (Thm. C.13). For transpositions (ab),

$$\min_{a\neq b} \Delta^{\mathrm{in}}_{(ab)}(\boldsymbol{H}) \sim \mu_{\mathbf{D}}^{-1}\gamma^2_{\mathrm{in}}, \quad \gamma_{\mathrm{in}} := \min_{a\neq b} \|\boldsymbol{\delta}^{\mathrm{out}}_{(ab)}\|_F. \tag{14}$$

For i.i.d. centered subgaussian fixed weights,

$$\gamma_{\mathrm{in}} = \Theta(\sigma_{\mathbf{F}}\sqrt{m}) \min_{a\neq b} \|\boldsymbol{U}_{b,:} - \boldsymbol{U}_{a,:}\|_2 \tag{15}$$

with high probability (Thm. C.15). $\|\boldsymbol{U}_{b,:} - \boldsymbol{U}_{a,:}\|_2$ measures how distinguishable a and b are on the input. If swapping acts almost trivially on $\mathcal{X}$, sensitivity is small. Since $\|\boldsymbol{U}_{a,:} - \boldsymbol{U}_{b,:}\|_2 \le 2\sqrt{\nu(\mathcal{U})}$, low coherence can mitigate sensitivity; asymmetry matters only along input-visible directions.

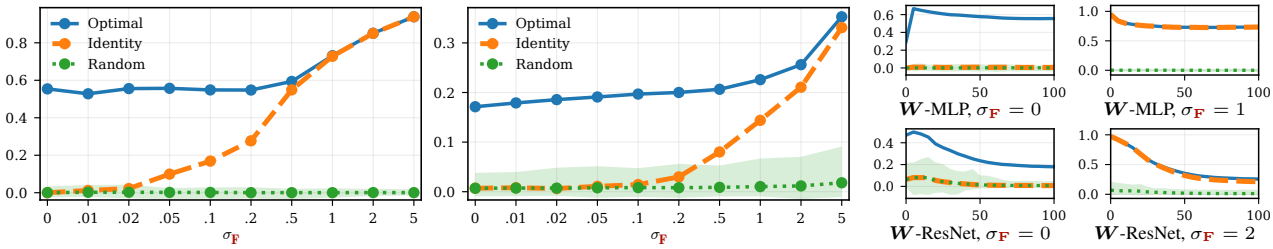
4. Linear Mode Connectivity

We also analyze how intermediate representations evolve when we linearly interpolate between two independently trained models. LMC asks whether the loss stays small along this segment. A natural sufficient condition is that the network's hidden representations vary approximately affinely along the interpolation path; recent work has observed this co-occurs with LMC (Zhou et al., 2023). We therefore quantify the nonlinearity induced by a single layer along the path, and relate it to an upper bound on the loss barrier. Given $\boldsymbol{W}^A, \boldsymbol{W}^B \in \mathbb{R}^{m\times d}$, consider $\boldsymbol{W}(\lambda) := (1-\lambda)\boldsymbol{W}^A + \lambda\boldsymbol{W}^B$ for $\lambda \in [0,1]$ and measure nonlinearity along it by the chord deviation $\xi_{\boldsymbol{H}}(\lambda;\boldsymbol{x}) := \boldsymbol{H}(\boldsymbol{W}(\lambda);\boldsymbol{x}) - \big((1-\lambda)\boldsymbol{H}(\boldsymbol{W}^A;\boldsymbol{x}) + \lambda\boldsymbol{H}(\boldsymbol{W}^B;\boldsymbol{x})\big)$.

For $\eta = \mathrm{ReLU}$, $\boldsymbol{x} \sim \mathcal{N}(0,\boldsymbol{\Sigma})$, in the center-dom. regime $\|\mathbf{d}_i \odot \boldsymbol{w}_i^X\|_{\boldsymbol{\Sigma}} \le \beta\,\|\mathbf{f}_i\|_{\boldsymbol{\Sigma}}$, $X \in \{A,B\}$, $\beta < 1$, we find that

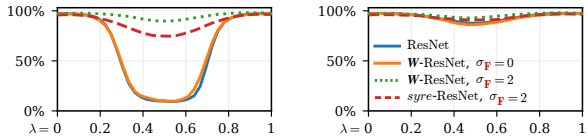
$$\sup_{\lambda\in[0,1]} \left\|\xi_{\boldsymbol{H}}(\lambda;\cdot)\right\|_{L^2(\mathbb{P};\mathbb{R}^m)} = \mathcal{O}(\beta^{3/2})\,\|\mathbf{F}\boldsymbol{\Sigma}^{1/2}\|_F, \tag{16}$$

i.e., the chord deviation goes to zero as $\beta \to 0$, and its scale depends on the total norm of the fixed part under the input covariance (Thm. D.1). With control over chord deviations, one can upper bound the loss along the linear path between two independently trained models (cf. Prop. D.3).



(a) $\boldsymbol{W}$-MLP, different $\sigma_{\mathbf{F}}$ at epoch 100 *(b)* $\boldsymbol{W}$-ResNet, different $\sigma_{\mathbf{F}}$ at epoch 100 *(c)* $\boldsymbol{W}$-asym. models over different epochs

Figure 2. Act. matching objs. of optimal, identity, and random permutations w/ diff. $\sigma_{\mathbf{F}}$, using post-norm, post-act. function values.



(a) ResNets on CIFAR-10, unaligned *(b)* ResNets on CIFAR-10, aligned

Figure 3. Aligned and unaligned LMC interpolation.

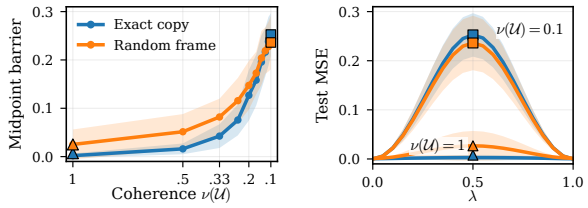


Figure 4. LMC barriers on synthetic dataset, $\mathbf{F} = 0$, varying $\nu(\mathcal{U})$.

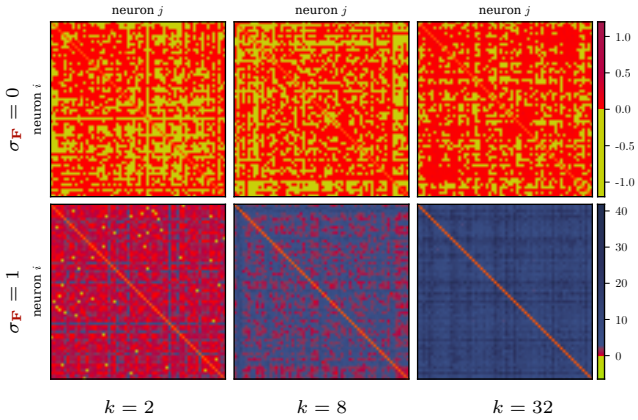


Figure 5. $\Delta^{\text{out}}_{(ij)}$ on synthetic Gaussian mixture data w/ varying intrinsic dim. k, positive/near-zero/negative values.

5. Experimental Results

Q1: *Do symmetry-breaking architectures improve unaligned LMC?* We compare standard, $\boldsymbol{W}$-asym., and *syre*-MLPs on MNIST, and the corresponding ResNets on CIFAR-10. We measure unaligned LMC and activation-matched LMC, using $\sigma_{\mathbf{F}} = 0$ and center-dom. $\boldsymbol{W}$-asym. settings ($\sigma_{\mathbf{F}} = 1$ for MLPs, 2 for ResNets).[2] Fig. 7 shows that large fixed weights yield unaligned LMC in MLPs and partly in ResNets, while standard and $\boldsymbol{W}$-asym. models with $\sigma_{\mathbf{F}} = 0$ have high unaligned barriers. After alignment, all barriers are low, including for $\boldsymbol{W}$-asym. models with $\sigma_{\mathbf{F}} = 0$ and *syre*-ResNets. Thus structural asymmetry alone is insufficient.

Q2: *When do $\boldsymbol{W}$-asym. networks learn the same assignment across runs?* We sweep $\sigma_{\mathbf{F}}$ and compare the activation-matching objective of the identity, the optimal permutation, and random permutations. Figs. 2a and 2b show that, once $\sigma_{\mathbf{F}}$ is large enough, the identity becomes nearly optimal while random permutations remain much worse; Fig. 2c shows the same trend over training. Hence sufficiently large fixed centers make neuron identities consistent across runs.

Q3: *Do swap costs depend on fixed scale and intrinsic dimension?* We estimate realization costs using (5) on the 90% PCA subspace and compute $\Delta^{\text{out}}_{(ij)}$. The theory predicts $\Delta^{\text{out}}_{\pi} \sim \mu_{\mathbf{D}}^{-1} \|\boldsymbol{\delta}^{\text{out}}_{\pi}\|_F^2$ and $\gamma_{\text{out}} = \Theta(\sigma_{\mathbf{F}} \sqrt{k} m^{-2/k})$. On Gaussian mixtures with width $m = 64$ and $k \in \{2, 8, 32\}$, Fig. 5 shows many cheap swaps for $\sigma_{\mathbf{F}} = 0$, but increasing separation for $\sigma_{\mathbf{F}} = 1$ as k grows. The unaligned midpoint accuracy drops are 46.4, 15.7, 6.1 for $k = 2, 8, 32$, confirming that low intrinsic dim. leaves approx. swaps accessible.

Q4: *How does subspace coherence affect symmetry breaking?* With $\mathbf{F} = 0$, diagonal masks should matter mainly when coordinate directions are visible on the data subspace. We test this by embedding $a, b \sim \mathcal{N}(0, 1)$ into $\mathbb{R}^{20}$ by copying each coordinate to $t \leq 10$ positions, giving $\nu(\mathcal{U}) = 1/t$, and train a one-hidden-layer $\boldsymbol{W}$-asym. MLP to fit $\mathrm{ReLU}(a + b)$. Fig. 4 shows low barriers at high coherence and high barriers at low coherence, also under a random-frame variant. Thus, diagonal asymmetry alone is effective mainly along data-visible coordinate directions.

6. Conclusion

We studied when parameter symmetry breaking yields neuron identifiability, i.e., a consistent assignment of features to neurons. This explains when asymmetry can be present in parameter space yet ineffective on data. We also showed that effective symmetry breaking can yield unaligned LMC. We leave extensions to richer symmetry groups, exact training dynamics, and more practical diagnostics for future work.

[2] For asymmetric models, alignment is applied to effective weights $\boldsymbol{W}_{\text{eff}} = \mathbf{F} + \mathbf{D} \odot \boldsymbol{W}$.

Acknowledgements

We thank the anonymous reviewers for their insightful feedback, and Valerie Engelmayer for providing feedback on an earlier version of this manuscript. VB acknowledges support from a PhD fellowship of the Munich Center for Machine Learning (MCML). DH acknowledges support from the Munich Data Science Institute (MDSI) via the MDSI Doctoral Fellowship. YEL acknowledges support from the Schmidt Futures Israeli Women's Postdoctoral Award. This research was supported by an Alexander von Humboldt Professorship.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Adilova, L., Andriushchenko, M., Kamp, M., Fischer, A., and Jaggi, M. Layer-wise linear mode connectivity. In *The Twelfth International Conference on Learning Representations*, 2024.

Ainsworth, S., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023.

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Bar, N. and Giryes, R. Pruning at initialization–a sketching perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Candes, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2009.

Candes, E. J. and Tao, T. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

Chen, A. M., Lu, H.-m., and Hecht-Nielsen, R. On the geometry of feedforward neural network error surfaces. *Neural computation*, 5(6):910–927, 1993.

Crisostomi, D., Fumero, M., Baieri, D., Bernard, F., and Rodolà, E. C^2M^3: Cycle-consistent multi-model merging. In *Advances in Neural Information Processing Systems*, 2024.

Dai, Z., Karzand, M., and Srebro, N. Representation costs of linear neural networks: Analysis and design. *Advances in Neural Information Processing Systems*, 34, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning*, pp. 1309–1318. PMLR, 2018.

Drineas, P., Magdon-Ismail, M., Mahoney, M. W., and Woodruff, D. P. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012.

Entezari, R., Sedghi, H., Saukh, O., and Neyshabur, B. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022.

Evci, U., Gale, T., Menick, J., Castro, P. S., and Elsen, E. Rigging the lottery: Making all tickets winners. In *International Conference on Machine Learning*, pp. 2943–2952. PMLR, 2020.

Ferbach, D., Goujaud, B., Gidel, G., and Dieuleveut, A. Proving linear mode connectivity of neural networks via optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pp. 3853–3861. PMLR, 2024.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

Frankle, J., Dziugaite, G. K., Roy, D. M., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, 2020.

Frankle, J., Dziugaite, G. K., Roy, D. M., and Carbin, M. Pruning neural networks at initialization: Why are we missing the mark? In *The Ninth International Conference on Learning Representations*, 2021.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059. PMLR, 2016.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of DNNs. *Advances in Neural Information Processing Systems*, 31, 2018.

Goodfellow, I. J., Vinyals, O., and Saxe, A. M. Qualitatively characterizing neural network optimization problems. *The Third International Conference on Learning Representations*, 2015.

Grigsby, E., Lindsey, K., and Rolnick, D. Hidden symmetries of relu networks. In *International Conference on Machine Learning*, pp. 11734–11760. PMLR, 2023.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Hecht-Nielsen, R. On the algebraic structure of feedforward network weight spaces. In *Advanced Neural Computers*, pp. 129–135. Elsevier, 1990.

Imfeld, M., Graldi, J., Giordano, M., Hofmann, T., Anagnostidis, S., and Singh, S. P. Transformer fusion with optimal transport. In *The Twelfth International Conference on Learning Representations*, 2024.

Ito, A., Yamada, M., Chijiwa, D., and Kumagai, A. Do we really need permutations? Impact of model width on linear mode connectivity. In *The Fourteenth International Conference on Learning Representations*, 2026.

Lee, N., Ajanthan, T., and Torr, P. H. S. SNIP: Single-shot network pruning based on connection sensitivity. In *The Seventh International Conference on Learning Representations*, 2019.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems*, 31, 2018.

Li, Y., Yosinski, J., Clune, J., Lipson, H., and Hopcroft, J. Convergent learning: Do different neural networks learn the same representations? In Storcheus, D., Rostamizadeh, A., and Kumar, S. (eds.), *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, volume 44 of *Proceedings of Machine Learning Research*, pp. 196–212, Montreal, Canada, 11 Dec 2015. PMLR.

Lim, D., Maron, H., Law, M. T., Lorraine, J., and Lucas, J. Graph metanetworks for processing diverse neural architectures. In *The Twelfth International Conference on Learning Representations*, 2024a.

Lim, D., Putterman, T., Walters, R., Maron, H., and Jegelka, S. The empirical impact of neural parameter symmetries, or lack thereof. *Advances in Neural Information Processing Systems*, 37:28322–28358, 2024b.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Nguyen, Q., Mukkamala, M. C., and Hein, M. On the loss landscape of a class of deep neural networks with no bad local valleys. In *International Conference on Learning Representations*, 2019.

Pittorino, F., Ferraro, A., Perugini, G., Feinauer, C., Baldassi, C., and Zecchina, R. Deep networks on toroids: Removing symmetries reveals the structure of flat regions in the landscape geometry. In *ICML*, pp. 17759–17781, 2022.

Rudelson, M. and Vershynin, R. Small ball probabilities for linear images of high-dimensional distributions. *International Mathematics Research Notices*, 19:9594–9617, 2015.

Savarese, P., Evron, I., Soudry, D., and Srebro, N. How do infinite width bounded norm networks look in function space? In *The Thirty Second Annual Conference on Learning Theory*. PMLR, 2019.

Simsek, B., Ged, F., Jacot, A., Spadaro, F., Hongler, C., Gerstner, W., and Brea, J. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *ICML*, pp. 9722–9732, 2021.

Singh, S. P. and Jaggi, M. Model fusion via optimal transport. *Advances in Neural Information Processing Systems*, 33:22045–22055, 2020.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Tanaka, H., Kunin, D., Yamins, D. L. K., and Ganguli, S. Pruning neural networks without any data by iteratively conserving synaptic flow. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Tatro, N., Chen, P.-Y., Das, P., Melnyk, I., Sattigeri, P., and Lai, R. Optimizing mode connectivity via neuron alignment. *Advances in Neural Information Processing Systems*, 33:15300–15311, 2020.

Theus, A., Cabodi, A., Anagnostidis, S., Orvieto, A., Singh, S. P., and Boeva, V. Generalized linear mode connectivity for transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Tropp, J. A. An introduction to matrix concentration inequalities. *Foundations and trends® in machine learning*, 8(1-2):1–230, 2015.

Verma, N. and Elbayad, M. Merging text transformer models from different initializations. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Vershynin, R. *High-Dimensional Probability*. Cambridge Series in Statistical and Probabilistic Mathematics. 2018.

Wager, S., Wang, S., and Liang, P. S. Dropout training as adaptive regularization. *Advances in Neural Information Processing Systems*, 26, 2013.

Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., and Fergus, R. Regularization of neural networks using DropConnect. In *International Conference on Machine Learning*, pp. 1058–1066. PMLR, 2013.

Wang, C., Zhang, G., and Grosse, R. Picking winning tickets before training by preserving gradient flow. In *The Eighth International Conference on Learning Representations*, 2020.

Xie, Y. and Smidt, T. A tale of two symmetries: Exploring the loss landscape of equivariant models. *Advances in Neural Information Processing Systems*, 38:53223–53255, 2025.

Yehudai, G. and Shamir, O. On the power and limitations of random features for understanding neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Zhao, B., Dehmamy, N., Walters, R., and Yu, R. Symmetry teleportation for accelerated optimization. *Advances in Neural Information Processing Systems*, 35:16679–16690, 2022.

Zhao, B., Ganev, I., Walters, R., Yu, R., and Dehmamy, N. Symmetries, flat minima, and the conserved quantities of gradient flow. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhao, B., Gower, R. M., Walters, R., and Yu, R. Improving convergence and generalization using parameter symmetries. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhao, B., Dehmamy, N., Walters, R., and Yu, R. Understanding mode connectivity via parameter space symmetry. In *Forty-second International Conference on Machine Learning*, 2025.

Zhao, B., Walters, R., and Yu, R. Symmetry in neural network parameter spaces. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Zhou, Z., Yang, Y., Yang, X., Yan, J., and Hu, W. Going beyond linear mode connectivity: The layerwise linear feature connectivity. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Ziyin, L., Xu, Y., and Chuang, I. L. Remove symmetries to control model expressivity and improve optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Appendices

A Notation **9**

B Related Work **10**

B.1 Parameter Symmetries and Mode Connectivity . . . 10

B.2 Fixed Masks, Pruning, and Sketching . . . 10

B.3 Representation Costs . . . 11

C Neuron Identifiability (§ 3) **12**

C.1 Realization Costs . . . 12

C.2 Neuron Reassignments . . . 15

C.3 Input Reindexings . . . 20

C.4 Identifiability vs. Feature Learning . . . 23

D Linear Mode Connectivity (§ 4) **25**

E Approximate Subspace Support and Relaxed Realization Costs **29**

E.1 Approximate Subspace Support . . . 29

E.2 Relaxed Realization Costs . . . 30

F Additional Figures **33**

F.1 Visualization of Effective Function Classes . . . 33

F.2 LMC Interpolation for MLPs and ResNets . . . 33

G Experimental Setup **34**

G.1 MLPs . . . 34

G.2 ResNets . . . 34

G.3 Datasets and Augmentations . . . 34

G.4 Activation Matching . . . 34

G.5 Coherence Experiment . . . 35

G.6 Gaussian Mixture Synthetic Dataset . . . 35

G.7 Empirical Neuron Realization and Swap Costs . . . 35

H Additional Experimental Results **37**

H.1 Layerwise Activation Matching Objectives . . . 37

H.2 Empirical Validation of Center Separation Rate . . . 39

H.3 Neuron Swap Costs via Ridge Regression and Additional Pairwise Neuron Swap Costs . . . 39

H.4 Transformers . . . 43

H.5 Subspace Coherence of MNIST Inputs and Hidden Representations . . . 43

A. Notation

Table 2. We list the most important symbols used throughout this work.

$\mathbb{N}, \mathbb{N}_0, \mathbb{Q}, \mathbb{R}$	Natural, non-negative integer, rational, real numbers.
$[n]$	Set $\{1, \dots, n\}$ for $n \in \mathbb{N}$.
$\mathbb{1}_A$	Indicator function in a set A.
$\mathcal{O}(\cdot), \Theta(\cdot), \Omega(\cdot)$	Landau symbols for asymptotic growth of a function.
S_n	Symmetric group $\{\pi : [n] \to [n] \mid \pi \text{ bijective}\}$.
$\mathrm{id}, (ij), \pi$	Identity, transposition $i \leftrightarrow j$, generic permutation $\in S_n$.
$\boldsymbol{P}, \boldsymbol{P}_\pi \in \{0,1\}^{n \times n}$	Generic permutation matrix, permutation matrix of $\pi \in S_n$.
$\odot$	Elementwise multiplication of matrices.
$\boldsymbol{I}_n$	Identity matrix in n dimensions.
$\mathbf{1}_n$	Vector of all ones in n dimensions.
$\boldsymbol{e}_i$	i-th canonical basis vector.
$\mathrm{Unif}(\cdot)$	Uniform distribution on a set.
$\mathrm{Bernoulli}(p)$	Bernoulli distribution with success probability p.
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ, variance σ^2.
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate normal distribution with mean vector $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$.
$\lVert\cdot\rVert_p, \lVert\cdot\rVert_{\mathrm{op}}, \lVert\cdot\rVert_F, \lVert\cdot\rVert_{\boldsymbol{A}}$	ℓ^p norms ($p \in [1, \infty]$), spectral norm, Frobenius norm, Mahalanobis norm w.r.t. $\boldsymbol{A}$.
$\boldsymbol{A}^\dagger$	Moore-Penrose pseudoinverse of $\boldsymbol{A}$.
$\succ, \succeq, \prec, \preceq$	Löwner partial order for matrices.
$L^p(\mu)$	p-integrable functions w.r.t. a measure μ, $p \in [1, \infty]$.
$\lVert\cdot\rVert_{L^p(\mu)}, \lVert\cdot\rVert_{L^p(\mu, \mathbb{R}^m)}$	L^p norms of functions mapping to $\mathbb{R}$, $\mathbb{R}^m$, $p \in [1, \infty]$.
$f * g$	Convolution of two functions f, g.
λ^n	n-dimensional Lebesgue measure.
B_t^n	n-dimensional ball in $\mathbb{R}^n$, i.e., $\{\boldsymbol{x} \in \mathbb{R}^n \mid \lVert\boldsymbol{x}\rVert_2 < t\}$.
V_n	Volume $\lambda^n(B_1^n)$ of the n-dimensional unit ball.
$\lVert\cdot\rVert_{\psi_2}, \lVert\cdot\rVert_{\psi_1}$	Subgaussian and subexponential norms.
$\mathbb{S}^{n-1}$	n-dimensional sphere $\{\boldsymbol{x} \in \mathbb{R}^n \mid \lVert\boldsymbol{x}\rVert_2 = 1\}$.
$\boldsymbol{\theta}, \Theta$	(Total) parameter vector and parameter space of a neural architecture.
$\boldsymbol{W}, \boldsymbol{w}$	Weight matrix of a layer, weight vector of a single neuron.
d, m	Ambient input dimension, output dimension (i.e., width) of a layer.
η	Activation function.
$\mathbf{F}, \mathbf{D} \in \mathbb{R}^{m \times d}$	Fixed offset and diagonal operator of symmetry breaking intervention.
$\mathbf{f}_i, \mathbf{d}_i, \boldsymbol{w}_i \in \mathbb{R}^d$	i-th rows of $\mathbf{F}, \mathbf{D}, \boldsymbol{W}$, $i \in [m]$.
$\sigma_{\mathbf{F}}$	Standard deviation (fixed weight scale) of entry distribution of $\mathbf{F}$.
$\mu_{\mathbf{D}}, \sigma_{\mathbf{D}}, b_{\mathbf{D}}$	Mean, standard deviation, a.s. deviation of entry distribution of $\mathbf{d}_i^2$.
$\boldsymbol{x}, \mathcal{X}, \mathbb{P}$	Input, input support, input distribution of a layer ($\mathcal{X} = \mathrm{supp}(\mathbb{P})$).
$\boldsymbol{H}(\boldsymbol{W}, \boldsymbol{x}), \boldsymbol{h}_i(\boldsymbol{w}_i, \boldsymbol{x})$	Layer map of a symmetry-broken layer, single neuron i.
$\mathcal{H}, \mathcal{H}_i, \mathcal{H}_i(\mathcal{X})$	Fct. class of a layer, fct. class of neuron i, fct. class of neuron i w/ explicit input supp. $\mathcal{X}$.
k	Intrinsic input dimension $\leq d$.
$\mathcal{U}, \boldsymbol{U}$	Input subspace, orthonormal basis of $\mathcal{U}$.
$\mathbf{v}_i \in \mathbb{R}^k, \mathbf{M}_i \in \mathbb{R}^{k \times d}$	$\mathbf{F}_i, \mathrm{Diag}(\mathbf{d}_i)$ in projected $\boldsymbol{U}$ coordinates.
$\mathbf{S}_i \in \mathbb{R}^{k \times k}$	Gram matrix $\mathbf{S}_i = \mathbf{M}_i \mathbf{M}_i^\top$.
$\lVert\boldsymbol{h}\rVert_{\mathcal{H}_i(\mathcal{X})}$	Realization cost of $\boldsymbol{h} : \mathcal{X} \to \mathbb{R}$ w.r.t. neuron i on input support $\mathcal{X}$.
$\nu(\mathcal{U})$	Subspace coherence.
$\Delta_{\pi,\tau}(\boldsymbol{H}; \mathcal{X})$	Realization cost sensitivity of $\boldsymbol{H} \in \mathcal{H}(\mathcal{X})$ under output-input permutation pair (π, τ).
$\boldsymbol{\delta}_\pi^{\mathrm{out}}, \boldsymbol{\delta}_\tau^{\mathrm{in}} \in \mathbb{R}^{m \times k}$	Projected center differences under output / input permutation π / τ.
$\gamma_{\mathrm{out}}, \gamma_{\mathrm{in}}$	Minimum proj. center distances for output / input.
$\xi_{\boldsymbol{H}}(\lambda; \boldsymbol{x})$	Chord deviation of layer $\boldsymbol{H}$ at interpolation parameter $\lambda \in [0, 1]$.
$\ell, \mathcal{L}$	Loss on one data point, population loss.

B. Related Work

B.1. Parameter Symmetries and Mode Connectivity

Mode connectivity of neural networks was introduced by early works which observed that trained neural network solutions can be connected by paths in parameter space where the performance (measured by training or validation loss or accuracy) does not significantly worsen (Garipov et al., 2018; Draxler et al., 2018). The focus soon shifted to linear mode connectivity (LMC) (Frankle et al., 2020), which has been widely interpreted to occur when two solutions are in the same basin of the optimization landscape, which is a desirable property for applications such as model merging. Several settings which admit LMC have been identified: Frankle et al. (2020) found LMC for networks trained with different optimization randomness but starting from the same partially converged solution. Later studies obtained LMC by aligning networks with respect to symmetries (Ainsworth et al., 2023), or removing symmetries (Lim et al., 2024b). Variants of LMC focusing on individual layers have been investigated (Adilova et al., 2024; Zhou et al., 2023).

Exact structural neural network parameter symmetries, in particular *permutation symmetries*, have been investigated as early as Hecht-Nielsen (1990); Chen et al. (1993), where it was observed that the whole parameter space can be mapped to a small cone via function-preserving permutations. Recent work has also studied hidden, local, and data-dependent symmetries, where equivalences may depend on the activation region, parameter point, or data distribution rather than preserving the realized function globally (Zhao et al., 2023; Grigsby et al., 2023; Xie & Smidt, 2025; Zhao et al., 2026).

Entezari et al. (2022) connected parameter symmetries to LMC and stated the conjecture that pairs of SGD solutions are likely linearly mode connected after applying a suitable permutation. A number of works have proposed permutation alignment algorithms, with the goal of obtaining interpolating paths of low barrier, based either on optimal transport (OT) of activations or weights (*activation* or *weight matching*), or directly optimizing for a permutation that decreases the midpoint barrier (*straight-through estimator* or *learned matching*) (Singh & Jaggi, 2020; Tatro et al., 2020; Pittorino et al., 2022; Ainsworth et al., 2023), alongside non-OT variants (Li et al., 2015; Crisostomi et al., 2024). Alignment methods have been primarily applied to MLPs (permuting neurons) and convolutional architectures such as ResNets (permuting channels), but have recently been successfully applied to Transformers (Imfeld et al., 2024; Verma & Elbayad, 2024; Theus et al., 2025), both for vision and language (where vision task models seem to be easier to align). Theoretical results in Ferbach et al. (2024) show aligned LMC can be obtained for trained wide one-hidden-layer neural networks in the mean-field regime.

There has also been interest in *unaligned LMC*, i.e., conditions under which LMC is given without post-hoc alignment. Recently, Lim et al. (2024b) proposed architectural symmetry breaking methods and obtained unaligned LMC with $\boldsymbol{W}$-*asymmetric models*, where parts of the weights of each neuron are fixed in a way that breaks the usual permutation symmetries. Not primarily motivated by LMC but in a similar fashion, Ziyin et al. (2025) add fixed noise to parameters. Ito et al. (2026) find that increasingly wide models tend to yield LMC without symmetry alignment or removal, and explain this by wide layers exhibiting reciprocally orthogonal null spaces, so that merged models increasingly behave like an ensemble.

B.2. Fixed Masks, Pruning, and Sketching

When $\mathbf{d}_i \in \{0,1\}^d$, the trainable part $\mathbf{d}_i \odot \boldsymbol{w}_i$ corresponds to a fixed sparse connectivity pattern, which is similar to a pruning-at-initialization or fixed DropConnect-style mask on the incoming weights (Wan et al., 2013). In contrast, standard dropout resamples activation masks during training (Srivastava et al., 2014). Dropout and related feature-noising schemes have been analyzed as induced regularizers (Wager et al., 2013) and, under variational assumptions, as approximate Bayesian inference (Gal & Ghahramani, 2016). The masks in this work, however, are sampled once and kept fixed across the independent training runs.

Prior work on pruning has studied post-training sparsification, lottery-ticket subnetworks (Frankle & Carbin, 2019), dynamic sparse training (Evci et al., 2020), and pruning at initialization (Lee et al., 2019; Wang et al., 2020; Tanaka et al., 2020; Frankle et al., 2021). Recent work has also connected pruning at initialization to randomized sketching (Bar & Giryes, 2025). Our use of this viewpoint is different: we do not use masks primarily to preserve an approximation of a dense network, but to understand when fixed diagonal operators make neurons functionally distinguishable on the input support. The relevant quantity for this distinction is the subspace coherence $\nu(\mathcal{U})$, equivalently the maximum leverage score of the input subspace (Drineas et al., 2012). Coherence is standard in matrix completion, compressed sensing, and randomized numerical linear algebra, where it controls whether uniform coordinate sampling preserves a low-dimensional structure (Candes & Tao, 2005; Candes & Recht, 2009). In our setting, the same quantity controls the concentration of $\mathbf{S}_i = \boldsymbol{U}^\top \mathrm{Diag}(\mathbf{d}_i)^2 \boldsymbol{U}$. I.e., we use coherence not for exact recovery, but to quantify when fixed masks are visible to the effective function classes of neurons.

B.3. Representation Costs

In (4), we define the *realization cost* of a neuron or layer as the minimum weight norm required to implement it. This coincides with the concept of *representation cost* used in prior work (Dai et al., 2021; Gunasekar et al., 2018; Savarese et al., 2019). We call it *realization cost* to avoid confusion with *representations* in a layer.

C. Neuron Identifiability (§ 3)

Assumption C.1 (Subspace support model)**.** The distribution $\mathbb{P}$ is supported on a k-dimensional subspace $\mathcal{U} \subset \mathbb{R}^d$, i.e., $\mathbb{P}_{\boldsymbol{x}}(\boldsymbol{x} \in \mathcal{U}) = 1$. Throughout, we fix an orthonormal basis $\boldsymbol{U} \in \mathbb{R}^{d\times k}$ of $\mathcal{U}$, i.e., $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_k$ and $\mathrm{im}(\boldsymbol{U}) = \mathcal{U}$.

Assumption C.2 (Non-degeneracy)**.** Under Assumption C.1, additionally assume the following. If η is injective (e.g., $\tanh$, LeakyReLU), assume $\mathrm{span}(\mathrm{supp}(\mathbb{P})) = \mathcal{U}$. If $s \mapsto (\eta(s), \eta(-s))$ is injective (e.g., ReLU, GELU), assume there is $T \subseteq \mathrm{supp}(\mathbb{P})$ s.t. $T = -T$ and $\mathrm{span}(T) = \mathcal{U}$.

C.1. Realization Costs

Lemma C.3 (Neuron identifiability from preactivations)**.** *Under Assumptions C.1 and C.2, any $\boldsymbol{h}_i$, $i \in [m]$, can be identified from its preactivation. Formally,*

$$\boldsymbol{h}_i(\boldsymbol{w}; \cdot) = \boldsymbol{h}_i(\boldsymbol{w}'; \cdot) \;\Leftrightarrow\; \boldsymbol{U}^\top \mathrm{Diag}(\mathbf{d}_i)(\boldsymbol{w} - \boldsymbol{w}') = 0, \tag{17}$$

where the l.h.s. is understood almost surely w.r.t. $\mathbb{P}$.

Proof. Fix $i \in [m]$, $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$, and write the (row-wise) preactivation vectors

$$\boldsymbol{w}_{\text{eff}} \;:=\; \mathbf{f}_i + \mathbf{d}_i \odot \boldsymbol{w} \;=\; \mathbf{f}_i + \mathrm{Diag}(\mathbf{d}_i)\boldsymbol{w}, \qquad \boldsymbol{w}'_{\text{eff}} \;:=\; \mathbf{f}_i + \mathrm{Diag}(\mathbf{d}_i)\boldsymbol{w}'. \tag{18}$$

By Assumption C.1, every input $\boldsymbol{x}$ satisfies $\boldsymbol{x} \in \mathcal{U} = \mathrm{im}(\boldsymbol{U})$, hence we can write $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{c}$ for some $\boldsymbol{c} \in \mathbb{R}^k$. Define the subspace coefficients $\boldsymbol{a} \;:=\; \boldsymbol{U}^\top \boldsymbol{w}_{\text{eff}} \in \mathbb{R}^k$ and $\boldsymbol{a}' \;:=\; \boldsymbol{U}^\top \boldsymbol{w}'_{\text{eff}} \in \mathbb{R}^k$. For any $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{c} \in \mathcal{U}$, we have the identity

$$\boldsymbol{w}_{\text{eff}}^\top \boldsymbol{x} \;=\; \boldsymbol{w}_{\text{eff}}^\top \boldsymbol{U}\boldsymbol{c} \;=\; (\boldsymbol{U}^\top \boldsymbol{w}_{\text{eff}})^\top \boldsymbol{c} \;=\; \boldsymbol{a}^\top \boldsymbol{c}, \tag{19}$$

and likewise $(\boldsymbol{w}'_{\text{eff}})^\top \boldsymbol{x} = \boldsymbol{a}'^\top \boldsymbol{c}$. Therefore, the neuron functions restricted to $\mathcal{U}$ can be written as

$$\boldsymbol{h}_i(\boldsymbol{w}; \boldsymbol{U}\boldsymbol{c}) \;=\; \eta(\boldsymbol{a}^\top \boldsymbol{c}), \qquad \boldsymbol{h}_i(\boldsymbol{w}'; \boldsymbol{U}\boldsymbol{c}) \;=\; \eta(\boldsymbol{a}'^\top \boldsymbol{c}). \tag{20}$$

Note that

$$\boldsymbol{a} - \boldsymbol{a}' \;=\; \boldsymbol{U}^\top \big(\boldsymbol{w}_{\text{eff}} - \boldsymbol{w}'_{\text{eff}}\big) \;=\; \boldsymbol{U}^\top \mathrm{Diag}(\mathbf{d}_i)(\boldsymbol{w} - \boldsymbol{w}'). \tag{21}$$

We can now prove the equivalence.

"$\Leftarrow$": If $\boldsymbol{U}^\top \mathrm{Diag}(\mathbf{d}_i)(\boldsymbol{w} - \boldsymbol{w}') = 0$, then (21) and (20) directly yield $\boldsymbol{h}_i(\boldsymbol{w}; \cdot) = \boldsymbol{h}_i(\boldsymbol{w}'; \cdot)$ on $\mathcal{U}$.

"$\Rightarrow$": Assume $\boldsymbol{h}_i(\boldsymbol{w}; \cdot) = \boldsymbol{h}_i(\boldsymbol{w}'; \cdot)$ on $\mathrm{supp}(\mathbb{P}) \subseteq \mathcal{U}$ (by Assumption C.1). Equivalently,

$$\eta(\boldsymbol{a}^\top \boldsymbol{c}) \;=\; \eta(\boldsymbol{a}'^\top \boldsymbol{c}) \qquad \text{for all } \boldsymbol{c} \in \boldsymbol{U}^\top(\mathrm{supp}(\mathbb{P})) \subseteq \mathbb{R}^k. \tag{22}$$

We show this forces $\boldsymbol{a} = \boldsymbol{a}'$ under Assumption C.2, distinguishing the two activation cases.

Case ①: η injective. Injectivity gives $\boldsymbol{a}^\top \boldsymbol{c} = \boldsymbol{a}'^\top \boldsymbol{c}$ for all $\boldsymbol{c} \in \boldsymbol{U}^\top(\mathrm{supp}(\mathbb{P}))$, hence $(\boldsymbol{a} - \boldsymbol{a}')^\top \boldsymbol{c} = 0$ for all such $\boldsymbol{c}$. By Assumption C.2, we have $\mathrm{span}(\mathrm{supp}(\mathbb{P})) = \mathcal{U}$, hence $\mathrm{span}\big(\boldsymbol{U}^\top \mathrm{supp}(\mathbb{P})\big) = \mathbb{R}^k$. Therefore, $(\boldsymbol{a} - \boldsymbol{a}')^\top \boldsymbol{c} = 0$ on a spanning set of $\mathbb{R}^k$, which implies $\boldsymbol{a} - \boldsymbol{a}' = \boldsymbol{U}^\top \mathrm{Diag}(\mathbf{d}_i)(\boldsymbol{w} - \boldsymbol{w}') = 0$.

Case ②: $s \mapsto (\eta(s), \eta(-s))$ injective. Let $T \subseteq \mathrm{supp}(\mathbb{P})$ be as in Assumption C.2, so $T = -T$ and $\mathrm{span}(T) = \mathcal{U}$. Set $T_{\boldsymbol{U}} := \boldsymbol{U}^\top T \subseteq \mathbb{R}^k$. Then $T_{\boldsymbol{U}} = -T_{\boldsymbol{U}}$ and $\mathrm{span}(T_{\boldsymbol{U}}) = \mathbb{R}^k$. Since $\eta(\boldsymbol{a}^\top \boldsymbol{c}) = \eta(\boldsymbol{a}'^\top \boldsymbol{c})$ for all $\boldsymbol{c} \in \boldsymbol{U}^\top(\mathrm{supp}(\mathbb{P}))$, in particular $\eta(\boldsymbol{a}^\top \boldsymbol{c}) = \eta(\boldsymbol{a}'^\top \boldsymbol{c})$ for all $\boldsymbol{c} \in T_{\boldsymbol{U}}$. Moreover, for any $\boldsymbol{c} \in T_{\boldsymbol{U}}$ we have $-\boldsymbol{c} \in T_{\boldsymbol{U}}$, hence also $\eta(-\boldsymbol{a}^\top \boldsymbol{c}) = \eta(-\boldsymbol{a}'^\top \boldsymbol{c})$. Therefore, for all $\boldsymbol{c} \in T_{\boldsymbol{U}}$, $(\eta(\boldsymbol{a}^\top \boldsymbol{c}), \eta(-\boldsymbol{a}^\top \boldsymbol{c})) = (\eta(\boldsymbol{a}'^\top \boldsymbol{c}), \eta(-\boldsymbol{a}'^\top \boldsymbol{c}))$, and injectivity yields $\boldsymbol{a}^\top \boldsymbol{c} = \boldsymbol{a}'^\top \boldsymbol{c}$ for all $\boldsymbol{c} \in T_{\boldsymbol{U}}$, i.e. $(\boldsymbol{a} - \boldsymbol{a}')^\top \boldsymbol{c} = 0$ for all $\boldsymbol{c} \in T_{\boldsymbol{U}}$. Since $\mathrm{span}(T_{\boldsymbol{U}}) = \mathbb{R}^k$, we get $\boldsymbol{a} = \boldsymbol{a}'$, which concludes the proof. □

Proposition C.4. *Define $\mathbf{v}_i$ as the projection of the fixed center of the i-th asymmetric neuron to the input subspace, and $\mathbf{M}_i$ as its projected diagonal operator, i.e.,*

$$\mathbf{v}_i := \boldsymbol{U}^\top \mathbf{f}_i \in \mathbb{R}^k, \quad \mathbf{M}_i := \boldsymbol{U}^\top \mathrm{Diag}(\mathbf{d}_i) \in \mathbb{R}^{k\times d}. \tag{23}$$

Under Assumptions C.1 and C.2, $\boldsymbol{h} \in \mathcal{H}_i(\mathcal{X})$ belonging to weight $\boldsymbol{w}$ can be written as $\boldsymbol{h}(\boldsymbol{x}) = \eta((\boldsymbol{U}\boldsymbol{a})^\top \boldsymbol{x})$, where

$$\boldsymbol{a} \;:=\; \boldsymbol{U}^\top \big(\mathbf{f}_i + \mathrm{Diag}(\mathbf{d}_i)\boldsymbol{w}\big) \;=\; \mathbf{v}_i + \mathbf{M}_i \boldsymbol{w} \;\in \mathbb{R}^k. \tag{24}$$

The mapping $\boldsymbol{h} \mapsto \boldsymbol{a}$ is a bijection that identifies $\mathcal{H}_i(\mathcal{X})$ with $\mathbf{v}_i + \mathrm{im}(\mathbf{M}_i) \subseteq \mathbb{R}^k$. In particular, $\mathcal{H}_i(\mathcal{X})$ is an affine subspace of $\mathbb{R}^k$ of dimension $\mathrm{rk}(\mathbf{M}_i) \le k$.

Proof. Fix $i \in [m]$. Under Assumption C.1, every $\boldsymbol{x} \in \mathcal{X}$ satisfies $\boldsymbol{x} \in \mathrm{im}(\boldsymbol{U})$, so we can write $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{c}$ with $\boldsymbol{c} = \boldsymbol{U}^\top \boldsymbol{x} \in \mathbb{R}^k$. For any $\boldsymbol{w} \in \mathbb{R}^d$, define the induced subspace preactivation coefficient

$$\boldsymbol{a}(\boldsymbol{w}) \;:=\; \boldsymbol{U}^\top \big(\mathbf{f}_i + \mathrm{Diag}(\mathbf{d}_i)\boldsymbol{w}\big) \;=\; \mathbf{v}_i + \mathbf{M}_i \boldsymbol{w} \;\in \mathbb{R}^k. \tag{25}$$

Then for every $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{c} \in \mathcal{X}$, $\boldsymbol{h}_i(\boldsymbol{w}; \boldsymbol{x}) = \eta\big((\boldsymbol{U}\boldsymbol{a}(\boldsymbol{w}))^\top \boldsymbol{x}\big)$. This yields

$$\mathcal{H}_i(\mathcal{X}) \subseteq \Big\{\eta\big((\boldsymbol{U}\boldsymbol{a})^\top \cdot\,\big) \;\big|\; \boldsymbol{a} \in \mathbf{v}_i + \mathrm{im}(\mathbf{M}_i)\Big\}. \tag{26}$$

Conversely, let $\boldsymbol{a} \in \mathbf{v}_i + \mathrm{im}(\mathbf{M}_i)$. Then there exists $\boldsymbol{w} \in \mathbb{R}^d$ such that $\mathbf{M}_i\boldsymbol{w} = \boldsymbol{a} - \mathbf{v}_i$, i.e., $\boldsymbol{a}(\boldsymbol{w}) = \boldsymbol{a}$ and $\boldsymbol{h}_i(\boldsymbol{w}; \cdot) = \eta((\boldsymbol{U}\boldsymbol{a})^\top \cdot)$ on $\mathcal{X}$, so the reverse inclusion in (26) holds as well.

It remains to argue that $\boldsymbol{a} \mapsto \eta((\boldsymbol{U}\boldsymbol{a})^\top \cdot)$ is injective on $\mathbb{R}^k$ under Assumption C.2. Let $\boldsymbol{a}, \boldsymbol{a}' \in \mathbb{R}^k$ satisfy $\eta((\boldsymbol{U}\boldsymbol{a})^\top \boldsymbol{x}) = \eta((\boldsymbol{U}\boldsymbol{a}')^\top \boldsymbol{x})$ for all $\boldsymbol{x} \in \mathrm{supp}(\mathbb{P})$. Set $\boldsymbol{d} := \boldsymbol{a} - \boldsymbol{a}'$. If η is injective, then $(\boldsymbol{U}\boldsymbol{d})^\top \boldsymbol{x} = 0$ for all $\boldsymbol{x} \in \mathrm{supp}(\mathbb{P})$, hence $\boldsymbol{d}^\top(\boldsymbol{U}^\top \boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathrm{supp}(\mathbb{P})$. Since $\mathrm{span}(\mathrm{supp}(\mathbb{P})) = \mathcal{U} = \mathrm{im}(\boldsymbol{U})$, we obtain $\boldsymbol{d} = 0$, i.e. $\boldsymbol{a} = \boldsymbol{a}'$. If $s \mapsto (\eta(s), \eta(-s))$ is injective, then for all $\boldsymbol{x} \in T \subseteq \mathrm{supp}(\mathbb{P})$, with $T = -T$, we have both

$$\eta((\boldsymbol{U}\boldsymbol{a})^\top \boldsymbol{x}) = \eta((\boldsymbol{U}\boldsymbol{a}')^\top \boldsymbol{x}) \quad \text{and} \quad \eta((\boldsymbol{U}\boldsymbol{a})^\top(-\boldsymbol{x})) = \eta((\boldsymbol{U}\boldsymbol{a}')^\top(-\boldsymbol{x})), \tag{27}$$

so the injectivity of $s \mapsto (\eta(s), \eta(-s))$ yields $(\boldsymbol{U}\boldsymbol{a})^\top \boldsymbol{x} = (\boldsymbol{U}\boldsymbol{a}')^\top \boldsymbol{x}$ for all $\boldsymbol{x} \in T$. Since $\mathrm{span}(T) = \mathcal{U}$, the same argument gives $\boldsymbol{a} = \boldsymbol{a}'$.

Having obtained the injectivity of $\boldsymbol{a} \mapsto \eta((\boldsymbol{U}\boldsymbol{a})^\top \cdot)$, the final claim follows immediately. □

Lemma C.5 (Realization cost in subspace coefficients)**.** *Let $i \in [m]$ and $\boldsymbol{a} \in \mathbb{R}^k$. Then,*

$$\big\|\eta((\boldsymbol{U}\boldsymbol{a})^\top \cdot)\big\|_{\mathcal{H}_i(\mathcal{X})} = \sqrt{(\boldsymbol{a} - \mathbf{v}_i)^\top \mathbf{S}_i^\dagger (\boldsymbol{a} - \mathbf{v}_i)} \tag{28}$$

$$=: \|\boldsymbol{a} - \mathbf{v}_i\|_{\mathbf{S}_i^\dagger} \tag{29}$$

for $\boldsymbol{a} - \mathbf{v}_i \in \mathrm{im}(\mathbf{S}_i) = \mathrm{im}(\mathbf{M}_i)$ (and $+\infty$ otherwise). Here, $\mathbf{S}_i^\dagger$ denotes the Moore-Penrose pseudoinverse of $\mathbf{S}_i$.

Proof. Fix $i \in [m]$ and $\boldsymbol{a} \in \mathbb{R}^k$. By Prop. C.4, we have

$$\eta((\boldsymbol{U}\boldsymbol{a})^\top \cdot) \in \mathcal{H}_i(\mathcal{X}) \quad \Leftrightarrow \quad \boldsymbol{a} - \mathbf{v}_i \in \mathrm{im}(\mathbf{M}_i), \tag{30}$$

and in that case $\eta((\boldsymbol{U}\boldsymbol{a})^\top \cdot) = \boldsymbol{h}_i(\boldsymbol{w}; \cdot)$ holds iff $\mathbf{M}_i\boldsymbol{w} = \boldsymbol{a} - \mathbf{v}_i$. Therefore, by definition of realization cost (4),

$$\big\|\eta((\boldsymbol{U}\boldsymbol{a})^\top \cdot)\big\|_{\mathcal{H}_i(\mathcal{X})} = \inf\{\|\boldsymbol{w}\|_2 \mid \mathbf{M}_i\boldsymbol{w} = \boldsymbol{a} - \mathbf{v}_i\}, \tag{31}$$

with the convention $+\infty$ if $\boldsymbol{a} - \mathbf{v}_i \notin \mathrm{im}(\mathbf{M}_i)$. Assume now that $\boldsymbol{a} - \mathbf{v}_i \in \mathrm{im}(\mathbf{M}_i)$. Recalling $\mathbf{S}_i = \mathbf{M}_i\mathbf{M}_i^\top$, the (unique) minimum ℓ^2 norm solution is

$$\boldsymbol{w}^\star = \mathbf{M}_i^\top(\mathbf{M}_i\mathbf{M}_i^\top)^\dagger(\boldsymbol{a} - \mathbf{v}_i) = \mathbf{M}_i^\top \mathbf{S}_i^\dagger(\boldsymbol{a} - \mathbf{v}_i), \tag{32}$$

and every solution can be written as $\boldsymbol{w} = \boldsymbol{w}^\star + \boldsymbol{w}_\perp$ with $\boldsymbol{w}_\perp \in \ker(\mathbf{M}_i)$, which implies $\|\boldsymbol{w}\|_2^2 = \|\boldsymbol{w}^\star\|_2^2 + \|\boldsymbol{w}_\perp\|_2^2 \geq \|\boldsymbol{w}^\star\|_2^2$. Hence, $\inf\{\|\boldsymbol{w}\|_2 \mid \mathbf{M}_i\boldsymbol{w} = \boldsymbol{a} - \mathbf{v}_i\} = \|\boldsymbol{w}^\star\|_2$. Moreover,

$$\|\boldsymbol{w}^\star\|_2^2 = (\boldsymbol{a} - \mathbf{v}_i)^\top \mathbf{S}_i^\dagger \mathbf{S}_i \mathbf{S}_i^\dagger (\boldsymbol{a} - \mathbf{v}_i) \tag{33}$$

$$= (\boldsymbol{a} - \mathbf{v}_i)^\top \mathbf{S}_i^\dagger (\boldsymbol{a} - \mathbf{v}_i), \tag{34}$$

where we used $\mathbf{S}_i^\dagger \mathbf{S}_i \mathbf{S}_i^\dagger = \mathbf{S}_i^\dagger$. This yields

$$\big\|\eta((\boldsymbol{U}\boldsymbol{a})^\top \cdot)\big\|_{\mathcal{H}_i(\mathcal{X})} = \sqrt{(\boldsymbol{a} - \mathbf{v}_i)^\top \mathbf{S}_i^\dagger (\boldsymbol{a} - \mathbf{v}_i)}, \tag{35}$$

which holds for $\boldsymbol{a} - \mathbf{v}_i \in \mathrm{im}(\mathbf{S}_i) = \mathrm{im}(\mathbf{M}_i)$ (and equals $+\infty$ otherwise). □

Theorem C.6 (Spectral concentration of $\mathbf{S}_i$)**.** *Let $i \in [m]$ and sample the entries of $\mathbf{d}_i$ i.i.d. with $(\mathbf{d}_i)_\ell^2$ having mean $\mu_{\mathbf{D}}$, variance $\sigma_{\mathbf{D}}^2$, and a.s. deviation $b_{\mathbf{D}}$ from its mean (all $< \infty$). For any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\|\mathbf{S}_i - \mu_{\mathbf{D}}\boldsymbol{I}_k\|_{\mathrm{op}} \leq C\left\{\sqrt{\sigma_{\mathbf{D}}^2 \nu(\mathcal{U})\Lambda} + b_{\mathbf{D}}\nu(\mathcal{U})\Lambda\right\}, \tag{36}$$

where $\Lambda := \log(2k/\delta)$, $\nu(\mathcal{U}) := \max_{\ell \in [d]} (\boldsymbol{U}\boldsymbol{U}^\top)_{\ell\ell}$ is the subspace coherence, and C is a universal constant.[3]

The proof of this theorem relies on an application of the matrix Bernstein inequality.

Lemma C.7 (Matrix Bernstein Inequality; Theorem 1.4 of Tropp (2015))**.** *Let $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_d \in \mathbb{R}^{k \times k}$ be independent,*

[3] One can also control $\max_i \|\mathbf{S}_i - \mu_{\mathbf{D}}\boldsymbol{I}_k\|_{\mathrm{op}}$ via a union bound at the cost of replacing $\log(2k/\delta)$ by $\log(2mk/\delta)$.

centered random symmetric matrices. Assume a uniform bound $\|\boldsymbol{A}_l\|_{\text{op}} \leq R$ almost surely for each addend $l \in [d]$. Let

$$\sigma^2 := \left\| \sum_{l=1}^{d} \mathbb{E}[\boldsymbol{A}_l^2] \right\|_{\text{op}}. \tag{37}$$

Then, for all $t \geq 0$,

$$\mathbb{P}\left(\left\| \sum_{l=1}^{d} \boldsymbol{A}_l \right\|_{\text{op}} \geq t \right) \leq 2k \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3} \right). \tag{38}$$

Proof of Thm. C.6. Denote the l-th row of $\boldsymbol{U}$ by $\boldsymbol{u}_l \in \mathbb{R}^k$. Observe that

$$(\boldsymbol{U}\boldsymbol{U}^\top)_{ll} \;=\; \boldsymbol{e}_l^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{e}_l \;=\; \|\boldsymbol{U}^\top \boldsymbol{e}_l\|_2^2 \;=\; \|\boldsymbol{u}_l\|_2^2, \tag{39}$$

where $\boldsymbol{e}_l$ denotes the l-th canonical basis vector in $\mathbb{R}^d$. The fluctuation of the Gram matrix $\mathbf{S}_i$ around its mean

$$\mathbb{E}[\mathbf{S}_i] \;=\; \boldsymbol{U}^\top \mathbb{E}[\text{diag}(\mathbf{d}_i)^2]\boldsymbol{U} \;=\; \boldsymbol{U}^\top(\mu_{\mathbf{D}}\boldsymbol{I}_d)\boldsymbol{U} \;=\; \mu_{\mathbf{D}}\boldsymbol{I}_k \tag{40}$$

can be written as a sum of independent rank 1 matrices

$$\mathbf{S}_i - \mu_{\mathbf{D}}\boldsymbol{I}_k \;=\; \sum_{l=1}^{d} \underbrace{((\mathbf{d}_i)_l^2 - \mu_{\mathbf{D}})\boldsymbol{u}_l\boldsymbol{u}_l^\top}_{=:\boldsymbol{A}_l} \;=\; \sum_{l=1}^{d} \boldsymbol{A}_l. \tag{41}$$

Clearly, the addends $\boldsymbol{A}_l$ are centered and symmetric. Furthermore, we can bound

$$\|\boldsymbol{A}_l\|_{\text{op}} \;=\; |(\mathbf{d}_i)_l^2 - \mu_{\mathbf{D}}| \cdot \|\boldsymbol{u}_l\boldsymbol{u}_l^\top\|_{\text{op}} \;\leq\; b_{\mathbf{D}}\|\boldsymbol{u}_l\|_2^2 \;\leq\; b_{\mathbf{D}}\nu(\mathcal{U}) \;=:\; R. \tag{42}$$

Next, we compute the variance parameter σ^2. Since

$$\boldsymbol{A}_l^2 = ((\mathbf{d}_i)_l^2 - \mu_{\mathbf{D}})^2(\boldsymbol{u}_l\boldsymbol{u}_l^\top)^2 = ((\mathbf{d}_i)_l^2 - \mu_{\mathbf{D}})^2\|\boldsymbol{u}_l\|_2^2(\boldsymbol{u}_l\boldsymbol{u}_l^\top) \tag{43}$$

and $\mathbb{E}[((\mathbf{d}_i)_l^2 - \mu_{\mathbf{D}})^2] = \sigma_{\mathbf{D}}^2$ by definition, we have

$$\begin{aligned}
\sigma^2 \;&=\; \left\| \sum_{l=1}^{d} \sigma_{\mathbf{D}}^2 \|\boldsymbol{u}_l\|_2^2 (\boldsymbol{u}_l\boldsymbol{u}_l^\top) \right\|_{\text{op}} && (44)\\
&\overset{(*)}{\leq}\; \sigma_{\mathbf{D}}^2 \underbrace{\left(\max_k \|\boldsymbol{u}_k\|_2^2 \right)}_{=\nu(\mathcal{U})} \left\| \sum_{l=1}^{d} \boldsymbol{u}_l\boldsymbol{u}_l^\top \right\|_{\text{op}} && (45)\\
&=\; \sigma_{\mathbf{D}}^2 \nu(\mathcal{U}) \| \underbrace{\boldsymbol{U}^\top\boldsymbol{U}}_{=\boldsymbol{I}_k} \|_{\text{op}} && (46)\\
&=\; \sigma_{\mathbf{D}}^2 \nu(\mathcal{U}). && (47)
\end{aligned}$$

In $(*)$, we have utilized that for positive semidefinite matrices $\boldsymbol{B}_l \succeq 0$ and scalars $c_l \in [0, C]$, one has $\|\sum_l c_l \boldsymbol{B}_l\|_{\text{op}} \leq C\|\sum_l \boldsymbol{B}_l\|_{\text{op}}$. We are now ready to apply Lem. C.7 to $\sum_{l=1}^{d} \boldsymbol{A}_l$. We set the target failure probability to $\delta \in (0, 1)$. We seek a threshold t such that the failure probability is at most δ. This holds if

$$2k \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3} \right) \;\leq\; \delta \quad \Leftrightarrow \quad \frac{t^2}{2} \;\geq\; \left(\sigma^2 + \frac{Rt}{3} \right) \log\left(\frac{2k}{\delta} \right). \tag{48}$$

Let $L := \log(2k/\delta)$. The inequality becomes $t^2 - \frac{2}{3}RLt - 2\sigma^2 L \geq 0$. Solving for t yields

$$t \;\geq\; \frac{RL}{3} + \sqrt{\left(\frac{RL}{3}\right)^2 + 2\sigma^2 L} \;\overset{(**)}{\leq}\; \frac{2}{3}RL + \sigma\sqrt{2L}, \tag{49}$$

where $(**)$ uses $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$ (setting t to this upper bound suffices). Finally, we can plug in $R = b_{\mathbf{D}}\nu(\mathcal{U})$ ((42)) and $\sigma^2 \leq \sigma_{\mathbf{D}}^2\nu(\mathcal{U})$ ((47)), and absorbing absolute constants into C, we obtain

$$t \;\geq\; C\left\{ \sqrt{\sigma_{\mathbf{D}}^2 \nu(\mathcal{U}) \log(2k/\delta)} + b_{\mathbf{D}}\nu(\mathcal{U}) \log(2k/\delta) \right\}, \tag{50}$$

which completes the proof. □

Lemma C.8. *Let $\boldsymbol{H} \in \mathcal{H}(\mathcal{X})$ with $\boldsymbol{h}_i = \eta((\boldsymbol{U}\boldsymbol{a}_i)^\top \cdot)$ and $\boldsymbol{r}_i := \boldsymbol{a}_i - \mathbf{v}_i$. Then,*

$$\Delta^{\text{out}}_{(ij)}(\boldsymbol{H}) \;=\; \underbrace{\|\mathbf{v}_j - \mathbf{v}_i\|^2_{\mathbf{S}_i^{-1}+\mathbf{S}_j^{-1}}}_{(i)\text{ center term}} + \underbrace{2\langle \mathbf{v}_j - \mathbf{v}_i, \mathbf{S}_i^{-1}\boldsymbol{r}_j - \mathbf{S}_j^{-1}\boldsymbol{r}_i\rangle}_{(ii)\text{ cross term}} + \underbrace{\|\boldsymbol{r}_j\|^2_{\mathbf{S}_i^{-1}-\mathbf{S}_j^{-1}} + \|\boldsymbol{r}_i\|^2_{\mathbf{S}_j^{-1}-\mathbf{S}_i^{-1}}}_{(iii)\text{ metric mismatch term}}. \tag{51}$$

Proof. Fix $i \neq j$ and write $\mathbf{v}_{ij} := \mathbf{v}_j - \mathbf{v}_i$. For a transposition (ij), only the i- and j-addends change, hence

$$\begin{aligned}\Delta^{\text{out}}_{(ij)}(\boldsymbol{H}) \;&=\; \|\mathbf{v}_j - \mathbf{v}_i + \boldsymbol{r}_j\|^2_{\mathbf{S}_i^{-1}} + \|\mathbf{v}_i - \mathbf{v}_j + \boldsymbol{r}_i\|^2_{\mathbf{S}_j^{-1}} - \|\boldsymbol{r}_i\|^2_{\mathbf{S}_i^{-1}} - \|\boldsymbol{r}_j\|^2_{\mathbf{S}_j^{-1}} \\ &=\; \|\mathbf{v}_{ij} + \boldsymbol{r}_j\|^2_{\mathbf{S}_i^{-1}} + \|-\mathbf{v}_{ij} + \boldsymbol{r}_i\|^2_{\mathbf{S}_j^{-1}} - \|\boldsymbol{r}_i\|^2_{\mathbf{S}_i^{-1}} - \|\boldsymbol{r}_j\|^2_{\mathbf{S}_j^{-1}}.\end{aligned} \tag{52}$$

Using $\|\boldsymbol{x}\|^2_{\boldsymbol{A}} = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ and expanding the quadratic forms in (52) yields

$$\|\mathbf{v}_{ij} + \boldsymbol{r}_j\|^2_{\mathbf{S}_i^{-1}} \;=\; \mathbf{v}_{ij}^\top \mathbf{S}_i^{-1}\mathbf{v}_{ij} + 2\,\mathbf{v}_{ij}^\top \mathbf{S}_i^{-1}\boldsymbol{r}_j + \boldsymbol{r}_j^\top \mathbf{S}_i^{-1}\boldsymbol{r}_j, \tag{53}$$

$$\|-\mathbf{v}_{ij} + \boldsymbol{r}_i\|^2_{\mathbf{S}_j^{-1}} \;=\; \mathbf{v}_{ij}^\top \mathbf{S}_j^{-1}\mathbf{v}_{ij} - 2\,\mathbf{v}_{ij}^\top \mathbf{S}_j^{-1}\boldsymbol{r}_i + \boldsymbol{r}_i^\top \mathbf{S}_j^{-1}\boldsymbol{r}_i. \tag{54}$$

Substituting (53) and (54) into (52) and regrouping gives

$$\begin{aligned}\Delta^{\text{out}}_{(ij)}(\boldsymbol{H}) = \mathbf{v}_{ij}^\top(\mathbf{S}_i^{-1} + \mathbf{S}_j^{-1})\mathbf{v}_{ij} + 2\,\mathbf{v}_{ij}^\top(\mathbf{S}_i^{-1}\boldsymbol{r}_j - \mathbf{S}_j^{-1}\boldsymbol{r}_i) \\ + \boldsymbol{r}_j^\top(\mathbf{S}_i^{-1} - \mathbf{S}_j^{-1})\boldsymbol{r}_j + \boldsymbol{r}_i^\top(\mathbf{S}_j^{-1} - \mathbf{S}_i^{-1})\boldsymbol{r}_i,\end{aligned} \tag{55}$$

which is precisely (51). □

C.2. Neuron Reassignments

Assumption C.9 (Spectral concentration of Gram anisotropies)**.** Fix $\varepsilon \in (0, \frac{\mu_{\mathbf{D}}}{2})$, and assume

$$\max_{i\in[m]} \|\mathbf{S}_i - \mu_{\mathbf{D}}\boldsymbol{I}_k\|_{\text{op}} \;\leq\; \varepsilon. \tag{56}$$

Theorem C.10 (Global output sensitivity)**.** *Consider a layer $\boldsymbol{H} = \eta\big((\mathbf{F} + \mathbf{D}\odot\boldsymbol{W})\,\cdot\big)$, and assume Assumption C.9 holds. Set $\boldsymbol{R} := (\mathbf{D}\odot\boldsymbol{W})\boldsymbol{U} \in \mathbb{R}^{m\times k}$ (proj. trainable parts). Let $\pi \in S_m$ and set $\boldsymbol{M}_\pi := \boldsymbol{I}_m - \mathrm{Diag}\big(\mathrm{diag}(\boldsymbol{P}_\pi)\big)$ to be the projection onto the coordinates affected by π. Assume that $\boldsymbol{\delta}^{\text{out}}_\pi \neq 0$, $\rho^{\text{out}}_\pi := \|\boldsymbol{M}_\pi\boldsymbol{R}\|_F / \|\boldsymbol{\delta}^{\text{out}}_\pi\|_F \leq 1$. Then,*

$$\Delta^{\text{out}}_\pi(\boldsymbol{H}) \;=\; \Big(1 \pm \mathcal{O}(\rho^{\text{out}}_\pi + \varepsilon\mu_{\mathbf{D}}^{-1})\Big)\cdot \mu_{\mathbf{D}}^{-1}\|\boldsymbol{\delta}^{\text{out}}_\pi\|^2_F. \tag{57}$$

Proof. First, note that by definition,

$$\Delta^{\text{out}}_\pi(\boldsymbol{H}, \mathcal{X}) \;=\; \sum_{i=1}^m \Big(\big\|(\boldsymbol{\delta}^{\text{out}}_\pi)_{i,:} + (\boldsymbol{P}_\pi\boldsymbol{R})_{i,:}\big\|^2_{\mathbf{S}_i^{-1}} - \|\boldsymbol{R}_{i,:}\|^2_{\mathbf{S}_i^{-1}}\Big). \tag{58}$$

Now if $\pi(i) = i$, then $(\boldsymbol{\delta}^{\text{out}}_\pi)_{i,:} = 0$ and $(\boldsymbol{P}_\pi\boldsymbol{R})_{i,:} = \boldsymbol{R}_{i,:}$, so the i-th summand in (58) vanishes. Since $\boldsymbol{M}_\pi$ is the (diagonal) projection onto the rows moved by π, we may write

$$\Delta^{\text{out}}_\pi(\boldsymbol{H}, \mathcal{X}) \;=\; \sum_{i=1}^m \left(\Big\|\big(\boldsymbol{M}_\pi(\boldsymbol{\delta}^{\text{out}}_\pi + \boldsymbol{P}_\pi\boldsymbol{R})\big)_{i,:}\Big\|^2_{\mathbf{S}_i^{-1}} - \|(\boldsymbol{M}_\pi\boldsymbol{R})_{i,:}\|^2_{\mathbf{S}_i^{-1}}\right). \tag{59}$$

Since $\varepsilon < \mu_{\mathbf{D}}$, the uniform bound $\|\mathbf{S}_i - \mu_{\mathbf{D}}\boldsymbol{I}_k\|_{\text{op}} \leq \varepsilon$ in i yields $(\mu_{\mathbf{D}} - \varepsilon)\boldsymbol{I}_k \preceq \mathbf{S}_i \preceq (\mu_{\mathbf{D}} + \varepsilon)\boldsymbol{I}_k$. Hence $\mathbf{S}_i$ is invertible, and

$$\big\|\mathbf{S}_i^{-1} - \mu_{\mathbf{D}}^{-1}\boldsymbol{I}_k\big\|_{\text{op}} \;\leq\; \frac{\varepsilon}{\mu_{\mathbf{D}}(\mu_{\mathbf{D}} - \varepsilon)} \;=:\; \beta. \tag{60}$$

Thus, for every $\boldsymbol{a} \in \mathbb{R}^k$,

$$\Big|\|\boldsymbol{a}\|^2_{\mathbf{S}_i^{-1}} - \mu_{\mathbf{D}}^{-1}\|\boldsymbol{a}\|^2_2\Big| \;\leq\; \beta\|\boldsymbol{a}\|^2_2. \tag{61}$$

Applying (61) to each term in (59) yields

$$\Delta^{\text{out}}_\pi(\boldsymbol{H}, \mathcal{X}) \;=\; \mu_{\mathbf{D}}^{-1}\left(\|\boldsymbol{M}_\pi(\boldsymbol{\delta}^{\text{out}}_\pi + \boldsymbol{P}_\pi\boldsymbol{R})\|^2_F - \|\boldsymbol{M}_\pi\boldsymbol{R}\|^2_F\right) + E_\pi, \tag{62}$$

with

$$|E_\pi| \;\leq\; \beta\left(\|\boldsymbol{M}_\pi(\boldsymbol{\delta}^{\text{out}}_\pi + \boldsymbol{P}_\pi\boldsymbol{R})\|^2_F + \|\boldsymbol{M}_\pi\boldsymbol{R}\|^2_F\right). \tag{63}$$

Because $\boldsymbol{M}_\pi\boldsymbol{\delta}^{\text{out}}_\pi = \boldsymbol{\delta}^{\text{out}}_\pi$ and $\boldsymbol{M}_\pi\boldsymbol{P}_\pi = \boldsymbol{P}_\pi\boldsymbol{M}_\pi$ implies $\|\boldsymbol{M}_\pi\boldsymbol{P}_\pi\boldsymbol{R}\|_F = \|\boldsymbol{P}_\pi\boldsymbol{M}_\pi\boldsymbol{R}\|_F = \|\boldsymbol{M}_\pi\boldsymbol{R}\|_F$, we can apply Cauchy-Schwarz to bound (63) as

$$|E_\pi| \;\leq\; \beta\Big(\|\boldsymbol{\delta}^{\text{out}}_\pi\|^2_F + 2\|\boldsymbol{\delta}^{\text{out}}_\pi\|_F\|\boldsymbol{M}_\pi\boldsymbol{R}\|_F + 2\|\boldsymbol{M}_\pi\boldsymbol{R}\|^2_F\Big). \tag{64}$$

It remains to simplify the isotropic term in (62), for which one can obtain

$$\|\boldsymbol{M}_\pi(\boldsymbol{\delta}_\pi^{\text{out}} + \boldsymbol{P}_\pi \boldsymbol{R})\|_F^2 - \|\boldsymbol{M}_\pi \boldsymbol{R}\|_F^2 \;=\; \|\boldsymbol{\delta}_\pi^{\text{out}}\|_F^2 + 2\langle \boldsymbol{\delta}_\pi^{\text{out}}, \boldsymbol{M}_\pi \boldsymbol{P}_\pi \boldsymbol{R}\rangle_F = \|\boldsymbol{\delta}_\pi^{\text{out}}\|_F^2 + 2\langle \boldsymbol{\delta}_\pi^{\text{out}}, \boldsymbol{P}_\pi \boldsymbol{R}\rangle_F, \tag{65}$$

where the last equality holds since $\boldsymbol{M}_\pi^\top = \boldsymbol{M}_\pi$ and $\boldsymbol{M}_\pi \boldsymbol{\delta}_\pi^{\text{out}} = \boldsymbol{\delta}_\pi^{\text{out}}$. Define

$$\boldsymbol{Q}_\pi \;:=\; \boldsymbol{I}_m - \text{ord}(\pi)^{-1} \sum_{t=0}^{\text{ord}(\pi)-1} \boldsymbol{P}_\pi^t. \tag{66}$$

Since $\boldsymbol{P}_\pi^{\text{ord}(\pi)} = \boldsymbol{I}_m$,

$$(\boldsymbol{I}_m - \boldsymbol{Q}_\pi)\boldsymbol{P}_\pi \;=\; \boldsymbol{P}_\pi(\boldsymbol{I}_m - \boldsymbol{Q}_\pi) \;=\; \boldsymbol{I}_m - \boldsymbol{Q}_\pi. \tag{67}$$

Thus, $\boldsymbol{I}_m - \boldsymbol{Q}_\pi$ is the orthogonal projection onto the fixed subspace of $\boldsymbol{P}_\pi$, and $\boldsymbol{Q}_\pi$ is the orthogonal projection onto its orthogonal complement. Moreover,

$$(\boldsymbol{I}_m - \boldsymbol{Q}_\pi)\boldsymbol{\delta}_\pi^{\text{out}} \;=\; (\boldsymbol{I}_m - \boldsymbol{Q}_\pi)(\boldsymbol{P}_\pi - \boldsymbol{I}_m)\mathbf{F}\boldsymbol{U} \;=\; 0. \tag{68}$$

Hence, $\boldsymbol{Q}_\pi \boldsymbol{\delta}_\pi^{\text{out}} = \boldsymbol{\delta}_\pi^{\text{out}}$ and, using $\boldsymbol{P}_\pi(\boldsymbol{I}_m - \boldsymbol{Q}_\pi) = \boldsymbol{I}_m - \boldsymbol{Q}_\pi$,

$$\left\langle \boldsymbol{\delta}_\pi^{\text{out}}, \boldsymbol{P}_\pi(\boldsymbol{I}_m - \boldsymbol{Q}_\pi)\boldsymbol{R}\right\rangle_F \;=\; \left\langle \boldsymbol{\delta}_\pi^{\text{out}}, (\boldsymbol{I}_m - \boldsymbol{Q}_\pi)\boldsymbol{R}\right\rangle_F \;=\; 0. \tag{69}$$

Therefore,

$$\left\langle \boldsymbol{\delta}_\pi^{\text{out}}, \boldsymbol{P}_\pi \boldsymbol{R}\right\rangle_F \;=\; \left\langle \boldsymbol{\delta}_\pi^{\text{out}}, \boldsymbol{P}_\pi \boldsymbol{Q}_\pi \boldsymbol{R}\right\rangle_F. \tag{70}$$

By Cauchy–Schwarz and the orthogonality of $\boldsymbol{P}_\pi$,

$$\left|\left\langle \boldsymbol{\delta}_\pi^{\text{out}}, \boldsymbol{P}_\pi \boldsymbol{R}\right\rangle_F\right| \;\le\; \|\boldsymbol{\delta}_\pi^{\text{out}}\|_F \|\boldsymbol{Q}_\pi \boldsymbol{R}\|_F. \tag{71}$$

Combining (62), (64), (65), and (71) gives

$$\begin{aligned}\left|\Delta_\pi^{\text{out}}(\boldsymbol{H}, \mathcal{X}) - \mu_{\mathbf{D}}^{-1}\|\boldsymbol{\delta}_\pi^{\text{out}}\|_F^2\right| \;\le\; & \frac{2}{\mu_{\mathbf{D}}} \|\boldsymbol{\delta}_\pi^{\text{out}}\|_F \|\boldsymbol{Q}_\pi \boldsymbol{R}\|_F \\ & + \frac{\varepsilon}{\mu_{\mathbf{D}}(\mu_{\mathbf{D}} - \varepsilon)}\Big(\|\boldsymbol{\delta}_\pi^{\text{out}}\|_F^2 + 2\|\boldsymbol{\delta}_\pi^{\text{out}}\|_F \|\boldsymbol{M}_\pi \boldsymbol{R}\|_F + 2\|\boldsymbol{M}_\pi \boldsymbol{R}\|_F^2\Big).\end{aligned} \tag{72}$$

For the final statement, note that $\boldsymbol{Q}_\pi = \boldsymbol{Q}_\pi \boldsymbol{M}_\pi$, and as $\boldsymbol{Q}_\pi$ is an orthogonal projection,

$$\|\boldsymbol{Q}_\pi \boldsymbol{R}\|_F \;=\; \|\boldsymbol{Q}_\pi \boldsymbol{M}_\pi \boldsymbol{R}\|_F \;\le\; \|\boldsymbol{M}_\pi \boldsymbol{R}\|_F. \tag{73}$$

Dividing (72) by $\mu_{\mathbf{D}}^{-1}\|\boldsymbol{\delta}_\pi^{\text{out}}\|_F^2$ therefore yields

$$\left|\frac{\mu_{\mathbf{D}} \Delta_\pi^{\text{out}}(\boldsymbol{H}, \mathcal{X})}{\|\boldsymbol{\delta}_\pi^{\text{out}}\|_F^2} - 1\right| \;\le\; 2\rho_\pi^{\text{out}} + \frac{\varepsilon}{\mu_{\mathbf{D}} - \varepsilon}\big(1 + 2\rho_\pi^{\text{out}} + 2(\rho_\pi^{\text{out}})^2\big). \tag{74}$$

If $\rho_\pi^{\text{out}} \le 1$ and $\varepsilon \le \mu_{\mathbf{D}}/2$, then

$$\frac{\varepsilon}{\mu_{\mathbf{D}} - \varepsilon} \;\le\; \frac{2\varepsilon}{\mu_{\mathbf{D}}}, \qquad 1 + 2\rho_\pi^{\text{out}} + 2(\rho_\pi^{\text{out}})^2 \;\le\; 5. \tag{75}$$

Hence, the r.h.s. is $\mathcal{O}(\rho_\pi^{\text{out}} + \varepsilon\mu_{\mathbf{D}}^{-1})$, and therefore

$$\Delta_\pi^{\text{out}}(\boldsymbol{H}, \mathcal{X}) \;=\; \Big(1 \pm \mathcal{O}(\rho_\pi^{\text{out}} + \varepsilon\mu_{\mathbf{D}}^{-1})\Big) \cdot \mu_{\mathbf{D}}^{-1}\|\boldsymbol{\delta}_\pi^{\text{out}}\|_F^2. \tag{76}$$

□

Corollary C.11 (Minimum output reassignment cost)**.** *In the setup of Thm. C.10, suppose further that $\rho_{(ij)}^{\text{out}} \le \rho^{\text{out}} \le 1$ for some ρ^{out}, uniformly in $i \ne j \in [m]$. Then,*

$$\min_{\text{id} \ne \pi \in S_m} \Delta_\pi^{\text{out}}(\boldsymbol{H}, \mathcal{X}) \;=\; \Big(1 \pm \mathcal{O}(\rho^{\text{out}} + \varepsilon\mu_{\mathbf{D}}^{-1})\Big) \cdot \mu_{\mathbf{D}}^{-1}\gamma_{\text{out}}^2. \tag{77}$$

Proof. Fix $\text{id} \ne \pi \in S_m$. Then,

$$\|\boldsymbol{\delta}_\pi^{\text{out}}\|_F^2 \;=\; \|(\boldsymbol{P}_\pi - \boldsymbol{I}_m)\mathbf{F}\boldsymbol{U}\|_F^2 \;=\; \sum_{i=1}^m \|\mathbf{v}_{\pi(i)} - \mathbf{v}_i\|_2^2. \tag{78}$$

Write π as a product of disjoint cycles, and consider one nontrivial cycle $(i_1 \dots i_\ell)$ of length $\ell \ge 2$, and use the convention $i_{\ell+1} := i_1$. Its contribution to the above sum is

$$\sum_{r=1}^{\ell} \|\mathbf{v}_{i_{r+1}} - \mathbf{v}_{i_r}\|_2^2 \;\ge\; \ell \min_{i \ne j} \|\mathbf{v}_j - \mathbf{v}_i\|_2^2 \;\ge\; 2 \min_{i \ne j} \|\mathbf{v}_j - \mathbf{v}_i\|_2^2. \tag{79}$$

On the other hand, for a transposition (ij) one has $\|\boldsymbol{\delta}^{\text{out}}_{(ij)}\|_F^2 \;=\; 2\|\mathbf{v}_j - \mathbf{v}_i\|_2^2$. Therefore,

$$\min_{\text{id}\neq\pi\in S_m} \|\boldsymbol{\delta}^{\text{out}}_{\pi}\|_F \;=\; \min_{i\neq j} \|\boldsymbol{\delta}^{\text{out}}_{(ij)}\|_F \;=\; \gamma_{\text{out}}. \tag{80}$$

Next, for every $i \neq j$,

$$\|\boldsymbol{R}_{i,:}\|_2^2 + \|\boldsymbol{R}_{j,:}\|_2^2 \;=\; \|\boldsymbol{M}_{(ij)}\boldsymbol{R}\|_F^2 \;=\; (\rho^{\text{out}}_{(ij)})^2\|\boldsymbol{\delta}^{\text{out}}_{(ij)}\|_F^2 \;\leq\; 2(\rho^{\text{out}})^2\|\mathbf{v}_j - \mathbf{v}_i\|_F^2. \tag{81}$$

Now let $\pi \neq \text{id}$ and decompose its moved indices into s nontrivial disjoint cycles $(i_1^{(s)} \dots i_{\ell_s}^{(s)})$. Then,

$$2\|\boldsymbol{M}_\pi\boldsymbol{R}\|_F^2 \;=\; \sum_s \sum_{r=1}^{\ell_s}\Big(\|\boldsymbol{R}_{i_r^{(s)},:}\|_2^2 + \|\boldsymbol{R}_{i_{r+1}^{(s)},:}\|_2^2\Big) \;\leq\; 2(\rho^{\text{out}})^2 \sum_s \sum_{r=1}^{\ell_s} \|\mathbf{v}_{i_{r+1}^{(s)}} - \mathbf{v}_{i_r^{(s)}}\|_2^2 \;=\; 2(\rho^{\text{out}})^2\|\boldsymbol{\delta}^{\text{out}}_\pi\|_F^2. \tag{82}$$

Hence, $\rho^{\text{out}}_\pi = \|\boldsymbol{M}_\pi\boldsymbol{R}\|_F/\|\boldsymbol{\delta}^{\text{out}}_\pi\|_F \leq \rho^{\text{out}}$ for all $\pi \neq \text{id}$. By Theorem C.10, it follows that

$$\Delta^{\text{out}}_\pi(\boldsymbol{H}, \mathcal{X}) \;=\; \Big(1 \pm \mathcal{O}(\rho^{\text{out}} + \varepsilon\mu_{\mathbf{D}}^{-1})\Big) \cdot \mu_{\mathbf{D}}^{-1}\|\boldsymbol{\delta}^{\text{out}}_\pi\|_F^2 \tag{83}$$

uniformly in $\pi \neq \text{id}$. Using (80) yields

$$\min_{\text{id}\neq\pi\in S_m} \Delta^{\text{out}}_\pi(\boldsymbol{H}, \mathcal{X}) \;=\; \Big(1 \pm \mathcal{O}(\rho^{\text{out}} + \varepsilon\mu_{\mathbf{D}}^{-1})\Big) \cdot \mu_{\mathbf{D}}^{-1}\gamma_{\text{out}}^2. \tag{84}$$

□

Theorem C.12 (Asymptotics of projected center gap γ_{out})**.** *Let $\mathbf{F}$ have i.i.d. entries of the form $B \cdot \sigma_{\mathbf{F}} Y_0$, where $B \sim$ Bernoulli(p), $p \in (0, 1]$, and Y_0 is independent of B with a Lebesgue density bounded by $M_0 < \infty$. Set $q := 2p - p^2$ and fix $\delta \in (0, 1)$. Let $m \geq 2$ and assume*

$$q \geq C\big\{\sqrt{q\,\nu(\mathcal{U})\Lambda} + \nu(\mathcal{U})\Lambda\big\}, \quad \Lambda := \log(4km^2/\delta). \tag{85}$$

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

$$\gamma_{\text{out}} \geq c\,\frac{\sqrt{q}}{M_0}\sigma_{\mathbf{F}} V_k^{-1/k}\Big(\frac{\delta}{m^2}\Big)^{1/k} = \Theta(\sigma_{\mathbf{F}}\sqrt{k}m^{-2/k}), \tag{86}$$

where V_k denotes the volume of the k-dim. unit ball, $V_k^{-1/k} = \Theta(\sqrt{k})$, and $c, C > 0$ are absolute constants.

Proof. Recall that $\mathbf{f}_i \in \mathbb{R}^d$ is the i-th row of $\mathbf{F}$ and $\mathbf{v}_i = \boldsymbol{U}^\top\mathbf{f}_i \in \mathbb{R}^k$, further $\gamma_{\text{out}} = \sqrt{2}\min_{i\neq j}\|\mathbf{v}_i - \mathbf{v}_j\|_2$ and $\mathbf{v}_{ij} = \mathbf{v}_i - \mathbf{v}_j = \boldsymbol{U}^\top(\mathbf{f}_i - \mathbf{f}_j) \in \mathbb{R}^k$. Moreover, define $Y := \sigma_{\mathbf{F}} Y_0$, which has a Lebesgue density bounded by $M := \sigma_{\mathbf{F}}^{-1} M_0$.

Step ①: Random mask concentration. Fix a pair (i, j). For each $\ell \in [d]$, let

$$\xi_\ell \;:=\; \mathbb{1}\{(B_{i\ell}, B_{j\ell}) \neq (0, 0)\} \in \{0, 1\}, \qquad \boldsymbol{P} \;:=\; \boldsymbol{U}^\top\text{Diag}(\xi)\boldsymbol{U} \in \mathbb{R}^{k\times k}. \tag{87}$$

Then, ξ_ℓ are i.i.d. Bernoulli variables with

$$\mathbb{E}[\xi_\ell] \;=\; \mathbb{P}(\xi_\ell = 1) \;=\; 1 - (1-p)^2 \;=\; 2p - p^2 \;=:\; q. \tag{88}$$

Let $\boldsymbol{u}_\ell := \boldsymbol{U}^\top\boldsymbol{e}_\ell \in \mathbb{R}^k$, so that $\|\boldsymbol{u}_\ell\|_2^2 = (\boldsymbol{U}\boldsymbol{U}^\top)_{\ell\ell} \leq \nu(\mathcal{U})$. Then,

$$\boldsymbol{P} - q\boldsymbol{I}_k \;=\; \sum_{\ell=1}^d (\xi_\ell - q)\,\boldsymbol{u}_\ell\boldsymbol{u}_\ell^\top. \tag{89}$$

Set $\boldsymbol{A}_\ell := (\xi_\ell - q)\,\boldsymbol{u}_\ell\boldsymbol{u}_\ell^\top$. Then, $\mathbb{E}[\boldsymbol{A}_\ell] = 0$ and

$$\|\boldsymbol{A}_\ell\|_{\text{op}} \;=\; |\xi_\ell - q|\,\|\boldsymbol{u}_\ell\|_2^2 \;\leq\; \nu(\mathcal{U}) \;=:\; R. \tag{90}$$

Further,

$$\boldsymbol{A}_\ell^2 \;=\; (\xi_\ell - q)^2(\boldsymbol{u}_\ell\boldsymbol{u}_\ell^\top)^2 \;=\; (\xi_\ell - q)^2\|\boldsymbol{u}_\ell\|_2^2\,\boldsymbol{u}_\ell\boldsymbol{u}_\ell^\top, \tag{91}$$

so, since $\mathbb{E}[(\xi_\ell - q)^2] = q(1-q) \leq q$,

$$\mathbb{E}[\boldsymbol{A}_\ell^2] \;\preceq\; q\,\|\boldsymbol{u}_\ell\|_2^2\,\boldsymbol{u}_\ell\boldsymbol{u}_\ell^\top \;\preceq\; q\,\nu(\mathcal{U})\,\boldsymbol{u}_\ell\boldsymbol{u}_\ell^\top. \tag{92}$$

Because $\sum_{\ell=1}^d \boldsymbol{u}_\ell\boldsymbol{u}_\ell^\top = \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_k$, it follows that

$$\sum_{\ell=1}^d \mathbb{E}[\boldsymbol{A}_\ell^2] \;\preceq\; q\,\nu(\mathcal{U})\,\boldsymbol{I}_k, \qquad \text{hence} \qquad \sigma^2 \;:=\; \Big\|\sum_{\ell=1}^d \mathbb{E}[\boldsymbol{A}_\ell^2]\Big\|_{\text{op}} \;\leq\; q\,\nu(\mathcal{U}). \tag{93}$$

By the matrix Bernstein inequality (Tropp (2015) and Lem. C.7), there exists an absolute constant $C_0 > 0$ such that for all $\eta \in (0,1)$, with probability at least $1-\eta$,

$$\|\boldsymbol{P} - q\boldsymbol{I}_k\|_{\mathrm{op}} \;\le\; C_0\Big(\sqrt{\sigma^2\log(2k/\eta)} + R\log(2k/\eta)\Big) \;\le\; C_0\Big(\sqrt{q\,\nu(\mathcal{U})\log(2k/\eta)} + \nu(\mathcal{U})\log(2k/\eta)\Big). \tag{94}$$

Now set $\eta := \delta/(2m^2)$, let

$$\Lambda \;:=\; \log(2k/\eta) \;=\; \log(4km^2/\delta), \tag{95}$$

and define the event

$$\mathcal{G} \;:=\; \Big\{\|\boldsymbol{P} - q\boldsymbol{I}_k\|_{\mathrm{op}} \le \frac{q}{2}\Big\}. \tag{96}$$

By (94) and the assumption $q \ge C\big(\sqrt{q\,\nu(\mathcal{U})\Lambda} + \nu(\mathcal{U})\Lambda\big)$, taken with $C := 2C_0$, we have

$$\mathbb{P}(\mathcal{G}) \;\ge\; 1 - \frac{\delta}{2m^2}. \tag{97}$$

On $\mathcal{G}$, we have $\boldsymbol{P} \succeq (q/2)\boldsymbol{I}_k$, and therefore

$$\det(\boldsymbol{P}) \;\ge\; (q/2)^k. \tag{98}$$

Step ②: Uniform conditional density bound for $\mathbf{v}_{ij}$. Fix $\xi^\star \in \{0,1\}^d$ such that $\mathbb{P}(\xi = \xi^\star) > 0$ and

$$\Big\|\boldsymbol{U}^\top \mathrm{Diag}(\xi^\star)\boldsymbol{U} - q\boldsymbol{I}_k\Big\|_{\mathrm{op}} \;\le\; \frac{q}{2}. \tag{99}$$

Let

$$S \;:=\; \{\ell \in [d] \mid \xi^\star_\ell = 1\} \;=:\; \{\ell_1, \dots, \ell_{|S|}\}, \tag{100}$$

define

$$\zeta \;:=\; \big((\mathbf{f}_i - \mathbf{f}_j)_{\ell_1}, \dots, (\mathbf{f}_i - \mathbf{f}_j)_{\ell_{|S|}}\big) \in \mathbb{R}^{|S|}, \tag{101}$$

and let $\boldsymbol{U}_S \in \mathbb{R}^{|S|\times k}$ be the submatrix of $\boldsymbol{U}$ with rows indexed by S. Then,

$$\boldsymbol{P}^\star \;:=\; \boldsymbol{U}^\top \mathrm{Diag}(\xi^\star)\boldsymbol{U} \;=\; \boldsymbol{U}_S^\top \boldsymbol{U}_S \;\succeq\; \frac{q}{2}\boldsymbol{I}_k. \tag{102}$$

In particular, $\boldsymbol{P}^\star$ is positive definite and $\mathrm{rk}(\boldsymbol{U}_S^\top) = k$. For each $\ell \in [d]$, write $X_\ell := (\mathbf{f}_i - \mathbf{f}_j)_\ell = B_{i\ell}Y_{i\ell} - B_{j\ell}Y_{j\ell}$ and let $\mathcal{E}_\ell := \{\xi_\ell = \xi^\star_\ell\}$. Then, $\{\xi = \xi^\star\} = \bigcap_{\ell=1}^d \mathcal{E}_\ell$, and each $\mathcal{E}_\ell$ is measurable w.r.t. $(B_{i\ell}, B_{j\ell})$. Since the families $\big(B_{i\ell}, B_{j\ell}, Y_{i\ell}, Y_{j\ell}\big)$ are independent across $\ell \in [d]$, it follows that for arbitrary Borel sets $A_1, \dots, A_{|S|} \subseteq \mathbb{R}$,

$$\mathbb{P}\left(\bigcap_{r=1}^{|S|}\{X_{\ell_r} \in A_r\} \,\middle|\, \xi = \xi^\star\right) \;=\; \frac{\mathbb{P}\Big(\bigcap_{r=1}^{|S|}\{X_{\ell_r} \in A_r\} \cap \bigcap_{\ell=1}^d \mathcal{E}_\ell\Big)}{\mathbb{P}(\xi = \xi^\star)} \tag{103}$$

$$= \frac{\prod_{r=1}^{|S|}\mathbb{P}(X_{\ell_r} \in A_r, \mathcal{E}_{\ell_r}) \prod_{\ell \notin S}\mathbb{P}(\mathcal{E}_\ell)}{\prod_{r=1}^{|S|}\mathbb{P}(\mathcal{E}_{\ell_r}) \prod_{\ell \notin S}\mathbb{P}(\mathcal{E}_\ell)} \tag{104}$$

$$= \prod_{r=1}^{|S|}\mathbb{P}(X_{\ell_r} \in A_r \mid \mathcal{E}_{\ell_r}). \tag{105}$$

Thus, the coordinates of ζ are independent under $\mathbb{P}(\cdot|\xi = \xi^\star)$. Now fix $\ell \in S$ for which $\xi^\star_\ell = 1$ and $\mathcal{E}_\ell = \{\xi_\ell = 1\}$. Let f_Y denote the Lebesgue density of Y, with $\|f_Y\|_{L^\infty(\lambda)} \le M$, and let Y' be an independent copy of Y. For every Borel set $A \subseteq \mathbb{R}$,

$$\mathbb{P}(X_\ell \in A \mid \mathcal{E}_\ell) \;=\; \frac{p(1-p)}{q}\mathbb{P}(Y \in A) \;+\; \frac{p(1-p)}{q}\mathbb{P}(-Y \in A) \;+\; \frac{p^2}{q}\mathbb{P}(Y - Y' \in A). \tag{106}$$

Hence, under $\mathbb{P}(\cdot|\mathcal{E}_\ell)$, X_ℓ has a Lebesgue density which is a convex combination of the densities of Y, $-Y$, and $Y - Y'$. $-Y$ has density $x \mapsto f_{-Y}(x) = f_Y(-x)$ and $Y - Y'$ has density $f_{Y-Y'} = f_Y * f_{-Y}$. Thus,

$$\|f_{Y-Y'}\|_{L^\infty(\lambda)} \;=\; \|f_Y * f_{-Y}\|_{L^\infty(\lambda)} \;\le\; \underbrace{\|f_Y\|_{L^\infty(\lambda)}}_{\le M}\underbrace{\|f_{-Y}\|_{L^1(\lambda)}}_{=1} \;\le\; M. \tag{107}$$

Therefore, for every $\ell \in S$, the conditional distribution of X_ℓ given $\mathcal{E}_\ell$ admits a Lebesgue density bounded by M. Together with (105), this shows that under $\mathbb{P}(\cdot|\xi = \xi^\star)$, $\zeta \in \mathbb{R}^{|S|}$ has independent coordinates, each with Lebesgue density bounded by M. Now set $\boldsymbol{A} := \boldsymbol{U}_S^\top \in \mathbb{R}^{k\times|S|}$, so that $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{P}^\star$. Since $\xi^\star_\ell = 0$ implies $B_{i\ell} = B_{j\ell} = 0$, we have $X_\ell = 0$ a.s. under $\mathbb{P}(\cdot|\xi = \xi^\star)$ for every $\ell \notin S$. Consequently,

$$\mathbf{v}_{ij} \;=\; \boldsymbol{U}^\top(\mathbf{f}_i - \mathbf{f}_j) \;=\; \boldsymbol{U}_S^\top\zeta \;=\; \boldsymbol{A}\zeta \qquad \text{a.s. under } \mathbb{P}(\cdot|\xi = \xi^\star). \tag{108}$$

Let $\boldsymbol{\Sigma} := (\boldsymbol{A}\boldsymbol{A}^\top)^{1/2} = (\boldsymbol{P}^\star)^{1/2}$ and $\boldsymbol{Q} := \boldsymbol{\Sigma}^{-1}\boldsymbol{A} \in \mathbb{R}^{k\times|S|}$. Then, $\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}_k$. Writing $\boldsymbol{\Pi} := \boldsymbol{Q}^\top\boldsymbol{Q}$, $\boldsymbol{\Pi}$ is the orthogonal projection onto the k-dim. subspace $\mathrm{im}(\boldsymbol{Q}^\top) \subseteq \mathbb{R}^{|S|}$, and $\boldsymbol{Q}\zeta = \boldsymbol{Q}\boldsymbol{\Pi}\zeta$. Let $\lambda_{\mathrm{im}(\boldsymbol{Q}^\top)}$ denote the k-dimensional Lebesgue measure on the subspace $\mathrm{im}(\boldsymbol{Q}^\top)$. By Rudelson & Vershynin (2015, Theorem 1.1), applied under the conditional probability measure $\mathbb{P}(\cdot|\xi = \xi^\star)$ to ζ and $\boldsymbol{\Pi}$, the random vector $\boldsymbol{\Pi}\zeta$ admits a density $f_{\boldsymbol{\Pi}\zeta|\xi=\xi^\star}$ w.r.t. $\lambda_{\mathrm{im}(\boldsymbol{Q}^\top)}$ satisfying

$$\|f_{\boldsymbol{\Pi}\zeta|\xi=\xi^\star}\|_{L^\infty(\lambda_{\mathrm{im}(\boldsymbol{Q}^\top)})} \;\le\; (C_1 M)^k \tag{109}$$

for an absolute constant $C_1 > 0$. Since the restriction $\boldsymbol{Q}\big|_{\mathrm{im}(\boldsymbol{Q}^\top)} : \mathrm{im}(\boldsymbol{Q}^\top) \to \mathbb{R}^k$ is an isometry, it preserves the k-dim. Lebesgue measure λ^k, and thus, $\boldsymbol{Q}\zeta$ admits a Lebesgue density $f_{\boldsymbol{Q}\zeta|\xi=\xi^\star}$ on $\mathbb{R}^k$ with

$$\|f_{\boldsymbol{Q}\zeta|\xi=\xi^\star}\|_{L^\infty(\lambda^k)} \;\le\; (C_1 M)^k. \tag{110}$$

Since $\mathbf{v}_{ij} = \boldsymbol{A}\zeta = \boldsymbol{\Sigma}(\boldsymbol{Q}\zeta)$ under $\mathbb{P}(\cdot|\xi = \xi^\star)$, a change of variables for the invertible linear map $\boldsymbol{\Sigma} : \mathbb{R}^k \to \mathbb{R}^k$ shows that $\mathbf{v}_{ij}$ admits a Lebesgue density $f_{\mathbf{v}_{ij}|\xi=\xi^\star}$ on $\mathbb{R}^k$ given by

$$f_{\mathbf{v}_{ij}|\xi=\xi^\star}(\boldsymbol{x}) \;=\; \frac{f_{\boldsymbol{Q}\zeta|\xi=\xi^\star}(\boldsymbol{\Sigma}^{-1}\boldsymbol{x})}{|\det \boldsymbol{\Sigma}|}, \qquad \boldsymbol{x} \in \mathbb{R}^k. \tag{111}$$

Therefore,

$$\|f_{\mathbf{v}_{ij}|\xi=\xi^\star}\|_{L^\infty(\lambda^k)} \;\le\; \frac{\|f_{\boldsymbol{Q}\zeta|\xi=\xi^\star}\|_{L^\infty(\lambda^k)}}{|\det \boldsymbol{\Sigma}|} \;\le\; \frac{(C_1 M)^k}{\sqrt{\det(\boldsymbol{A}\boldsymbol{A}^\top)}} \;=\; \frac{(C_1 M)^k}{\sqrt{\det(\boldsymbol{P}^\star)}} \;\le\; (C_1 M)^k \Big(\frac{2}{q}\Big)^{k/2}. \tag{112}$$

Consequently, for every $t \ge 0$,

$$\mathbb{P}(\|\mathbf{v}_{ij}\|_2 \le t \,|\, \xi = \xi^\star) \;=\; \int_{B_t^k} f_{\mathbf{v}_{ij}|\xi=\xi^\star}\, \mathrm{d}\lambda^k \tag{113}$$

$$\le\; \lambda^k(B_t^k)\, \|f_{\mathbf{v}_{ij}|\xi=\xi^\star}\|_{L^\infty(\lambda^k)} \tag{114}$$

$$\le\; V_k t^k (C_1 M)^k \Big(\frac{2}{q}\Big)^{k/2} \;=\; \Big(C_2 V_k^{1/k} \frac{Mt}{\sqrt{q}}\Big)^k, \tag{115}$$

where we set $C_2 := \sqrt{2}\, C_1$.

The bound (115) is uniform over all $\xi^\star \in \{0,1\}^d$ with $\mathbb{P}(\xi = \xi^\star) > 0$ and $\|\boldsymbol{U}^\top \mathrm{Diag}(\xi^\star)\boldsymbol{U} - q\boldsymbol{I}_k\|_{\mathrm{op}} \le q/2$. Since the event $\mathcal{G}$ is measurable w.r.t. ξ, and ξ takes values in the finite set $\{0,1\}^d$, conditioning on $\mathcal{G}$ yields

$$\mathbb{P}(\|\mathbf{v}_{ij}\|_2 \le t \,|\, \mathcal{G}) \;=\; \sum_{\substack{\xi^\star \in \{0,1\}^d \\ \mathbb{P}(\xi=\xi^\star|\mathcal{G})>0}} \mathbb{P}(\xi = \xi^\star \mid \mathcal{G})\, \mathbb{P}(\|\mathbf{v}_{ij}\|_2 \le t \,|\, \xi = \xi^\star) \;\le\; \Big(C_2 V_k^{1/k} \frac{Mt}{\sqrt{q}}\Big)^k. \tag{116}$$

Combining (116) with $\mathbb{P}(\mathcal{G}^\mathsf{c}) \le \delta/(2m^2)$ from Step ①, we obtain

$$\mathbb{P}(\|\mathbf{v}_{ij}\|_2 \le t) \;\le\; \mathbb{P}(\mathcal{G}^\mathsf{c}) + \mathbb{P}(\|\mathbf{v}_{ij}\|_2 \le t \,|\, \mathcal{G}) \;\le\; \frac{\delta}{2m^2} + \Big(C_2 V_k^{1/k} \frac{Mt}{\sqrt{q}}\Big)^k. \tag{117}$$

Step ③: Union bound over all pairs. Choose

$$t \;:=\; c\, \frac{\sqrt{q}}{M} V_k^{-1/k} \left(\frac{\delta}{m^2}\right)^{1/k} \tag{118}$$

with $c := (2C_2)^{-1}$. Then,

$$\Big(C_2 V_k^{1/k} \frac{Mt}{\sqrt{q}}\Big)^k \;=\; (C_2 c)^k \frac{\delta}{m^2} \;=\; \Big(\frac{1}{2}\Big)^k \frac{\delta}{m^2} \;\le\; \frac{1}{2}\frac{\delta}{m^2}, \tag{119}$$

so (117) gives for every fixed pair (i,j)

$$\mathbb{P}(\|\mathbf{v}_{ij}\|_2 \le t) \;\le\; \frac{\delta}{2m^2} + \frac{1}{2}\frac{\delta}{m^2} \;=\; \frac{\delta}{m^2}. \tag{120}$$

By a union bound over the $\binom{m}{2} \le m^2/2$ pairs,

$$\mathbb{P}\left(\min_{i\ne j} \|\mathbf{v}_{ij}\| \le t\right) \;\le\; \sum_{i<j} \mathbb{P}(\|\mathbf{v}_{ij}\|_2 \le t) \;\le\; \frac{m^2}{2} \cdot \frac{\delta}{m^2} \;\le\; \delta. \tag{121}$$

Hence, recalling that $M = \sigma_{\mathbf{F}}^{-1} M_0$,

$$\gamma_{\text{out}} \;=\; \sqrt{2}\min_{i\neq j}\|\mathbf{v}_{ij}\| \;\geq\; \sqrt{2}t \;=\; \sqrt{2}c\,\frac{\sqrt{q}}{M_0}\sigma_{\mathbf{F}} V_k^{-1/k}\left(\frac{\delta}{m^2}\right)^{1/k} \tag{122}$$

with probability at least $1-\delta$, which is precisely the first inequality of (86), absorbing $\sqrt{2}$ into c.

Step ④: Stirling approximation. Fixing δ, q, M_0 and using the Stirling formula

$$\Gamma(x+1) \;\sim\; \sqrt{2\pi x}\left(\frac{x}{e}\right)^x \qquad \text{as } x\to\infty, \tag{123}$$

we obtain

$$V_k^{-1/k} \;=\; \left(\frac{\Gamma(k/2+1)}{\pi^{k/2}}\right)^{1/k} \;\sim\; \left(\frac{\sqrt{\pi k}\left(\frac{k}{2e}\right)^{k/2}}{\pi^{k/2}}\right)^{1/k} \;=\; \underbrace{(\pi k)^{1/(2k)}}_{\to 1,\ k\to\infty}\sqrt{\frac{k}{2\pi e}} \;\sim\; \sqrt{\frac{k}{2\pi e}}, \tag{124}$$

where $\Gamma(x) := \int_0^\infty t^{x-1}e^{-t}\,\mathrm{d}t$ denotes the gamma function. Hence, $V_k^{-1/k} = \Theta(\sqrt{k})$, and absorbing the constant $(2\pi e)^{-1/2}$ into c yields that the r.h.s. of (122) is $\Theta\big(\sigma_{\mathbf{F}}\sqrt{k}\,m^{-2/k}\big)$ in $m, k, \sigma_{\mathbf{F}}$. □

C.3. Input Reindexings

Theorem C.13 (Global input sensitivity)**.** *In the setup of Thm. C.10, fix an input reindexing $\tau \in S_d$. Assume that the concentration event Assumption C.9 holds both for $\mathbf{S}_i$ and $\mathbf{S}_i^\tau := \boldsymbol{U}^\top \boldsymbol{P}_\tau^\top \mathrm{Diag}(\mathbf{d}_i)^2 \boldsymbol{P}_\tau \boldsymbol{U}$.[4] Also assume that $\boldsymbol{\delta}_\tau^{\text{in}} \neq 0$ and $\rho_\tau^{\text{in}} := \|\boldsymbol{R}\|_F / \|\boldsymbol{\delta}_\tau^{\text{in}}\|_F \leq 1$. Then,*

$$\Delta_\tau^{\text{in}}(\boldsymbol{H}) \;=\; \Big(1 \pm \mathcal{O}(\rho_\tau^{\text{in}} + \varepsilon\mu_{\mathbf{D}}^{-1})\Big)\cdot \mu_{\mathbf{D}}^{-1}\|\boldsymbol{\delta}_\tau^{\text{in}}\|_F^2. \tag{125}$$

Proof. The argument is partly analogous to the proof of Thm. C.10. For each $i \in [m]$, let $\boldsymbol{a}_i := \mathbf{v}_i + \boldsymbol{R}_{i,:} \in \mathbb{R}^k$ denote the coefficient vector realized by the i-th neuron on $\mathcal{X}$. Under the input permutation τ, the transformed function is evaluated on the reindexed support $\boldsymbol{P}_\tau\mathcal{X} \subseteq \boldsymbol{P}_\tau\mathcal{U}$, which has $\boldsymbol{P}_\tau\boldsymbol{U}$ as an orthonormal basis. W.r.t. this basis, the same transformed neuron is still represented by the coefficient vector $\boldsymbol{a}_i$, while the corresponding projected center becomes

$$\mathbf{v}_i^\tau \;:=\; (\mathbf{F}\boldsymbol{P}_\tau\boldsymbol{U})_{i,:} \;=\; \mathbf{v}_i + (\boldsymbol{\delta}_\tau^{\text{in}})_{i,:}. \tag{126}$$

Hence,

$$\Delta_\tau^{\text{in}}(\boldsymbol{H},\mathcal{X}) \;=\; \sum_{i=1}^m \left(\big\|\boldsymbol{R}_{i,:} - (\boldsymbol{\delta}_\tau^{\text{in}})_{i,:}\big\|^2_{(\mathbf{S}_i^\tau)^{-1}} - \|\boldsymbol{R}_{i,:}\|^2_{\mathbf{S}_i^{-1}}\right). \tag{127}$$

Since $\varepsilon < \mu_{\mathbf{D}}$, the assumptions imply

$$(\mu_{\mathbf{D}} - \varepsilon)\boldsymbol{I}_k \;\preceq\; \mathbf{S}_i, \mathbf{S}_i^\tau \;\preceq\; (\mu_{\mathbf{D}} + \varepsilon)\boldsymbol{I}_k \qquad \text{for all } i\in[m]. \tag{128}$$

Thus, both $\mathbf{S}_i$ and $\mathbf{S}_i^\tau$ are invertible, and

$$\big\|\mathbf{S}_i^{-1} - \mu_{\mathbf{D}}^{-1}\boldsymbol{I}_k\big\|_{\text{op}}, \big\|(\mathbf{S}_i^\tau)^{-1} - \mu_{\mathbf{D}}^{-1}\boldsymbol{I}_k\big\|_{\text{op}} \;\leq\; \frac{\varepsilon}{\mu_{\mathbf{D}}(\mu_{\mathbf{D}} - \varepsilon)} \;=:\; \beta. \tag{129}$$

Therefore, for every $\boldsymbol{a} \in \mathbb{R}^k$,

$$\Big|\|\boldsymbol{a}\|^2_{\mathbf{S}_i^{-1}} - \mu_{\mathbf{D}}^{-1}\|\boldsymbol{a}\|_2^2\Big| \;\leq\; \beta\|\boldsymbol{a}\|_2^2, \qquad \Big|\|\boldsymbol{a}\|^2_{(\mathbf{S}_i^\tau)^{-1}} - \mu_{\mathbf{D}}^{-1}\|\boldsymbol{a}\|_2^2\Big| \;\leq\; \beta\|\boldsymbol{a}\|_2^2. \tag{130}$$

Applying this to each term in (127) yields

$$\Delta_\tau^{\text{in}}(\boldsymbol{H},\mathcal{X}) \;=\; \mu_{\mathbf{D}}^{-1}\Big(\|\boldsymbol{R} - \boldsymbol{\delta}_\tau^{\text{in}}\|_F^2 - \|\boldsymbol{R}\|_F^2\Big) + E_\tau, \tag{131}$$

with

$$|E_\tau| \;\leq\; \beta\Big(\|\boldsymbol{R} - \boldsymbol{\delta}_\tau^{\text{in}}\|_F^2 + \|\boldsymbol{R}\|_F^2\Big). \tag{132}$$

By Cauchy–Schwarz,

$$|E_\tau| \;\leq\; \beta\Big(\|\boldsymbol{\delta}_\tau^{\text{in}}\|_F^2 + 2\|\boldsymbol{\delta}_\tau^{\text{in}}\|_F\|\boldsymbol{R}\|_F + 2\|\boldsymbol{R}\|_F^2\Big). \tag{133}$$

[4] When $\mathbf{D}$ has i.i.d. entries, $\mathbf{S}_i^\tau = \boldsymbol{U}^\top\boldsymbol{P}_\tau^\top\mathrm{Diag}(\mathbf{d}_i)^2\boldsymbol{P}_\tau\boldsymbol{U} \overset{d}{=} \mathbf{S}_i$.

Moreover,

$$\|\boldsymbol{R} - \boldsymbol{\delta}_\tau^{\text{in}}\|_F^2 - \|\boldsymbol{R}\|_F^2 \;=\; \|\boldsymbol{\delta}_\tau^{\text{in}}\|_F^2 - 2\left\langle \boldsymbol{\delta}_\tau^{\text{in}}, \boldsymbol{R}\right\rangle_F, \tag{134}$$

and by Cauchy–Schwarz,

$$\left|\left\langle \boldsymbol{\delta}_\tau^{\text{in}}, \boldsymbol{R}\right\rangle_F\right| \;\le\; \|\boldsymbol{\delta}_\tau^{\text{in}}\|_F \|\boldsymbol{R}\|_F. \tag{135}$$

Combining (131), (133), (134), and (135) gives

$$\left|\Delta_\tau^{\text{in}}(\boldsymbol{H}, \mathcal{X}) - \mu_{\mathbf{D}}^{-1}\|\boldsymbol{\delta}_\tau^{\text{in}}\|_F^2\right| \;\le\; \frac{2}{\mu_{\mathbf{D}}}\|\boldsymbol{\delta}_\tau^{\text{in}}\|_F\|\boldsymbol{R}\|_F + \frac{\varepsilon}{\mu_{\mathbf{D}}(\mu_{\mathbf{D}} - \varepsilon)}\Big(\|\boldsymbol{\delta}_\tau^{\text{in}}\|_F^2 + 2\|\boldsymbol{\delta}_\tau^{\text{in}}\|_F\|\boldsymbol{R}\|_F + 2\|\boldsymbol{R}\|_F^2\Big). \tag{136}$$

Finally, divide (136) by $\mu_{\mathbf{D}}^{-1}\|\boldsymbol{\delta}_\tau^{\text{in}}\|_F^2$ to obtain

$$\left|\frac{\mu_{\mathbf{D}}\Delta_\tau^{\text{in}}(\boldsymbol{H}, \mathcal{X})}{\|\boldsymbol{\delta}_\tau^{\text{in}}\|_F^2} - 1\right| \;\le\; 2\rho_\tau^{\text{in}} + \frac{\varepsilon}{\mu_{\mathbf{D}} - \varepsilon}\big(1 + 2\rho_\tau^{\text{in}} + 2(\rho_\tau^{\text{in}})^2\big). \tag{137}$$

If $\rho_\tau^{\text{in}} \le 1$ and $\varepsilon \le \mu_{\mathbf{D}}/2$, then

$$\frac{\varepsilon}{\mu_{\mathbf{D}} - \varepsilon} \;\le\; \frac{2\varepsilon}{\mu_{\mathbf{D}}}, \qquad 1 + 2\rho_\tau^{\text{in}} + 2(\rho_\tau^{\text{in}})^2 \;\le\; 5. \tag{138}$$

Thus the r.h.s. is $\mathcal{O}(\rho_\tau^{\text{in}} + \varepsilon\mu_{\mathbf{D}}^{-1})$, and therefore

$$\Delta_\tau^{\text{in}}(\boldsymbol{H}, \mathcal{X}) \;=\; \Big(1 \pm \mathcal{O}(\rho_\tau^{\text{in}} + \varepsilon\mu_{\mathbf{D}}^{-1})\Big) \cdot \mu_{\mathbf{D}}^{-1}\|\boldsymbol{\delta}_\tau^{\text{in}}\|_F^2. \tag{139}$$

□

Corollary C.14 (Minimum input transposition cost)**.** *Assume that the assumptions of Thm. C.13 hold for every transposition $(ab) \in S_d$, and that $\rho_{(ab)}^{\text{in}} \le \rho^{\text{in}} \le 1$ for some ρ^{in} uniformly in $a \ne b$. Then,*

$$\min_{a \ne b} \Delta_{(ab)}^{\text{in}}(\boldsymbol{H}, \mathcal{X}) \;=\; \Big(1 \pm \mathcal{O}(\rho^{\text{in}} + \varepsilon\mu_{\mathbf{D}}^{-1})\Big) \cdot \mu_{\mathbf{D}}^{-1}\gamma_{\text{in}}^2. \tag{140}$$

Proof. Since $\rho_{(ab)}^{\text{in}} \le \rho^{\text{in}}$ for every transposition (ab), by Thm. C.13,

$$\Delta_{(ab)}^{\text{in}}(\boldsymbol{H}, \mathcal{X}) \;=\; \Big(1 \pm \mathcal{O}(\rho^{\text{in}} + \varepsilon\mu_{\mathbf{D}}^{-1})\Big) \cdot \mu_{\mathbf{D}}^{-1}\|\boldsymbol{\delta}_{(ab)}^{\text{in}}\|_F^2 \tag{141}$$

uniformly in $a \ne b$. Taking the minimum over all transpositions gives

$$\min_{a \ne b} \Delta_{(ab)}^{\text{in}}(\boldsymbol{H}, \mathcal{X}) \;=\; \Big(1 \pm \mathcal{O}(\rho^{\text{in}} + \varepsilon\mu_{\mathbf{D}}^{-1})\Big) \cdot \mu_{\mathbf{D}}^{-1}\gamma_{\text{in}}^2, \tag{142}$$

which is exactly the claim. □

Theorem C.15 (Asymptotics of γ_{in})**.** *Let $\mathbf{F}_{:,\ell} \in \mathbb{R}^m$ denote the ℓ-th column of $\mathbf{F}$. Let $a \ne b \in [d]$. Then,*

$$\|\boldsymbol{\delta}_{(ab)}^{\text{in}}\|_F \;=\; \|\mathbf{F}_{:,b} - \mathbf{F}_{:,a}\|_2\, \|\boldsymbol{U}_{b,:} - \boldsymbol{U}_{a,:}\|_2. \tag{143}$$

If further $\mathbf{F}$ has i.i.d. centered subgaussian entries with variance $\sigma_{\mathbf{F}}^2$, subgaussian norm uniformly bounded by $C\sigma_{\mathbf{F}}$, and $m \gtrsim C^4 \log(d^2/\delta)$, then with probability at least $1 - \delta$,

$$\gamma_{\text{in}} \;=\; \Theta(\sigma_{\mathbf{F}}\sqrt{m}) \cdot \min_{a \ne b} \|\boldsymbol{U}_{b,:} - \boldsymbol{U}_{a,:}\|_2. \tag{144}$$

Proof. **Step ①: Proof of (143).** Fix $a \ne b$. For a transposition (ab), we have for any $\boldsymbol{x} \in \mathbb{R}^d$

$$(\boldsymbol{P}_{(ab)} - \boldsymbol{I}_d)\boldsymbol{x} \;=\; (x_b - x_a)(\boldsymbol{e}_a - \boldsymbol{e}_b). \tag{145}$$

Thus,

$$(\boldsymbol{\delta}_{(ab)}^{\text{in}}) \;=\; \mathbf{F}(\boldsymbol{P}_{(ab)} - \boldsymbol{I}_d)\boldsymbol{U} \;=\; \mathbf{F}(\boldsymbol{e}_a - \boldsymbol{e}_b)(\boldsymbol{U}_{b,:} - \boldsymbol{U}_{a,:}) \;=\; (\mathbf{F}_{:,a} - \mathbf{F}_{:,b})(\boldsymbol{U}_{b,:} - \boldsymbol{U}_{a,:}), \tag{146}$$

which is a matrix of rank ≤ 1. Using $\|\boldsymbol{x}\boldsymbol{y}^\top\|_F = \|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2$ gives (143).

Step ②: Proof of (144). We assume that the entries of $\mathbf{F}$ are i.i.d., centered, have variance $\sigma_{\mathbf{F}}^2$, and subgaussian norm $K := \|\mathbf{F}_{ij}\|_{\psi_2}$. Fix $a < b$ and define $Z_i := \mathbf{F}_{ib} - \mathbf{F}_{ia}$, for $i \in [m]$. Then, $Z_1, \ldots, Z_m$ are independent and centered, with

$$\mathbb{E}Z_i^2 \;=\; \mathbb{E}\mathbf{F}_{ib}^2 + \mathbb{E}\mathbf{F}_{ia}^2 \;=\; 2\sigma_{\mathbf{F}}^2. \tag{147}$$

Moreover, by the triangle inequality for the subgaussian norm,

$$\|Z_i\|_{\psi_2} \;\le\; \|\mathbf{F}_{ib}\|_{\psi_2} + \|\mathbf{F}_{ia}\|_{\psi_2} \;=\; 2K. \tag{148}$$

Therefore, for $Y_i := Z_i^2 - \mathbb{E}Z_i^2$, the random variables $Y_1, \dots, Y_m$ are independent, centered, and subexponential. In particular, there exists a universal constant $C_0 > 0$ such that

$$\|Y_i\|_{\psi_1} \;\le\; C_0 \|Z_i\|_{\psi_2}^2 \;\le\; 4C_0K^2. \tag{149}$$

Hence, by the Bernstein inequality for sums of independent subexponential random variables (e.g., Vershynin (2018, Thm. 2.8.2)), and absorbing the universal constant in the bound on $\|Y_i\|_{\psi_1}$, there exists a universal constant $c_1 > 0$ such that for every $t \ge 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^m Y_i\right| \ge t\right) \;\le\; 2\exp\left(-c_1 \min\left\{\frac{t^2}{mK^4}, \frac{t}{K^2}\right\}\right). \tag{150}$$

Set $L := \log(d^2/\delta)$ and choose

$$t \;:=\; C_1K^2\left(\sqrt{mL} + L\right), \tag{151}$$

where $C_1 > 0$ is a universal constant yet to be determined. Then,

$$\frac{t}{K^2} \;=\; C_1\left(\sqrt{mL} + L\right) \;\ge\; C_1L, \qquad \frac{t^2}{mK^4} \;=\; C_1^2\frac{(\sqrt{mL}+L)^2}{m} \;\ge\; C_1^2L. \tag{152}$$

Thus,

$$\min\left\{\frac{t^2}{mK^4}, \frac{t}{K^2}\right\} \;\ge\; \min\{C_1^2, C_1\}L. \tag{153}$$

Choosing C_1 large enough so that $c_1 \min\{C_1^2, C_1\} \ge 1$, we obtain

$$\mathbb{P}\left(\left|\sum_{i=1}^m Y_i\right| \ge t\right) \;\le\; 2\exp\big(-c_1\min\{C_1^2, C_1\}\cdot L\big) \;\le\; 2e^{-L} \;=\; \frac{2\delta}{d^2}. \tag{154}$$

Since

$$\sum_{i=1}^m Y_i \;=\; \sum_{i=1}^m Z_i^2 - 2m\sigma_{\mathbf{F}}^2 \;=\; \|\mathbf{F}_{:,b} - \mathbf{F}_{:,a}\|_2^2 - 2m\sigma_{\mathbf{F}}^2, \tag{155}$$

we have shown that, for this fixed pair $a < b$,

$$\mathbb{P}\left(\left|\|\mathbf{F}_{:,b} - \mathbf{F}_{:,a}\|_2^2 - 2m\sigma_{\mathbf{F}}^2\right| \ge t\right) \;\le\; \frac{2\delta}{d^2}. \tag{156}$$

Taking a union bound over the $\binom{d}{2} \le d^2/2$ pairs, we conclude that, with probability at least $1-\delta$, simultaneously for all $a < b$,

$$\left|\|\mathbf{F}_{:,b} - \mathbf{F}_{:,a}\|_2^2 - 2m\sigma_{\mathbf{F}}^2\right| \;\le\; C_1K^2\left(\sqrt{m\log(d^2/\delta)} + \log(d^2/\delta)\right). \tag{157}$$

On this event, if $t < 2m\sigma_{\mathbf{F}}^2$, then for every $a \ne b$,

$$\sqrt{2m\sigma_{\mathbf{F}}^2 - t} \;\le\; \|\mathbf{F}_{:,b} - \mathbf{F}_{:,a}\|_2 \;\le\; \sqrt{2m\sigma_{\mathbf{F}}^2 + t}. \tag{158}$$

Combining this with (143) gives

$$\sqrt{2m\sigma_{\mathbf{F}}^2 - t}\,\min_{a\ne b}\|\boldsymbol{U}_{a,:} - \boldsymbol{U}_{b,:}\|_2 \;\le\; \gamma_{\text{in}} \;\le\; \sqrt{2m\sigma_{\mathbf{F}}^2 + t}\,\min_{a\ne b}\|\boldsymbol{U}_{a,:} - \boldsymbol{U}_{b,:}\|_2. \tag{159}$$

It remains to translate (159) into the asymptotic statement. By assumption, $K = \|\mathbf{F}_{ij}\|_{\psi_2} \le C\sigma_{\mathbf{F}}$. Increasing C if necessary, we may assume $C \ge 1$. We obtain

$$t \;\le\; C_1C^2\sigma_{\mathbf{F}}^2\left(\sqrt{mL} + L\right). \tag{160}$$

We now choose a universal constant $C_2 \ge \max\{4C_1^2, 2C_1\}$. If $m \ge C_2C^4L$, then

$$C_1C^2\sqrt{mL} \;\le\; m/2, \qquad C_1C^2L \;\le\; m/2. \tag{161}$$

Indeed, the first inequality follows from $C_1C^2\sqrt{mL} \le m/2 \Leftrightarrow m \ge 4C_1^2C^4L$; the second follows since $C \ge 1$ and $C_2 \ge 2C_1$. Therefore, $t \le m\sigma_{\mathbf{F}}^2$, and substituting this into (159) gives

$$\sigma_{\mathbf{F}}\sqrt{m}\,\min_{a\ne b}\|\boldsymbol{U}_{b,:} - \boldsymbol{U}_{a,:}\|_2 \;\le\; \gamma_{\text{in}} \;\le\; \sqrt{3}\,\sigma_{\mathbf{F}}\sqrt{m}\,\min_{a\ne b}\|\boldsymbol{U}_{b,:} - \boldsymbol{U}_{a,:}\|_2. \tag{162}$$

Equivalently,

$$\gamma_{\text{in}} \;=\; \Theta(\sigma_{\mathbf{F}}\sqrt{m})\cdot\min_{a\ne b}\|\boldsymbol{U}_{b,:} - \boldsymbol{U}_{a,:}\|_2 \tag{163}$$

with probability at least $1-\delta$. □

C.4. Identifiability vs. Feature Learning

Let $\boldsymbol{x} \sim \mathbb{P} := \mathcal{N}(0, \boldsymbol{I}_d)$ and let $\eta = \mathrm{ReLU}$. For any measurable $h : \mathbb{R}^d \to \mathbb{R}$, write $\|h\|^2_{L^2(\mathbb{P})} := \mathbb{E}_{\boldsymbol{x}}[h(\boldsymbol{x})^2]$. For the symmetric full feature learning regime, consider a two-layer network of width m as

$$f_{\mathrm{sym}}(\boldsymbol{x}) \;:=\; \boldsymbol{\alpha}^\top \eta(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}), \tag{164}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{W} \in \mathbb{R}^{m\times d}$, and $\boldsymbol{b} \in \mathbb{R}^m$ are trainable and η acts elementwise. For an asymmetric model, fix centers $\mathbf{F} \in \mathbb{R}^{m\times d}$ and consider

$$f_{\mathrm{asym}}(\boldsymbol{x}) \;:=\; \boldsymbol{\alpha}^\top \eta((\mathbf{F} + \boldsymbol{W})\boldsymbol{x}), \tag{165}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^m$ and $\boldsymbol{W} \in \mathbb{R}^{m\times d}$ are trainable.

Lemma C.16 (Lipschitz stability of ridge features)**.** *Let $\boldsymbol{x} \sim \mathbb{P} = \mathcal{N}(0, \boldsymbol{I}_d)$ and $\eta = \mathrm{ReLU}$. Then for any $\mathbf{f} \in \mathbb{R}^d$ and any perturbation $\boldsymbol{w} \in \mathbb{R}^d$,*

$$\big\|\eta\big((\mathbf{f} + \boldsymbol{w})^\top \boldsymbol{x}\big) - \eta\big(\mathbf{f}^\top \boldsymbol{x}\big)\big\|_{L^2(\mathbb{P})} \;\le\; \|\boldsymbol{w}\|_2. \tag{166}$$

Proof. Since $\eta = \mathrm{ReLU}$ is 1-Lipschitz, we have pointwise for all $\boldsymbol{x} \in \mathbb{R}^d$,

$$\big|\eta\big((\mathbf{f} + \boldsymbol{w})^\top \boldsymbol{x}\big) - \eta\big(\mathbf{f}^\top \boldsymbol{x}\big)\big| \;\le\; \big|(\mathbf{f} + \boldsymbol{w})^\top \boldsymbol{x} - \mathbf{f}^\top \boldsymbol{x}\big| \;=\; |\boldsymbol{w}^\top \boldsymbol{x}|. \tag{167}$$

Taking the expectation under $\boldsymbol{x}$ yields

$$\big\|\eta\big((\mathbf{f} + \boldsymbol{w})^\top \boldsymbol{x}\big) - \eta\big(\mathbf{f}^\top \boldsymbol{x}\big)\big\|^2_{L^2(\mathbb{P})} \;\le\; \mathbb{E}_{\boldsymbol{x}}\big[(\boldsymbol{w}^\top \boldsymbol{x})^2\big] \;=\; \boldsymbol{w}^\top \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^\top]\, \boldsymbol{w} \;=\; \|\boldsymbol{w}\|_2^2, \tag{168}$$

since $\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{I}_d$. □

Lemma C.17 (Asymmetric network stability under bounded perturbations)**.** *Let $\boldsymbol{x} \sim \mathbb{P} = \mathcal{N}(0, \boldsymbol{I}_d)$ and $\eta = \mathrm{ReLU}$. Fix $\mathbf{F} \in \mathbb{R}^{m\times d}$ and consider the following two networks with the same output weights $\boldsymbol{\alpha} \in \mathbb{R}^m$*

$$f_{\mathrm{asym}}(\boldsymbol{x}) := \boldsymbol{\alpha}^\top \eta\big((\mathbf{F} + \boldsymbol{W})\boldsymbol{x}\big), \qquad f_0(\boldsymbol{x}) := \boldsymbol{\alpha}^\top \eta\big(\mathbf{F}\boldsymbol{x}\big), \tag{169}$$

where $\boldsymbol{W} \in \mathbb{R}^{m\times d}$ is arbitrary. Then,

$$\|f_{\mathrm{asym}} - f_0\|_{L^2(\mathbb{P})} \;\le\; \|\boldsymbol{\alpha}\|_2\, \|\boldsymbol{W}\|_F. \tag{170}$$

Proof. Write $\mathbf{f}_i \in \mathbb{R}^d$ and $\boldsymbol{w}_i \in \mathbb{R}^d$ for the i-th rows of $\mathbf{F}$ and $\boldsymbol{W}$, respectively, so that

$$f_{\mathrm{asym}}(\boldsymbol{x}) - f_0(\boldsymbol{x}) \;=\; \sum_{i=1}^m \alpha_i \Big(\eta\big((\mathbf{f}_i + \boldsymbol{w}_i)^\top \boldsymbol{x}\big) - \eta\big(\mathbf{f}_i^\top \boldsymbol{x}\big)\Big). \tag{171}$$

Let $\delta_i(\boldsymbol{x}) := \eta\big((\mathbf{f}_i + \boldsymbol{w}_i)^\top \boldsymbol{x}\big) - \eta\big(\mathbf{f}_i^\top \boldsymbol{x}\big)$. By Cauchy–Schwarz applied pointwise in $\boldsymbol{x}$,

$$\big|f_{\mathrm{asym}}(\boldsymbol{x}) - f_0(\boldsymbol{x})\big| \;\le\; \|\boldsymbol{\alpha}\|_2\, \|\boldsymbol{\delta}(\boldsymbol{x})\|_2, \qquad \boldsymbol{\delta}(\boldsymbol{x}) \;:=\; (\delta_1(\boldsymbol{x}), \ldots, \delta_m(\boldsymbol{x}))^\top. \tag{172}$$

Squaring and taking expectation gives

$$\|f_{\mathrm{asym}} - f_0\|^2_{L^2(\mathbb{P})} \;\le\; \|\boldsymbol{\alpha}\|_2^2\, \mathbb{E}_{\boldsymbol{x}}\big[\|\boldsymbol{\delta}(\boldsymbol{x})\|_2^2\big] \;=\; \|\boldsymbol{\alpha}\|_2^2 \sum_{i=1}^m \|\delta_i\|^2_{L^2(\mathbb{P})}. \tag{173}$$

Applying Lem. C.16 to each $(\mathbf{f}_i, \boldsymbol{w}_i)$ yields $\|\delta_i\|_{L^2(\mathbb{P})} \le \|\boldsymbol{w}_i\|_2$, hence

$$\|f_{\mathrm{asym}} - f_0\|^2_{L^2(\mathbb{P})} \;\le\; \|\boldsymbol{\alpha}\|_2^2 \sum_{i=1}^m \|\boldsymbol{w}_i\|_2^2 \;=\; \|\boldsymbol{\alpha}\|_2^2\, \|\boldsymbol{W}\|_F^2. \tag{174}$$

□

Theorem C.18 (Random feature hardness for single ReLU neuron (Yehudai & Shamir, 2019), specialization of Theorem 4.2)**.** *There exists a universal constant $c > 0$ such that for all $d > 40$, the following holds: For every $\boldsymbol{w}_\star \in \mathbb{R}^d$ with $\|\boldsymbol{w}_\star\|_2 = d^2$, there exists a bias $b_\star \in \mathbb{R}$ with $|b_\star| \le 6d^3 + 1$ such that for any coefficients $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$, with probability at least $1 - \exp(-cd)$ over i.i.d. $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m \sim \mathrm{Unif}(\mathbb{S}^{d-1})$,*

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\Big(\sum_{i=1}^m \alpha_i\, \eta(\boldsymbol{b}_i^\top \boldsymbol{x}) \;-\; \eta(\boldsymbol{w}_\star^\top \boldsymbol{x} + b_\star)\Big)^2\right] \;\le\; \frac{1}{50} \quad\Longrightarrow\quad m \cdot \max_{i\in[m]} |\alpha_i| \;\ge\; \frac{1}{48d^2}\exp(cd). \tag{175}$$

Theorem C.19 (Hardness in center-dominated regime)**.** *Let $\boldsymbol{x} \sim \mathbb{P} = \mathcal{N}(0, \boldsymbol{I}_d)$, $\eta = \mathrm{ReLU}$, and fix $d > 40$, $m \in \mathbb{N}$. Draw $\mathbf{F} \in \mathbb{R}^{m\times d}$ with i.i.d. $\mathbf{f}_i \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. The following holds for every $\boldsymbol{w}_\star \in \mathbb{R}^d$ with $\|\boldsymbol{w}_\star\|_2 = d^2$: There exists a bias $b_\star \in \mathbb{R}$ with $|b_\star| \le 6d^3 + 1$ such that for all $\boldsymbol{\alpha} \in \mathbb{R}^m$ and $\boldsymbol{W} \in \mathbb{R}^{m\times d}$ satisfying $\|\boldsymbol{\alpha}\|_\infty < \frac{1}{48md^2}\exp(cd)$ and*

$\|\boldsymbol{\alpha}\|_2\|\boldsymbol{W}\|_F \le \frac{1}{10\sqrt{2}}$, *with probability at least* $1-\exp(-cd)$ *over sampling* $\mathbf{F}$, *the asymmetric network* $f_{\text{asym}}(\boldsymbol{x}) := \boldsymbol{\alpha}^\top \eta\big((\mathbf{F}+\boldsymbol{W})\boldsymbol{x}\big)$ *incurs the population error lower bound*

$$\|f_{\text{asym}} - \eta(\boldsymbol{w}_\star^\top(\cdot)+b_\star)\|_{L^2(\mathbb{P})} \;>\; 1/(10\sqrt{2}). \tag{176}$$

Proof. Work on the event $\mathcal{E}$ (of probability at least $1-\exp(-cd)$) on which Thm. C.18 holds for $\{\mathbf{f}_i\}_{i=1}^m$. Fix any $\boldsymbol{w}_\star \in \mathbb{R}^d$ with $\|\boldsymbol{w}_\star\|_2 = d^2$. On $\mathcal{E}$, Thm. C.18 provides a bias $b_\star \in \mathbb{R}$ with $|b_\star| \le 6d^3+1$ such that for every choice of coefficients $\alpha_1,\dots,\alpha_m \in \mathbb{R}$,

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\Big[\Big(\sum_{i=1}^m \alpha_i\,\eta(\mathbf{f}_i^\top\boldsymbol{x}) - \eta(\boldsymbol{w}_\star^\top\boldsymbol{x}+b_\star)\Big)^2\Big] \le \frac{1}{50} \implies m\max_{i\in[m]}|\alpha_i| \;\ge\; \frac{1}{48d^2}\exp(cd). \tag{177}$$

Now fix arbitrary $\boldsymbol{\alpha}\in\mathbb{R}^m$ and $\boldsymbol{W}\in\mathbb{R}^{m\times d}$ satisfying

$$\|\boldsymbol{\alpha}\|_\infty \;<\; \frac{1}{48md^2}\exp(cd), \qquad \|\boldsymbol{\alpha}\|_2\|\boldsymbol{W}\|_F \;\le\; \frac{1}{10\sqrt{2}}. \tag{178}$$

Define

$$f_{\text{asym}}(\boldsymbol{x}) \;:=\; \boldsymbol{\alpha}^\top\eta\big((\mathbf{F}+\boldsymbol{W})\boldsymbol{x}\big), \qquad f_0(\boldsymbol{x}) \;:=\; \boldsymbol{\alpha}^\top\eta(\mathbf{F}\boldsymbol{x}) \;=\; \sum_{i=1}^m \alpha_i\,\eta(\mathbf{f}_i^\top\boldsymbol{x}). \tag{179}$$

By Lem. C.17,

$$\|f_{\text{asym}} - f_0\|_{L^2(\mathbb{P})} \;\le\; \|\boldsymbol{\alpha}\|_2\|\boldsymbol{W}\|_F \;\le\; \frac{1}{10\sqrt{2}}. \tag{180}$$

Next,

$$m\max_{i\in[m]}|\alpha_i| \;<\; \frac{1}{48d^2}\exp(cd) \tag{181}$$

as required in the theorem. Therefore, by the contrapositive of (177), we must have

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\big[(f_0(\boldsymbol{x}) - \eta(\boldsymbol{w}_\star^\top\boldsymbol{x}+b_\star))^2\big] > \frac{1}{50}, \qquad \text{i.e.} \qquad \|f_0 - \eta(\boldsymbol{w}_\star^\top\boldsymbol{x}+b_\star)\|_{L^2(\mathbb{P})} > \frac{1}{\sqrt{50}} = \frac{1}{5\sqrt{2}}. \tag{182}$$

Finally, by the triangle inequality and (180),

$$\|f_{\text{asym}} - \eta(\boldsymbol{w}_\star^\top\boldsymbol{x}+b_\star)\|_{L^2(\mathbb{P})} \ge \|f_0 - \eta(\boldsymbol{w}_\star^\top\boldsymbol{x}+b_\star)\|_{L^2(\mathbb{P})} - \|f_{\text{asym}} - f_0\|_{L^2(\mathbb{P})} > \frac{1}{5\sqrt{2}} - \frac{1}{10\sqrt{2}} = \frac{1}{10\sqrt{2}}. \tag{183}$$

□

D. Linear Mode Connectivity (§ 4)

Theorem D.1 (Chord dev. in center-dom. regime, ReLU)**.** *Let* $\eta = \mathrm{ReLU}$ *and* $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ *with* $\boldsymbol{\Sigma} \succeq 0$*. Fix* $\boldsymbol{W}^A, \boldsymbol{W}^B \in \mathbb{R}^{m\times d}$*. Suppose there exists* $\beta \in [0,1)$ *such that for all* $i \in [m]$*,*

$$\|\mathbf{d}_i \odot \boldsymbol{w}_i^A\|_{\boldsymbol{\Sigma}} \le \beta \, \|\mathbf{f}_i\|_{\boldsymbol{\Sigma}}, \quad \|\mathbf{d}_i \odot \boldsymbol{w}_i^B\|_{\boldsymbol{\Sigma}} \le \beta \, \|\mathbf{f}_i\|_{\boldsymbol{\Sigma}}, \tag{184}$$

where $\|u\|_{\boldsymbol{\Sigma}} := \sqrt{u^\top \boldsymbol{\Sigma} u}$*. Then,*

$$\sup_{\lambda\in[0,1]} \left\|\xi_{\boldsymbol{H}}(\lambda;\cdot)\right\|_{L^2(\mathbb{P};\mathbb{R}^m)} \;=\; \mathcal{O}(\beta^{3/2}) \, \|\mathbf{F}\boldsymbol{\Sigma}^{1/2}\|_F. \tag{185}$$

Proof. Fix $i \in [m]$. Write

$$\boldsymbol{w}_{\mathrm{eff},i}^A \;:=\; \mathbf{f}_i + \mathbf{d}_i \odot \boldsymbol{w}_i^A, \qquad \boldsymbol{w}_{\mathrm{eff},i}^B \;:=\; \mathbf{f}_i + \mathbf{d}_i \odot \boldsymbol{w}_i^B \tag{186}$$

for the effective weights and

$$Z_i^A \;:=\; (\boldsymbol{w}_{\mathrm{eff},i}^A)^\top \boldsymbol{x}, \qquad Z_i^B \;:=\; (\boldsymbol{w}_{\mathrm{eff},i}^B)^\top \boldsymbol{x} \tag{187}$$

for the preactivation random variables under the input distribution. If $\|\mathbf{f}_i\|_{\boldsymbol{\Sigma}} = 0$, then (184) forces $\|\mathbf{d}_i \odot \boldsymbol{w}_i^A\|_{\boldsymbol{\Sigma}} = \|\mathbf{d}_i \odot \boldsymbol{w}_i^B\|_{\boldsymbol{\Sigma}} = 0$, hence $Z_i^A = Z_i^B = 0$ a.s. and $\xi_i(\lambda;\boldsymbol{x}) \equiv 0$; in this case the desired bound holds trivially for neuron i. Assume from now on that $\|\mathbf{f}_i\|_{\boldsymbol{\Sigma}} > 0$. Then, $Z_i^A \sim \mathcal{N}(0,(\sigma_i^A)^2)$, $Z_i^B \sim \mathcal{N}(0,(\sigma_i^B)^2)$ with $\sigma_i^A := \|\boldsymbol{w}_{\mathrm{eff},i}^A\|_{\boldsymbol{\Sigma}}$, $\sigma_i^B := \|\boldsymbol{w}_{\mathrm{eff},i}^B\|_{\boldsymbol{\Sigma}}$. Since $\beta < 1$, we have $\sigma_i^A, \sigma_i^B > 0$. Further, the correlation of Z_i^A and Z_i^B is

$$\rho_i \;:=\; \frac{\langle \boldsymbol{w}_{\mathrm{eff},i}^A, \boldsymbol{w}_{\mathrm{eff},i}^B\rangle_{\boldsymbol{\Sigma}}}{\sigma_i^A \sigma_i^B}, \qquad \theta_i \;:=\; \arccos(\rho_i). \tag{188}$$

For ReLU, one obtains

$$\begin{aligned}
\xi_i(\lambda;\boldsymbol{x})^2 \;&=\; \Big(\mathrm{ReLU}\big((1-\lambda)Z_i^A + \lambda Z_i^B\big) - \big((1-\lambda)\mathrm{ReLU}(Z_i^A) + \lambda \mathrm{ReLU}(Z_i^B)\big)\Big)^2 && (189)\\
&=\; \min\{(1-\lambda)^2 (Z_i^A)^2,\; \lambda^2 (Z_i^B)^2\}\, \mathbb{1}\{Z_i^A Z_i^B \le 0\} && (190)\\
&\le\; (1-\lambda)^2 (Z_i^A)^2 \mathbb{1}\{Z_i^A Z_i^B \le 0\} + \lambda^2 (Z_i^B)^2 \mathbb{1}\{Z_i^A Z_i^B \le 0\}. && (191)
\end{aligned}$$

For centered bivariate Gaussian random variables with variances $(\sigma_i^A)^2, (\sigma_i^B)^2$ and correlation $\rho_i = \cos(\theta_i)$,

$$\mathbb{E}\big[(Z_i^A)^2 \mathbb{1}\{Z_i^A Z_i^B \le 0\}\big] = (\sigma_i^A)^2 \, \frac{\theta_i - \sin(\theta_i)\cos(\theta_i)}{\pi}, \tag{192}$$

and likewise

$$\mathbb{E}\big[(Z_i^B)^2 \mathbb{1}\{Z_i^A Z_i^B \le 0\}\big] = (\sigma_i^B)^2 \, \frac{\theta_i - \sin(\theta_i)\cos(\theta_i)}{\pi}. \tag{193}$$

Indeed, if $(G_1, G_2) \sim \mathcal{N}(0, \boldsymbol{I}_2)$, then

$$(Z_i^A/\sigma_i^A, Z_i^B/\sigma_i^B) \;\overset{d}{=}\; (G_1, \cos(\theta_i)G_1 + \sin(\theta_i)G_2), \tag{194}$$

where $\theta_i = \arccos(\rho_i) \in [0,\pi]$. Since (192) only depends on the joint distribution, we may compute it under this representation. Write $(G_1, G_2) = R(\cos\varphi, \sin\varphi)$ in polar coordinates. Since the standard Gaussian distribution is invariant under rotations, $\varphi \sim \mathrm{Unif}\big([0, 2\pi)\big)$ is independent of R, and $\mathbb{E}[R^2] = \mathbb{E}[G_1^2 + G_2^2] = 2$. In polar coordinates, we have

$$\begin{aligned}
Z_i^A/\sigma_i^A \;&=\; R\cos(\varphi), && (195)\\
Z_i^B/\sigma_i^B \;&=\; \cos(\theta_i) R\cos(\varphi) + \sin(\theta_i) R\sin(\varphi) \;=\; R\cos(\varphi - \theta_i), && (196)
\end{aligned}$$

where the last equality uses $\cos(\varphi - \theta_i) = \cos(\varphi)\cos(\theta_i) + \sin(\varphi)\sin(\theta_i)$. Thus, since $\sigma_i^A, \sigma_i^B > 0$ and $R > 0$ a.s.,

$$\begin{aligned}
Z_i^A Z_i^B \le 0 \;&\Leftrightarrow\; \sigma_i^A \sigma_i^B R^2 \cos(\varphi)\cos(\varphi - \theta_i) \le 0 && (197)\\
&\Leftrightarrow\; \cos(\varphi)\cos(\varphi - \theta_i) \le 0. && (198)
\end{aligned}$$

For $\theta_i \in [0,\pi]$, this holds exactly for

$$\varphi \in \left[\frac{\pi}{2}, \frac{\pi}{2} + \theta_i\right] \cup \left[\frac{3\pi}{2}, \frac{3\pi}{2} + \theta_i\right], \tag{199}$$

with endpoints modulo 2π. Therefore,

$$\begin{aligned}
\mathbb{E}\big[(Z_i^A)^2 \mathbb{1}\{Z_i^A Z_i^B \le 0\}\big] &= (\sigma_i^A)^2\, \mathbb{E}\big[R^2 \cos^2(\varphi) \mathbb{1}\{\cos(\varphi)\cos(\varphi-\theta_i) \le 0\}\big] && (200)\\
&= (\sigma_i^A)^2\, \underbrace{\mathbb{E}\big[R^2\big]}_{=2}\, \mathbb{E}\big[\cos^2(\varphi) \mathbb{1}\{\cos(\varphi)\cos(\varphi-\theta_i) \le 0\}\big] && (201)\\
&= (\sigma_i^A)^2\, \frac{2}{2\pi} \int_{\{\cos(\varphi)\cos(\varphi-\theta_i)\le 0\}} \cos^2(\varphi)\, \mathrm{d}\varphi && (202)\\
&= (\sigma_i^A)^2\, \frac{2}{\pi} \int_{\pi/2}^{\pi/2+\theta_i} \cos^2(\varphi)\, \mathrm{d}\varphi && (203)\\
&= (\sigma_i^A)^2\, \frac{\theta_i - \sin(\theta_i)\cos(\theta_i)}{\pi}, && (204)
\end{aligned}$$

where the last equality follows from $\int \cos^2(\varphi)\,\mathrm{d}\varphi = \varphi/2 + \sin(2\varphi)/4$ and $\sin(2\theta_i) = 2\sin(\theta_i)\cos(\theta_i)$. Hence,

$$\|\xi_i(\lambda;\cdot)\|_{L^2(\mathbb{P})}^2 \;\le\; \Big((1-\lambda)^2(\sigma_i^A)^2 + \lambda^2(\sigma_i^B)^2\Big)\frac{\theta_i - \sin(\theta_i)\cos(\theta_i)}{\pi}. \tag{205}$$

It remains to bound the size and angular terms in (205) using (184). For the size term, write $\boldsymbol{w}_{\mathrm{eff},i}^A = \mathbf{f}_i + \boldsymbol{r}_i^A$ and $\boldsymbol{w}_{\mathrm{eff},i}^B = \mathbf{f}_i + \boldsymbol{r}_i^B$, where $\boldsymbol{r}_i^A := \mathbf{d}_i \odot \boldsymbol{w}_i^A$ and $\boldsymbol{r}_i^B := \mathbf{d}_i \odot \boldsymbol{w}_i^B$. Then

$$\sigma_i^A = \|\mathbf{f}_i + \boldsymbol{r}_i^A\|_{\boldsymbol{\Sigma}} \le \|\mathbf{f}_i\|_{\boldsymbol{\Sigma}} + \|\boldsymbol{r}_i^A\|_{\boldsymbol{\Sigma}} \le (1+\beta)\|\mathbf{f}_i\|_{\boldsymbol{\Sigma}}, \tag{206}$$

and similarly $\sigma_i^B \le (1+\beta)\|\mathbf{f}_i\|_{\boldsymbol{\Sigma}}$. Hence, since $(1-\lambda)^2 + \lambda^2 \le 1$,

$$(1-\lambda)^2(\sigma_i^A)^2 + \lambda^2(\sigma_i^B)^2 \;\le\; (1+\beta)^2\, \|\mathbf{f}_i\|_{\boldsymbol{\Sigma}}^2. \tag{207}$$

To bound θ_i, define

$$\boldsymbol{u}_i^0 := \frac{\mathbf{f}_i}{\|\mathbf{f}_i\|_{\boldsymbol{\Sigma}}}, \qquad \boldsymbol{u}_i^A := \frac{\mathbf{f}_i + \boldsymbol{r}_i^A}{\|\mathbf{f}_i + \boldsymbol{r}_i^A\|_{\boldsymbol{\Sigma}}}, \qquad \boldsymbol{u}_i^B := \frac{\mathbf{f}_i + \boldsymbol{r}_i^B}{\|\mathbf{f}_i + \boldsymbol{r}_i^B\|_{\boldsymbol{\Sigma}}}. \tag{208}$$

These vectors are normalized in $\|\cdot\|_{\boldsymbol{\Sigma}}$, and $\rho_i = \langle \boldsymbol{u}_i^A, \boldsymbol{u}_i^B\rangle_{\boldsymbol{\Sigma}}$. Set

$$\eta_i^A \;:=\; \frac{\|\boldsymbol{r}_i^A\|_{\boldsymbol{\Sigma}}}{\|\mathbf{f}_i\|_{\boldsymbol{\Sigma}}}, \qquad s_i^A \;:=\; \frac{\langle \boldsymbol{r}_i^A, \mathbf{f}_i\rangle_{\boldsymbol{\Sigma}}}{\|\mathbf{f}_i\|_{\boldsymbol{\Sigma}}^2}. \tag{209}$$

Then $\eta_i^A \le \beta$ and $s_i^A \ge -\eta_i^A$ by Cauchy-Schwarz. Since $\beta < 1$, this gives $1 + s_i^A > 0$, and

$$\langle \boldsymbol{u}_i^A, \boldsymbol{u}_i^0\rangle_{\boldsymbol{\Sigma}} = \frac{1+s_i^A}{\sqrt{1+2s_i^A+(\eta_i^A)^2}}. \tag{210}$$

Moreover,

$$\frac{(1+s_i^A)^2}{1+2s_i^A+(\eta_i^A)^2} - \big(1-(\eta_i^A)^2\big) \;=\; \frac{(s_i^A+(\eta_i^A)^2)^2}{1+2s_i^A+(\eta_i^A)^2} \;\ge\; 0. \tag{211}$$

Thus,

$$\langle \boldsymbol{u}_i^A, \boldsymbol{u}_i^0\rangle_{\boldsymbol{\Sigma}} \;\ge\; \sqrt{1-(\eta_i^A)^2} \;\ge\; \sqrt{1-\beta^2}. \tag{212}$$

The same argument gives $\langle \boldsymbol{u}_i^B, \boldsymbol{u}_i^0\rangle_{\boldsymbol{\Sigma}} \ge \sqrt{1-\beta^2}$. Now write $a_i^A := \langle \boldsymbol{u}_i^A, \boldsymbol{u}_i^0\rangle_{\boldsymbol{\Sigma}}$ and $a_i^B := \langle \boldsymbol{u}_i^B, \boldsymbol{u}_i^0\rangle_{\boldsymbol{\Sigma}}$. Since $\|\boldsymbol{u}_i^A - a_i^A\boldsymbol{u}_i^0\|_{\boldsymbol{\Sigma}}^2 = 1-(a_i^A)^2$ and similarly for B, Cauchy-Schwarz gives

$$\begin{aligned}
\rho_i \;=\; \langle \boldsymbol{u}_i^A, \boldsymbol{u}_i^B\rangle_{\boldsymbol{\Sigma}} \;&=\; a_i^A a_i^B + \big\langle \boldsymbol{u}_i^A - a_i^A\boldsymbol{u}_i^0, \boldsymbol{u}_i^B - a_i^B\boldsymbol{u}_i^0\big\rangle_{\boldsymbol{\Sigma}} && (213)\\
&\ge\; a_i^A a_i^B - \sqrt{1-(a_i^A)^2}\sqrt{1-(a_i^B)^2}. && (214)
\end{aligned}$$

The function

$$(a,b) \mapsto ab - \sqrt{1-a^2}\sqrt{1-b^2} \tag{215}$$

is increasing in each argument on $[0,1]^2$. Since $a_i^A, a_i^B \ge \sqrt{1-\beta^2}$, we obtain

$$\rho_i \;\ge\; (1-\beta^2) - \beta^2 = 1 - 2\beta^2. \tag{216}$$

Therefore,

$$\theta_i = \arccos(\rho_i) \;\le\; \arccos(1-2\beta^2) = 2\arcsin(\beta) \;\le\; \pi\beta. \tag{217}$$

Thus,

$$\frac{\theta_i - \sin(\theta_i)\cos(\theta_i)}{\pi} \;=\; \frac{1}{\pi}\int_0^{\theta_i} 2\underbrace{\sin^2(s)}_{\le s^2}\,\mathrm{d}s \;\le\; \frac{2}{3\pi}\theta_i^3 \;\le\; \frac{2\pi^2}{3}\beta^3. \tag{218}$$

Combining (205), (207), and (218) yields

$$\|\xi_i(\lambda;\cdot)\|^2_{L^2(\mathbb{P})} \;\le\; \frac{2\pi^2}{3}\,(1+\beta)^2\beta^3\,\|\mathbf{f}_i\|^2_{\boldsymbol{\Sigma}}. \tag{219}$$

Summing over i and using $\|\xi_{\boldsymbol{H}}(\lambda;\cdot)\|^2_{L^2(\mathbb{P};\mathbb{R}^m)} = \sum_{i=1}^m \|\xi_i(\lambda;\cdot)\|^2_{L^2(\mathbb{P})}$ gives

$$\|\xi_{\boldsymbol{H}}(\lambda;\cdot)\|^2_{L^2(\mathbb{P};\mathbb{R}^m)} \;\le\; \frac{2\pi^2}{3}\,(1+\beta)^2\beta^3\,\|\mathbf{F}\boldsymbol{\Sigma}^{1/2}\|^2_F. \tag{220}$$

Finally, taking $\sqrt{\cdot}$ and the supremum over $\lambda \in [0,1]$ on the l.h.s. proves the claim. □

We also record a version of Thm. D.1 for smooth activation functions.

Theorem D.2 (Chord deviation in center-dominated regime, C^2 activations)**.** *Let $\boldsymbol{x} \sim \mathcal{N}(0,\boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \succeq 0$, and let $\eta \in C^2(\mathbb{R})$ with $L_\eta := \sup_{z\in\mathbb{R}} |\eta''(z)| < \infty$. Fix $\boldsymbol{W}^A, \boldsymbol{W}^B \in \mathbb{R}^{m\times d}$. Suppose there exists $\beta \in [0,1)$ such that, for all $i \in [m]$,*

$$\|\mathbf{d}_i \odot \boldsymbol{w}_i^A\|_{\boldsymbol{\Sigma}} \;\le\; \beta\|\mathbf{f}_i\|_{\boldsymbol{\Sigma}}, \qquad \|\mathbf{d}_i \odot \boldsymbol{w}_i^B\|_{\boldsymbol{\Sigma}} \;\le\; \beta\|\mathbf{f}_i\|_{\boldsymbol{\Sigma}}. \tag{221}$$

Then

$$\sup_{\lambda\in[0,1]} \big\|\xi_{\boldsymbol{H}}(\lambda;\cdot)\big\|_{L^2(\mathbb{P};\mathbb{R}^m)} \;\le\; \frac{\sqrt{3}}{2}\,L_\eta\,\beta^2\|\mathbf{F}\boldsymbol{\Sigma}^{1/2}\|^2_F. \tag{222}$$

Proof. Fix $i \in [m]$ and write $\boldsymbol{r}_i^A := \mathbf{d}_i \odot \boldsymbol{w}_i^A$ and $\boldsymbol{r}_i^B := \mathbf{d}_i \odot \boldsymbol{w}_i^B$. Set

$$Z_i^A \;:=\; (\mathbf{f}_i + \boldsymbol{r}_i^A)^\top \boldsymbol{x}, \qquad Z_i^B \;:=\; (\mathbf{f}_i + \boldsymbol{r}_i^B)^\top \boldsymbol{x}, \qquad \Delta Z_i \;:=\; Z_i^B - Z_i^A. \tag{223}$$

For fixed $\boldsymbol{x}$, define $h_i(s) := \eta(Z_i^A + s\Delta Z_i)$ for $s \in [0,1]$. Then

$$\xi_i(\lambda;\boldsymbol{x}) \;=\; h_i(\lambda) - \big((1-\lambda)h_i(0) + \lambda h_i(1)\big). \tag{224}$$

We obtain

$$\xi_i(\lambda;\boldsymbol{x}) \;=\; (1-\lambda)\int_0^\lambda h_i'(s)\,ds - \lambda\int_\lambda^1 h_i'(s)\,ds \tag{225}$$

$$\;=\; (1-\lambda)\int_0^\lambda \big(h_i'(s) - h_i'(\lambda)\big)\,\mathrm{d}s \;-\; \lambda\int_\lambda^1 \big(h_i'(s) - h_i'(\lambda)\big)\,\mathrm{d}s. \tag{226}$$

Moreover, $h_i''(s) = \eta''(Z_i^A + s\Delta Z_i)(\Delta Z_i)^2$, and hence $|h_i''(s)| \le L_\eta|\Delta Z_i|^2$ for all $s \in [0,1]$. Therefore,

$$|\xi_i(\lambda;\boldsymbol{x})| \;\le\; L_\eta|\Delta Z_i|^2\left((1-\lambda)\int_0^\lambda (\lambda - s)\,\mathrm{d}s + \lambda\int_\lambda^1 (s-\lambda)\,\mathrm{d}s\right) \tag{227}$$

$$\;=\; \frac{L_\eta}{2}\lambda(1-\lambda)|\Delta Z_i|^2. \tag{228}$$

Thus

$$\|\xi_i(\lambda;\cdot)\|^2_{L^2(\mathbb{P})} \;\le\; \frac{L_\eta^2}{4}\lambda^2(1-\lambda)^2\mathbb{E}[(\Delta Z_i)^4]. \tag{229}$$

Since $\Delta Z_i = (\boldsymbol{r}_i^B - \boldsymbol{r}_i^A)^\top\boldsymbol{x}$ is centered Gaussian,

$$\mathbb{E}[(\Delta Z_i)^4] \;=\; 3\|\boldsymbol{r}_i^B - \boldsymbol{r}_i^A\|^4_{\boldsymbol{\Sigma}}. \tag{230}$$

By (221),

$$\|\boldsymbol{r}_i^B - \boldsymbol{r}_i^A\|_{\boldsymbol{\Sigma}} \;\le\; \|\boldsymbol{r}_i^B\|_{\boldsymbol{\Sigma}} + \|\boldsymbol{r}_i^A\|_{\boldsymbol{\Sigma}} \;\le\; 2\beta\|\mathbf{f}_i\|_{\boldsymbol{\Sigma}}. \tag{231}$$

Hence

$$\|\xi_i(\lambda;\cdot)\|^2_{L^2(\mathbb{P})} \;\le\; 12\,L_\eta^2\,\lambda^2(1-\lambda)^2\beta^4\|\mathbf{f}_i\|^4_{\boldsymbol{\Sigma}}. \tag{232}$$

Summing over i gives

$$\big\|\xi_{\boldsymbol{H}}(\lambda;\cdot)\big\|^2_{L^2(\mathbb{P};\mathbb{R}^m)} \;\le\; 12\,L_\eta^2\,\lambda^2(1-\lambda)^2\beta^4\sum_{i=1}^m \|\mathbf{f}_i\|^4_{\boldsymbol{\Sigma}}. \tag{233}$$

Taking $\sqrt{\cdot}$ and using $\lambda(1-\lambda) \le 1/4$ yields

$$\sup_{\lambda\in[0,1]} \left\|\xi_{\boldsymbol{H}}(\lambda;\cdot)\right\|_{L^2(\mathbb{P};\mathbb{R}^m)} \;\le\; \frac{\sqrt{3}}{2}\, L_\eta\, \beta^2 \left(\sum_{i=1}^{m} \|\mathbf{f}_i\|_{\boldsymbol{\Sigma}}^4\right)^{1/2}. \tag{234}$$

Finally,

$$\left(\sum_{i=1}^{m} \|\mathbf{f}_i\|_{\boldsymbol{\Sigma}}^4\right)^{1/2} \;\le\; \sum_{i=1}^{m} \|\mathbf{f}_i\|_{\boldsymbol{\Sigma}}^2 \;=\; \|\mathbf{F}\boldsymbol{\Sigma}^{1/2}\|_F^2, \tag{235}$$

which proves the claim. □

Proposition D.3 (LMC bound by chord deviation)**.** *Consider a dataset $\mathcal{X}\times\mathcal{Y} \ni (\boldsymbol{x},\boldsymbol{y}) \sim \mathbb{P}$, a neural network $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}^{d_{\text{out}}}$, and define the population loss $\mathcal{L}(\boldsymbol{\theta}) := \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}[\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}),\boldsymbol{y})] < \infty$. Assume that for every $\boldsymbol{y}$, $\boldsymbol{z} \mapsto \ell(\boldsymbol{z},\boldsymbol{y})$ is convex and L_ℓ-Lipschitz w.r.t. $\|\cdot\|_2$. Fix two parameter vectors $\boldsymbol{\theta}^A, \boldsymbol{\theta}^B$ and define the linear path $\boldsymbol{\theta}(\lambda) := (1-\lambda)\boldsymbol{\theta}^A + \lambda\boldsymbol{\theta}^B$, $\lambda \in [0,1]$. Define further*

$$\xi_f(\lambda;\boldsymbol{x}) := f_{\boldsymbol{\theta}(\lambda)}(\boldsymbol{x}) - \big((1-\lambda)f_{\boldsymbol{\theta}^A}(\boldsymbol{x}) + \lambda f_{\boldsymbol{\theta}^B}(\boldsymbol{x})\big). \tag{236}$$

Then, for all $\lambda \in [0,1]$, we obtain

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}(\lambda)) &- \big((1-\lambda)\mathcal{L}(\boldsymbol{\theta}^A) + \lambda\mathcal{L}(\boldsymbol{\theta}^B)\big)\\ &\le\; L_\ell \left\|\xi_f(\lambda;\cdot)\right\|_{L^2(\mathbb{P}_{\boldsymbol{x}},\mathbb{R}^{d_{\text{out}}})}.\end{aligned} \tag{237}$$

Proof. Fix $\lambda \in [0,1]$. By convexity of $\boldsymbol{z} \mapsto \ell(\boldsymbol{z},\boldsymbol{y})$, for every $(\boldsymbol{x},\boldsymbol{y})$,

$$\ell\big((1-\lambda)f_{\boldsymbol{\theta}^A}(\boldsymbol{x}) + \lambda f_{\boldsymbol{\theta}^B}(\boldsymbol{x}),\boldsymbol{y}\big) \;\le\; (1-\lambda)\ell(f_{\boldsymbol{\theta}^A}(\boldsymbol{x}),\boldsymbol{y}) + \lambda\ell(f_{\boldsymbol{\theta}^B}(\boldsymbol{x}),\boldsymbol{y}). \tag{238}$$

Moreover, by the definition of $\xi_f(\lambda,\boldsymbol{x})$ and since $\ell(\cdot,\boldsymbol{y})$ is L_ℓ-Lipschitz w.r.t. $\|\cdot\|_2$,

$$\ell(f_{\boldsymbol{\theta}(\lambda)}(\boldsymbol{x}),\boldsymbol{y}) \;\le\; \ell\big((1-\lambda)f_{\boldsymbol{\theta}^A}(\boldsymbol{x}) + \lambda f_{\boldsymbol{\theta}^B}(\boldsymbol{x}),\boldsymbol{y}\big) + L_\ell \left\|\xi_f(\lambda,\boldsymbol{x})\right\|_2. \tag{239}$$

Combining (238) and (239) and taking expectations w.r.t. $(\boldsymbol{x},\boldsymbol{y}) \sim \mathbb{P}$ gives

$$\mathcal{L}(\boldsymbol{\theta}(\lambda)) \;\le\; (1-\lambda)\mathcal{L}(\boldsymbol{\theta}^A) + \lambda\mathcal{L}(\boldsymbol{\theta}^B) + L_\ell\, \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\boldsymbol{x}}}\left[\|\xi_f(\lambda,\boldsymbol{x})\|_2\right]. \tag{240}$$

Finally, by Jensen's inequality,

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\boldsymbol{x}}}\left[\|\xi_f(\lambda,\boldsymbol{x})\|_2\right] \;\le\; \left(\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\boldsymbol{x}}}\left[\|\xi_f(\lambda,\boldsymbol{x})\|_2^2\right]\right)^{1/2} \;=\; \|\xi_f(\lambda,\cdot)\|_{L^2(\mathbb{P}_{\boldsymbol{x}},\mathbb{R}^{d_{\text{out}}})}. \tag{241}$$

Substituting this into (240) and rearranging yields the claim. □

E. Approximate Subspace Support and Relaxed Realization Costs

In this section, we make rigorous how the exact subspace support model (Assumptions C.1 and C.2) can be interpreted as a low-rank approximation of a general data distribution: Namely, after projection on the top k principal components, exact realizability in the projected model implies approximate realizability on the original distribution, with the error controlled by the discarded second-moment tail energy. In this sense, the results from § 3 should not only be read as applicable to this idealized model, but as a tractable approximation to the observable realizability structure when the input distribution to a layer is only approximately low-rank. We also explain a natural relaxed variant of the realization cost framework from § 3, which does not require exact realizability on the data support, and show that the hard realization cost from § 3 arises as its ridgeless limit.

E.1. Approximate Subspace Support

Let the input $\boldsymbol{x}$ to a layer be drawn from a distribution $\mathbb{P}$ on $\mathbb{R}^d$ with finite second moment $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\|\boldsymbol{x}\|_2^2 < \infty$, and define

$$\boldsymbol{\Sigma} \;:=\; \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}[\boldsymbol{x}\boldsymbol{x}^\top] \;\succeq\; 0. \tag{242}$$

Let $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ be an eigendecomposition, i.e., $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_1, \dots, \lambda_d)$ with $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ and $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_d$. For a target rank k, let $\boldsymbol{U}_k \in \mathbb{R}^{d\times k}$ be the top k eigenvectors and define the tail energy

$$\tau_k \;:=\; \sum_{j>k} \lambda_j. \tag{243}$$

In this context, Assumption C.1 for a given k means precisely that $\tau_k = 0$ with $\mathcal{U} := \operatorname{span} \boldsymbol{U}_k$, and the non-degeneracy assumption (Assumption C.2) is equivalent to $\tau_{k-1} > 0$ for injective η. More generally, one can show the following:

Lemma E.1 (Subspace support and second-moment tail energy)**.** *Assume $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\|\boldsymbol{x}\|_2^2 < \infty$. Then, $\tau_k = 0$ iff $\mathbb{P}(\boldsymbol{x} \in \mathrm{im}(\boldsymbol{U}_k)) = 1$. Moreover, setting $\boldsymbol{P}_k := \boldsymbol{U}_k\boldsymbol{U}_k^\top$, the mean-squared residual outside the top-k subspace is $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\left[\left\|\boldsymbol{x} - \boldsymbol{P}_k\boldsymbol{x}\right\|_2^2\right] = \tau_k$.*

Proof. One can directly calculate

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\left[\|\boldsymbol{x} - \boldsymbol{P}_k\boldsymbol{x}\|_2^2\right] \;=\; \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}[\boldsymbol{x}^\top(\boldsymbol{I} - \boldsymbol{P}_k)\boldsymbol{x}] \;=\; \mathrm{tr}((\boldsymbol{I} - \boldsymbol{P}_k)\boldsymbol{\Sigma}) \;=\; \sum_{j>k}\lambda_j \;=\; \tau_k. \tag{244}$$

If $\tau_k = 0$, then $\|\boldsymbol{x} - \boldsymbol{P}_k\boldsymbol{x}\|_2^2 \geq 0$ has expectation 0, hence $\boldsymbol{x} = \boldsymbol{P}_k\boldsymbol{x}$ a.s., which means $\boldsymbol{x} \in \mathrm{im}(\boldsymbol{U}_k)$ a.s. Conversely, if $\boldsymbol{x} \in \mathrm{im}(\boldsymbol{U}_k)$ a.s., then $\boldsymbol{x} = \boldsymbol{P}_k\boldsymbol{x}$ a.s., so $\tau_k = 0$. □

Lem. E.1 shows that τ_k is the residual incurred by projecting $\mathbb{P}$ onto its top-k principal subspace. This makes the projected distribution $\mathbb{P}_k := (\boldsymbol{P}_k)_*\mathbb{P}$ a natural low-rank surrogate for the original input distribution. One may therefore ask whether coincidence of neuron functions on this projected model still implies approximate coincidence on the original distribution. The next proposition answers this affirmatively: for Lipschitz neuron functions, the resulting $L^2(\mathbb{P})$ error can be controlled by the second-moment tail energy τ_k. In this sense, the subspace support model from § 3 can also be seen as a tractable approximation whenever the input distribution is only approximately low-rank.

Proposition E.2 (Projected coincidence implies approximate coincidence)**.** *Assume $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\|\boldsymbol{x}\|_2^2 < \infty$. Let $\boldsymbol{h}_1, \boldsymbol{h}_2 : \mathbb{R}^d \to \mathbb{R}$ be L_1- and L_2-Lipschitz, respectively, i.e., $|\boldsymbol{h}_i(\boldsymbol{x}) - \boldsymbol{h}_i(\boldsymbol{y})| \leq L_i\|\boldsymbol{x} - \boldsymbol{y}\|_2$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $i \in \{1, 2\}$. Assume that $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ coincide on the projected distribution $\mathbb{P}_k$, i.e. $\boldsymbol{h}_1(\boldsymbol{P}_k\boldsymbol{x}) = \boldsymbol{h}_2(\boldsymbol{P}_k\boldsymbol{x})$ for $\mathbb{P}$-a.e. $\boldsymbol{x}$. Then,*

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\left[(\boldsymbol{h}_1(\boldsymbol{x}) - \boldsymbol{h}_2(\boldsymbol{x}))^2\right] \;\leq\; 2(L_1^2 + L_2^2)\,\tau_k. \tag{245}$$

This means that if τ_k is small, any two neuron functions that agree on the top-k projected model are close in $L^2(\mathbb{P})$ on the original distribution.

Proof. Since $\boldsymbol{h}_1(\boldsymbol{P}_k\boldsymbol{x}) = \boldsymbol{h}_2(\boldsymbol{P}_k\boldsymbol{x})$ for $\mathbb{P}$-a.e. $\boldsymbol{x}$,

$$\boldsymbol{h}_1(\boldsymbol{x}) - \boldsymbol{h}_2(\boldsymbol{x}) = \big(\boldsymbol{h}_1(\boldsymbol{x}) - \boldsymbol{h}_1(\boldsymbol{P}_k\boldsymbol{x})\big) + \big(\boldsymbol{h}_2(\boldsymbol{P}_k\boldsymbol{x}) - \boldsymbol{h}_2(\boldsymbol{x})\big). \tag{246}$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$, we obtain

$$\begin{aligned}(\boldsymbol{h}_1(\boldsymbol{x}) - \boldsymbol{h}_2(\boldsymbol{x}))^2 \;&\leq\; 2\big(\boldsymbol{h}_1(\boldsymbol{x}) - \boldsymbol{h}_1(\boldsymbol{P}_k\boldsymbol{x})\big)^2 + 2\big(\boldsymbol{h}_2(\boldsymbol{x}) - \boldsymbol{h}_2(\boldsymbol{P}_k\boldsymbol{x})\big)^2 \\ &\leq\; 2(L_1^2 + L_2^2)\|\boldsymbol{x} - \boldsymbol{P}_k\boldsymbol{x}\|_2^2.\end{aligned} \tag{247}$$

Taking expectations and using Lem. E.1,

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\big[(\boldsymbol{h}_1(\boldsymbol{x})-\boldsymbol{h}_2(\boldsymbol{x}))^2\big] \;\le\; 2(L_1^2+L_2^2)\,\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\|\boldsymbol{x}-\boldsymbol{P}_k\boldsymbol{x}\|_2^2 \;=\; 2(L_1^2+L_2^2)\tau_k, \tag{248}$$

which proves (245). □

In particular, if η is L-Lipschitz and $\boldsymbol{h}_i(\boldsymbol{x}) = \eta(\boldsymbol{w}_i^\top \boldsymbol{x})$ for $\boldsymbol{w}_i \in \mathbb{R}^d$, $i \in \{1,2\}$, then each $\boldsymbol{h}_i$ is $L\|\boldsymbol{w}_i\|_2$-Lipschitz (by Cauchy-Schwarz). Hence, Prop. E.2 yields

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\big[(\boldsymbol{h}_1(\boldsymbol{x})-\boldsymbol{h}_2(\boldsymbol{x}))^2\big] \;\le\; 2L^2(\|\boldsymbol{w}_1\|_2^2+\|\boldsymbol{w}_2\|_2^2)\tau_k \tag{249}$$

whenever $\boldsymbol{h}_1(\boldsymbol{P}_k\boldsymbol{x}) = \boldsymbol{h}_2(\boldsymbol{P}_k\boldsymbol{x})$ for $\mathbb{P}$-a.e. $\boldsymbol{x}$.

E.2. Relaxed Realization Costs

In the previous subsection, we have seen in what sense the subspace support model can be viewed as an approximation to a general input distribution. For realization costs, a complementary question is how one should formulate the cost when exact realizability is not imposed. A natural choice is to relax the hard constraint by penalizing squared $L^2(\mathbb{P})$ approximation error and weight norm. We record it here to show that the realization cost from § 3 arises as its ridgeless limit.

Just as in § 3, consider a single asymmetric layer in isolation, and define $\boldsymbol{H}(\boldsymbol{W};\boldsymbol{x}) := \eta((\mathbf{F}+\mathbf{D}\odot\boldsymbol{W})\boldsymbol{x}) \in \mathbb{R}^m$, where $\boldsymbol{W} \in \mathbb{R}^{m\times d}$, $\boldsymbol{x} \in \mathbb{R}^d$, $\mathbf{F}, \mathbf{D} \in \mathbb{R}^{m\times d}$ are fixed architectural components, and $\eta : \mathbb{R} \to \mathbb{R}$ acts elementwise.

Definition E.3 (Relaxed L^2 realization cost)**.** Fix neuron i and $\beta > 0$. For any $\boldsymbol{h} \in L^2(\mathbb{P})$, define

$$\|\boldsymbol{h}\|^2_{\mathcal{H}_i(\mathbb{P}),\beta} \;=\; \|\boldsymbol{h}\|^2_{\mathcal{H}_i,\beta} \;:=\; \inf_{\boldsymbol{w}\in\mathbb{R}^d}\Big\{\beta^{-1}\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\big[(\boldsymbol{h}(\boldsymbol{x})-\boldsymbol{h}_i(\boldsymbol{w};\boldsymbol{x}))^2\big]+\|\boldsymbol{w}\|_2^2\Big\}. \tag{250}$$

In cases where the distribution $\mathbb{P}$ might vary, we will make the dependence explicit through writing $\mathcal{H}_i(\mathbb{P})$. It is straightforward to see that for Lipschitz η and any $\boldsymbol{h} \in L^2(\mathbb{P})$, $\|\boldsymbol{h}\|^2_{\mathcal{H}_i,\beta} < \infty$ is finite. A canonical family of targets we will look at are functions of the form $\boldsymbol{h}_{\boldsymbol{u}}(\boldsymbol{x}) := \eta(\boldsymbol{u}^\top\boldsymbol{x})$, $\boldsymbol{u} \in \mathbb{R}^d$, i.e., precisely *realizable* target features, since this is precisely the setting in § 3.

Theorem E.4 (Explicit upper bound via a quadratic surrogate)**.** *Assume η is L-Lipschitz. Fix neuron i and define $\mathbf{D}_i := \mathrm{Diag}(\mathbf{d}_i) \in \mathbb{R}^{d\times d}$. For any $\boldsymbol{u} \in \mathbb{R}^d$,*

$$\|\boldsymbol{h}_{\boldsymbol{u}}\|^2_{\mathcal{H}_i,\beta} \;\le\; \inf_{\boldsymbol{w}\in\mathbb{R}^d}\left\{\beta^{-1}L^2\,\|\boldsymbol{u}-\mathbf{f}_i-\mathbf{D}_i\boldsymbol{w}\|^2_{\boldsymbol{\Sigma}}+\|\boldsymbol{w}\|_2^2\right\}. \tag{251}$$

Moreover, the infimum in (251) has a unique minimizer

$$\boldsymbol{w}_i^\star(\boldsymbol{u}) = \Big(\beta\,\boldsymbol{I}+L^2\,\mathbf{D}_i\boldsymbol{\Sigma}\mathbf{D}_i\Big)^{-1}L^2\,\mathbf{D}_i\boldsymbol{\Sigma}(\boldsymbol{u}-\mathbf{f}_i), \tag{252}$$

and the corresponding minimized value equals

$$\frac{L^2}{\beta}\|\boldsymbol{u}-\mathbf{f}_i\|^2_{\boldsymbol{\Sigma}} - \frac{L^4}{\beta^2}\,(\boldsymbol{u}-\mathbf{f}_i)^\top\boldsymbol{\Sigma}\mathbf{D}_i\left(\boldsymbol{I}+\frac{L^2}{\beta}\mathbf{D}_i\boldsymbol{\Sigma}\mathbf{D}_i\right)^{-1}\mathbf{D}_i\boldsymbol{\Sigma}(\boldsymbol{u}-\mathbf{f}_i). \tag{253}$$

Proof. Fix $\boldsymbol{w} \in \mathbb{R}^d$ and abbreviate

$$\ell_{\boldsymbol{u}}(\boldsymbol{x}) \;:=\; \boldsymbol{u}^\top\boldsymbol{x}, \qquad \ell_i(\boldsymbol{w};\boldsymbol{x}) \;:=\; (\mathbf{f}_i+\mathbf{D}_i\boldsymbol{w})^\top\boldsymbol{x}. \tag{254}$$

Since η is L-Lipschitz,

$$(\boldsymbol{h}_{\boldsymbol{u}}(\boldsymbol{x})-\boldsymbol{h}_i(\boldsymbol{w};\boldsymbol{x}))^2 \;\le\; L^2\,(\ell_{\boldsymbol{u}}(\boldsymbol{x})-\ell_i(\boldsymbol{w};\boldsymbol{x}))^2. \tag{255}$$

Taking expectations yields

$$\mathbb{E}[(\boldsymbol{h}_{\boldsymbol{u}}(\boldsymbol{x})-\boldsymbol{h}_i(\boldsymbol{w};\boldsymbol{x}))^2] \;\le\; L^2\,\|\boldsymbol{u}-\mathbf{f}_i-\mathbf{D}_i\boldsymbol{w}\|^2_{\boldsymbol{\Sigma}}. \tag{256}$$

Plugging into Def. E.3 and then taking the infimum over $\boldsymbol{w}$ proves (251). To compute the infimum, define $\alpha := L^2/\beta$ and consider the strictly convex quadratic

$$J(\boldsymbol{w}) \;:=\; \|\boldsymbol{w}\|_2^2+\alpha\,(\boldsymbol{u}-\mathbf{f}_i-\mathbf{D}_i\boldsymbol{w})^\top\boldsymbol{\Sigma}(\boldsymbol{u}-\mathbf{f}_i-\mathbf{D}_i\boldsymbol{w}). \tag{257}$$

Differentiating gives

$$\nabla J(\boldsymbol{w}) \;=\; 2\boldsymbol{w}-2\alpha\,\mathbf{D}_i\boldsymbol{\Sigma}(\boldsymbol{u}-\mathbf{f}_i-\mathbf{D}_i\boldsymbol{w}) \;=\; 2\Big(\boldsymbol{I}+\alpha\,\mathbf{D}_i\boldsymbol{\Sigma}\mathbf{D}_i\Big)\boldsymbol{w}-2\alpha\,\mathbf{D}_i\boldsymbol{\Sigma}(\boldsymbol{u}-\mathbf{f}_i). \tag{258}$$

Since $\boldsymbol{I} + \alpha\, \mathbf{D}_i \boldsymbol{\Sigma} \mathbf{D}_i \succ 0$, setting $\nabla J(\boldsymbol{w}) = 0$ yields the unique minimizer (252). It remains to compute the minimum value. Write $\boldsymbol{r} := \boldsymbol{u} - \mathbf{f}_i$ and $\boldsymbol{A} := \boldsymbol{I} + \alpha\, \mathbf{D}_i \boldsymbol{\Sigma} \mathbf{D}_i \succ 0$. Expanding J yields

$$J(\boldsymbol{w}) \;=\; \boldsymbol{w}^\top \boldsymbol{A} \boldsymbol{w} - 2\alpha\, \boldsymbol{w}^\top \mathbf{D}_i \boldsymbol{\Sigma} \boldsymbol{r} + \alpha\, \boldsymbol{r}^\top \boldsymbol{\Sigma} \boldsymbol{r}. \tag{259}$$

Let $\boldsymbol{c} := \alpha\, \mathbf{D}_i \boldsymbol{\Sigma} \boldsymbol{r}$. Then, $J(\boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{A} \boldsymbol{w} - 2\boldsymbol{w}^\top \boldsymbol{c} + \alpha\, \boldsymbol{r}^\top \boldsymbol{\Sigma} \boldsymbol{r}$. Since $\boldsymbol{A} \succ 0$, we have

$$\boldsymbol{w}^\top \boldsymbol{A} \boldsymbol{w} - 2\boldsymbol{w}^\top \boldsymbol{c} \;=\; (\boldsymbol{w} - \boldsymbol{A}^{-1}\boldsymbol{c})^\top \boldsymbol{A} (\boldsymbol{w} - \boldsymbol{A}^{-1}\boldsymbol{c}) - \boldsymbol{c}^\top \boldsymbol{A}^{-1} \boldsymbol{c}, \tag{260}$$

which implies that the minimum is attained at $\boldsymbol{w}^\star := \boldsymbol{A}^{-1}\boldsymbol{c}$ (which coincides with (252)) and equals

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} J(\boldsymbol{w}) \;=\; \alpha\, \boldsymbol{r}^\top \boldsymbol{\Sigma} \boldsymbol{r} - \boldsymbol{c}^\top \boldsymbol{A}^{-1} \boldsymbol{c} \;=\; \alpha\, \boldsymbol{r}^\top \boldsymbol{\Sigma} \boldsymbol{r} - \alpha^2\, \boldsymbol{r}^\top \boldsymbol{\Sigma} \mathbf{D}_i (\boldsymbol{I} + \alpha\, \mathbf{D}_i \boldsymbol{\Sigma} \mathbf{D}_i)^{-1} \mathbf{D}_i \boldsymbol{\Sigma} \boldsymbol{r}, \tag{261}$$

which is exactly (253). □

If η is bi-Lipschitz, one can prove a lower bound in a similar way. We will now show that the $\beta \to 0$ (ridgeless) limit recovers the hard realization cost from § 3, where non-realizable targets have infinite cost.

Theorem E.5 (Ridgeless limit)**.** *Assume η is Lipschitz and $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\|\boldsymbol{x}\|_2^2 < \infty$. Fix neuron i and let*

$$\|\boldsymbol{h}\|^2_{\mathcal{H}_i(\mathbb{P})} \;=\; \|\boldsymbol{h}\|^2_{\mathcal{H}_i} \;:=\; \inf\left\{\|\boldsymbol{w}\|_2^2 \;\middle|\; \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\big[(\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{h}_i(\boldsymbol{w};\boldsymbol{x}))^2\big] = 0\right\} \in [0,\infty] \tag{262}$$

with the convention $\inf \varnothing := +\infty$*. Then for any $\boldsymbol{h} \in L^2(\mathbb{P})$,*

$$\lim_{\beta\to 0} \|\boldsymbol{h}\|^2_{\mathcal{H}_i,\beta} \;=\; \|\boldsymbol{h}\|^2_{\mathcal{H}_i}. \tag{263}$$

Proof. Write $f(\boldsymbol{w}) := \mathbb{E}[(\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{h}_i(\boldsymbol{w};\boldsymbol{x}))^2] \geq 0$ and $\phi(\beta) := \inf_{\boldsymbol{w}}\{f(\boldsymbol{w}) + \beta\|\boldsymbol{w}\|_2^2\}$ so that $\|\boldsymbol{h}\|^2_{\mathcal{H}_i,\beta} = \beta^{-1}\phi(\beta)$ by Def. E.3.

First, suppose $\|\boldsymbol{h}\|^2_{\mathcal{H}_i} = +\infty$. We show $\|\boldsymbol{h}\|^2_{\mathcal{H}_i,\beta} \to +\infty$. Assume for contradiction that there exists $0 < \beta_n \to 0$ and $C < \infty$ with $\|\boldsymbol{h}\|^2_{\mathcal{H}_i,\beta_n} \leq C$ for all n. Then there exist $\boldsymbol{w}_n$ such that

$$f(\boldsymbol{w}_n) + \beta_n \|\boldsymbol{w}_n\|_2^2 \;\leq\; \beta_n (C+1). \tag{264}$$

In particular, this implies that $\|\boldsymbol{w}_n\|_2^2 \leq C+1$ and $f(\boldsymbol{w}_n) \to 0$ as $n \to \infty$. Since $(\boldsymbol{w}_n)_n$ is bounded, we can extract a convergent subsequence $(\boldsymbol{w}_{n_k})_k$ with $\boldsymbol{w}_{n_k} \to \boldsymbol{w}^*$ as $k \to \infty$. By Lipschitzness of η and $\mathbb{E}\|\boldsymbol{x}\|_2^2 < \infty$, the map $\boldsymbol{w} \mapsto \boldsymbol{h}_i(\boldsymbol{w};\cdot)$ is continuous from $(\mathbb{R}^d, \|\cdot\|_2)$ to $L^2(\mathbb{P})$, hence $f(\boldsymbol{w}_{n_k}) \to f(\boldsymbol{w}^*)$ and therefore $f(\boldsymbol{w}^*) = 0$. This contradicts $\|\boldsymbol{h}\|^2_{\mathcal{H}_i} = +\infty$.

Assume now $\|\boldsymbol{h}\|^2_{\mathcal{H}_i} < \infty$. Then, for every $\beta > 0$,

$$\phi(\beta) \;=\; \inf_{\boldsymbol{w}}\{f(\boldsymbol{w}) + \beta\|\boldsymbol{w}\|_2^2\} \;\leq\; \inf_{\boldsymbol{w}: f(\boldsymbol{w})=0}\{\beta\|\boldsymbol{w}\|_2^2\} \;=\; \beta\, \|\boldsymbol{h}\|^2_{\mathcal{H}_i}, \tag{265}$$

and

$$\limsup_{\beta\to 0}\; \beta^{-1}\phi(\beta) \;\leq\; \|\boldsymbol{h}\|^2_{\mathcal{H}_i}. \tag{266}$$

For the reverse inequality, let $0 < \beta_n \to 0$ and pick $\boldsymbol{w}_n$ such that $f(\boldsymbol{w}_n) + \beta_n\|\boldsymbol{w}_n\|_2^2 \leq \phi(\beta_n) + \beta_n/n$. By the previous upper bound, $\phi(\beta_n) \leq \beta_n(\|\boldsymbol{h}\|^2_{\mathcal{H}_i} + 1)$ for all large enough n, hence $(\|\boldsymbol{w}_n\|_2)_n$ is bounded and $f(\boldsymbol{w}_n) \to 0$. Similar as above, any limit point $\boldsymbol{w}^*$ of a convergent subsequence satisfies $f(\boldsymbol{w}^*) = 0$ as $n \to \infty$. Therefore $\|\boldsymbol{w}^*\|_2^2 \geq \|\boldsymbol{h}\|^2_{\mathcal{H}_i}$ and thus $\liminf_{n\to\infty} \|\boldsymbol{w}_n\|_2^2 \geq \|\boldsymbol{h}\|^2_{\mathcal{H}_i}$. Since $\beta_n^{-1}\phi(\beta_n) \geq \|\boldsymbol{w}_n\|_2^2 - 1/n$, we can conclude

$$\liminf_{\beta\to 0}\; \beta^{-1}\phi(\beta) \;\geq\; \|\boldsymbol{h}\|^2_{\mathcal{H}_i}. \tag{267}$$

Combining (266) and (267) gives the claim. □

When $\eta = \mathrm{id}$, the relaxed cost in Def. E.3 reduces to a strictly convex quadratic ridge objective in $\boldsymbol{w}$, so the minimizer and optimum admit explicit closed forms for every $\beta > 0$. For general nonlinear $\eta \neq \mathrm{id}$, the same objective is typically nonconvex in $\boldsymbol{w}$ and no simple closed-form solution is available, so apart from the ridgeless limit, one would have to work with tractable surrogates such as Thm. E.4. By contrast, in the ridgeless limit the problem simplifies conceptually: Thm. E.5 shows that as $\beta \to 0$, one recovers the realization cost from § 3, which admits the explicit closed form (5). At the same time, because the limit enforces exact realizability, it is singular under perturbations of the *support*, i.e., even a small amount of mass outside the top-k subspace can destroy feasibility on the original distribution and make the hard cost jump to $+\infty$.

For fixed $\beta > 0$, however, the relaxed objective remains stable under such perturbations. The next corollary records the corresponding quantitative upper bound.

Corollary E.6 (Projected realizability yields small relaxed cost)**.** *Fix neuron i and $\beta > 0$, and let $\boldsymbol{h} \in L^2(\mathbb{P})$. Assume that the hard realization cost of $\boldsymbol{h}$ with respect to $\mathbb{P}_k$ is finite and attained at some $\boldsymbol{w}^* \in \mathbb{R}^d$, i.e.*

$$\boldsymbol{h}(\boldsymbol{P}_k\boldsymbol{x}) = \boldsymbol{h}_i(\boldsymbol{w}^*; \boldsymbol{P}_k\boldsymbol{x}) \quad \text{for } \mathbb{P}\text{-a.e. } \boldsymbol{x}, \qquad \|\boldsymbol{w}^*\|_2 = \|\boldsymbol{h}\|_{\mathcal{H}_i(\mathbb{P}_k)}. \tag{268}$$

Assume also that $\boldsymbol{h}$, $\boldsymbol{h}_i(\boldsymbol{w}^*; \cdot)$ are $L_{\boldsymbol{h}}$- and $L_i(\boldsymbol{w}^*)$-Lipschitz, respectively. Then,

$$\|\boldsymbol{h}\|^2_{\mathcal{H}_i(\mathbb{P}),\beta} \;\leq\; \|\boldsymbol{h}\|^2_{\mathcal{H}_i(\mathbb{P}_k)} + 2(L^2_{\boldsymbol{h}} + L_i(\boldsymbol{w}^*)^2)\,\beta^{-1}\tau_k. \tag{269}$$

Proof. By Def. E.3, evaluating the infimum at $\boldsymbol{w}^*$ and applying Prop. E.2 gives

$$\|\boldsymbol{h}\|^2_{\mathcal{H}_i(\mathbb{P}),\beta} \;\leq\; \beta^{-1}\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\big[(\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{h}_i(\boldsymbol{w}^*; \boldsymbol{x}))^2\big] + \|\boldsymbol{w}^*\|_2^2 \;\leq\; 2(L^2_{\boldsymbol{h}} + L_i(\boldsymbol{w}^*)^2)\,\beta^{-1}\tau_k + \|\boldsymbol{w}^*\|_2^2, \tag{270}$$

which yields the claim since $\|\boldsymbol{w}^*\|_2 = \|\boldsymbol{h}\|_{\mathcal{H}_i(\mathbb{P}_k)}$. □

Corollary E.6 links the approximate subspace support from § E.1 and the relaxed realization costs we introduced in § E.2. Namely, the first term on the r.h.s. of (269) is the hard realization cost on the projected model $\mathbb{P}_k$. The second term is the approximation error incurred when passing back to the original distribution. Thus, for fixed $\beta > 0$, the projected subspace support model yields a quantitative upper bound on the relaxed realization cost under the true distribution.

F. Additional Figures

F.1. Visualization of Effective Function Classes

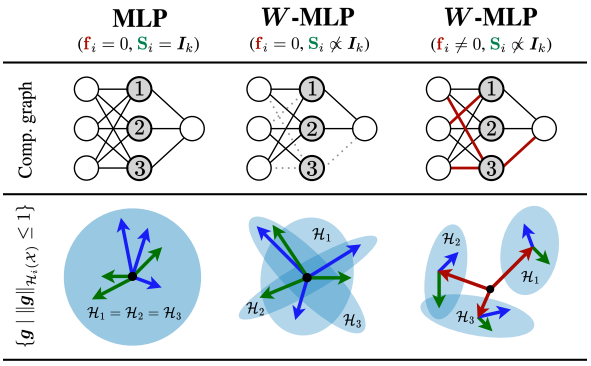


Figure 6. Eff. fct. classes and learned features (**run A**/**run B**) w/ varying levels of symmetry. *Col. 1:* In a fully symmetric MLP, each neuron can implement the same functions on $\mathcal{X}$. *Col. 2:* Pruning weights affects $\mathbf{S}_i$ and can introduce anisotropy. *Col. 3:* Fixed weights via $\mathbf{F}$ introduce functional biases.

F.2. LMC Interpolation for MLPs and ResNets

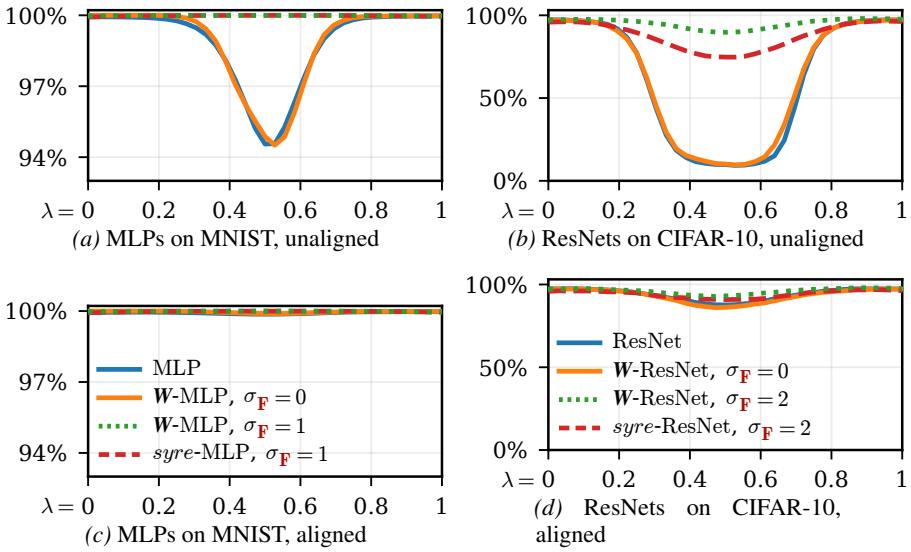


Figure 7. Aligned and unaligned training accuracy LMC interpolation for standard models and asymmetric models (average of 8 model pairs each). Alignment using activation matching.

G. Experimental Setup

G.1. MLPs

For MLP experiments, we use MLPs with 4 layers, input dimension 784 (MNIST input dimension), hidden dimensions 512, layer norms, and GELU activation functions. For $\boldsymbol{W}$-asymmetric MLPs ($\boldsymbol{W}$-MLPs), following Lim et al. (2024b), we fix 64 entries per linear layer row (= neuron) for the first three layers, and 256 entries per row for the final layer. Fixed entries are randomly sampled, making use of a mask seed that ensures the same mask in different model instances. Whenever we perform parameter interpolation between models, the models have the same mask (which is considered a fixed part of the architecture).

For the fixed weight scale $\sigma_{\mathbf{F}}$, we use $\sigma_{\mathbf{F}} = 1$ as the default choice for effective symmetry breaking $\boldsymbol{W}$-MLPs, and $\sigma_{\mathbf{F}} = 0$ for the variant that does not exhibit effective symmetry breaking, as well as a range of values for $\sigma_{\mathbf{F}}$ between 0.0 and 5.0 for one $\sigma_{\mathbf{F}}$ sweep experiment. In contrast to Lim et al. (2024b), we always apply the same value of $\sigma_{\mathbf{F}}$ at all layers.

We train the MLPs, $\boldsymbol{W}$-MLPs, and *syre*-MLPs on MNIST for 100 epochs with AdamW (Loshchilov & Hutter, 2019), a learning rate of 0.001, and a weight decay of 0.01.

G.2. ResNets

For our ResNet experiments, we use ResNet20 networks with $8\times$ width multiplier (following the setup of Lim et al. (2024b)), ReLU activation functions, and group norms. The architecture contains an initial convolution ($3 \to 128$ channels), three block groups ($128, 128 \to 256, 256 \to 512$ channels), and a final linear layer. For $\boldsymbol{W}$-asymmetric ResNets, in each convolution, a number of fixed weights are introduced per input channel (spread randomly over the output channels and kernel dimensions). The number of fixed parameters varies per block group, with the shortcut convolution in block groups 2 and 3 having a separate value for the number of fixed values. Table 3 gives the values across the network.

Table 3. Number of fixed parameters per layer.

Layer	# fixed
Initial convolution	12
Block group 1	108
Block group 2	162
Shortcut	18
Block group 3	216
Shortcut	24
Final linear	24

The default fixed weight scale in our $\boldsymbol{W}$-ResNet experiments is $\sigma_{\mathbf{F}} = 2$ for the effective symmetry breaking setting, and $\sigma_{\mathbf{F}} = 0$ for the ineffective symmetry breaking setting. We train the ResNets on CIFAR-10 for 100 epochs using AdamW (Loshchilov & Hutter, 2019) with weight decay of 0.01, and learning rate warm up from 0.0001 to 0.01 in the first 20 epochs ($\gamma = 1.259$).

G.3. Datasets and Augmentations

We use MNIST and CIFAR-10 datasets with a random 80%/10%/10% train/val/test split. On MNIST, we only use a normalization transform. On CIFAR-10, we additionally use a RandomHorizontalFlip transform, and a RandomCrop transform with 4 pixels of padding and a 32x32 output size.

G.4. Activation Matching

We implement activation matching as described in § 5. For MLPs, this is straightforward: we forward the full dataset through the models and record intermediate activations $\boldsymbol{Z}^A, \boldsymbol{Z}^B$ at each layer, which we use to compute the similarity matrix $\boldsymbol{S}^{\text{act}}$. For ResNets, some subtleties are involved. The permutable dimension is the convolution channel dimension, which takes the role of the hidden representation dimension in $\boldsymbol{Z}$. We could average over the pixel dimensions to have $\boldsymbol{Z} \in \mathbb{R}^{n \times d}$. However we find it beneficial for LMC to treat each pixel as an independent data sample, producing $\boldsymbol{Z} \in \mathbb{R}^{(nk^2) \times d}$. Furthermore, in ResNets, the residual connections cause certain activation permutations to affect several layer inputs: For example, the input

to block 1 in block group 1 is also indirectly fed into block 2 via the additive residual connection, without anything yielding an independent permutation symmetry in between. Each block group exhibits one such a global permutation that affects several layer inputs, and hence should be estimated jointly, yielding $\boldsymbol{Z} \in \mathbb{R}^{(nk^2c)\times d}$ (for $c = 4$ in our case). Besides, each block admits an inner permutation symmetry between its two convolution layers, which can be estimated independently. Due to memory constraints, we perform activation matching on 10% of our CIFAR-10 training set, which is sufficient to yield an alignment of low barrier. For our plots in § 5, we average the activation matching objective across all layers.

G.5. Coherence Experiment

We record details for the experiment isolating dependence on the coherence of the data subspace (see Fig. 4) in a teacher-student setup. The latent input variable is $(a, b) \sim \mathcal{N}(0, \boldsymbol{I}_2)$ and we aim to learn the teacher $f(a, b) := \mathrm{ReLU}(a + b)$. We sample 1024 training examples and 2048 test examples from this latent distribution, which are used across all coherence levels. The latent (a, b) is embedded into the ambient space $\mathbb{R}^d$, $d = 20$, as

$$\boldsymbol{x} \;=\; \boldsymbol{U}_k(a, b)^\top, \qquad \boldsymbol{U}_k \in \mathbb{R}^{d\times 2}, \qquad \boldsymbol{U}_k^\top \boldsymbol{U}_k = \boldsymbol{I}_2, \tag{271}$$

where $k \in \{1, \dots, 10\}$ controls the active ambient support. For the "exact-copy" family, all rows of $\boldsymbol{U}_k$ are zero except

$$U_{2r-1,1} \;=\; \frac{1}{\sqrt{k}}, \qquad U_{2r,2} \;=\; \frac{1}{\sqrt{k}}, \qquad r = 1, \dots, k. \tag{272}$$

I.e., each latent coordinate is copied uniformly over k ambient coordinates. For the "random-frame" family, we sample a permutation π and angles $\theta_1, \dots, \theta_{d/2} \sim \mathrm{Unif}(0, \pi)$. These sampled values are reused for all k. For a given k, all rows of $\boldsymbol{U}_k$ are zero except, for $r = 1, \dots, k$,

$$\boldsymbol{U}_{\pi(2r-1),:} \;=\; \frac{1}{\sqrt{k}}(\cos\theta_r,\, \sin\theta_r), \qquad \boldsymbol{U}_{\pi(2r),:} \;=\; \frac{1}{\sqrt{k}}(-\sin\theta_r,\, \cos\theta_r). \tag{273}$$

The student is a one-hidden-layer asymmetric ReLU network of width 128 with $\mathbf{F} = 0$ and $\mathbf{D} \sim \mathrm{Bernoulli}(0.15)$ i.i.d. Further, the output weights are fixed to 1. Hidden weights are initialized independently from $\mathcal{N}(0, 0.05^2)$ and hidden biases are initialized to zero. Each model is trained for 100 epochs on MSE on a batch size of 512 using AdamW (Loshchilov & Hutter, 2019) (learning rate 10^{-2}, weight decay 10^{-2}). For each k and subspace embedding, we sample 5 outer seeds for $\mathbf{D}$, and 10 initializations of student pairs for interpolation each, yielding in total 50 interpolation pairs per k and subspace embedding. For each k and subspace embedding, the reported midpoint barrier is the mean over the 50 pair-level midpoint barriers $\pm$ standard deviation over these 50 values.

G.6. Gaussian Mixture Synthetic Dataset

In our experiments (**Q3** in § 5, and § H.2, § H.3 in the appendix) we use a synthetic Gaussian mixture model dataset that allows us to control the intrinsic data dimension parameter k. We sample class centers $\boldsymbol{\mu}_c \in \mathbb{R}^k$, $(\boldsymbol{\mu}_c)_i \sim \mathcal{N}(0, \sigma_{\text{sep}}^2)$, where σ_{sep} is a class separation parameter. We generate k-dimensional class k latent samples as

$$\boldsymbol{z}_i \;=\; \boldsymbol{Q}_c(\boldsymbol{\mu}_c + \boldsymbol{\varepsilon}_i) \in \mathbb{R}^k, \tag{274}$$

where $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \mathrm{Diag}(s_1^2, \dots, s_k^2))$ is anisotropic Gaussian noise (with direction-wise scaling vector $\boldsymbol{s} \in \mathbb{R}^k$) and $\boldsymbol{Q}_c$ is a class-specific orthogonal transformation, and embed them into the ambient space as $\boldsymbol{x}_i = \boldsymbol{P}\boldsymbol{z}_i \in \mathbb{R}^d$, where $\boldsymbol{P} \in \mathbb{R}^{d\times k}$. We generate 50 000 training samples using five classes in an ambient space of dimension $d = 128$, varying the intrinsic dimension k depending on the experiment.

G.7. Empirical Neuron Realization and Swap Costs

We estimate neuron realization costs in trained models, and pairwise neuron swap costs derived from them (**Q3** in § 5, and § H.3 in the appendix) using two methods: Using the Mahalanobis distance of (5), and using a ridge ridge regression setup (see Def. E.3).

For the Mahalanobis distance estimation, we record activations during a forward pass on the whole dataset, and use them to estimate an input subspace for each layer as a PCA subspace that explains at least 90% of the variance. We then follow (5) to compute per neuron the induced subspace preactivation coefficient $\boldsymbol{a}$, the Gram matrix $\mathbf{S}_i$, and finally the realization cost (Mahalanobis distance) under the subspace support model.

For the ridge regression realization cost estimation, we optimize (250) for 15 000 iterations using AdamW (learning rate 0.005, no weight decay). We sweep $\beta \in \{0.001, 0.01, 0.1, 1.0\}$, and report results for $\beta = 0.01$. For the ridge regression

experiments, we re-train models using batch norm instead of layer norm, consider the neuron function to include the normalization and activation function, but leave the batch norm parameters fixed. The optimization for neuron i's realization cost with respect to neuron j's class starts at neuron i's trained weights.

H. Additional Experimental Results

H.1. Layerwise Activation Matching Objectives

In addition to Fig. 2, we also report activation matching objectives per layer. In Fig. 8, we plot layerwise activation matching objectives by epoch for MLPs on MNIST. This reveals that for all observed cases, the patterns are consistent across layers and, therefore, the aggregate over all layers from Fig. 2 is informative.

We also report activation matching objectives for σ-asymmetric networks (Lim et al., 2024b). This different symmetry breaking intervention operates by replacing the nonlinearity with one that does not act elementwise. This specific approach has been found to not break symmetries as effectively as $\boldsymbol{W}$-asym. in the context of unaligned LMC. While σ-asym. networks do not fit into our framework of (1), one can still perform activation matching to qualitatively analyze neuron identifiability. Fig. 8 reveals that, in fact, the identity does not produce a higher objective than random permutations, and therefore, similar to $\boldsymbol{W}$-asym. networks with $\sigma_{\mathbf{F}} = 0$, symmetries are not broken effectively.

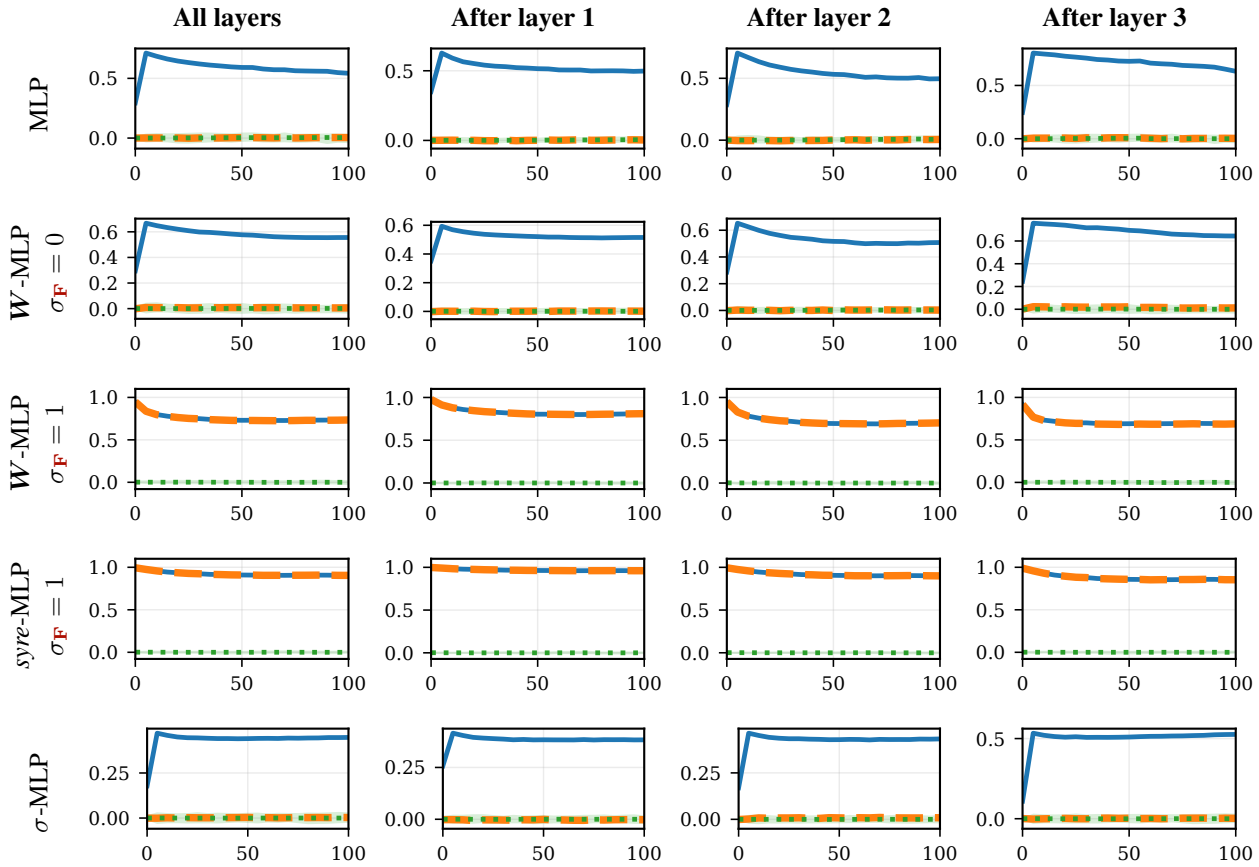
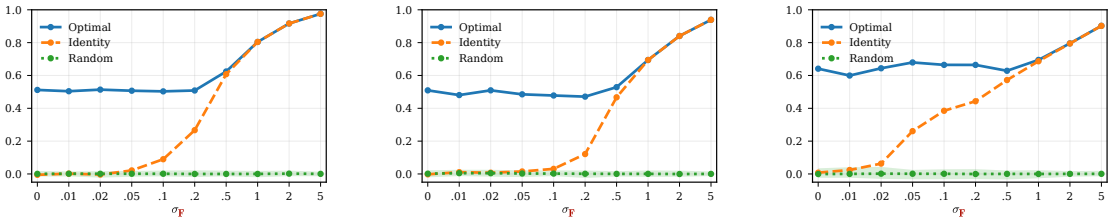


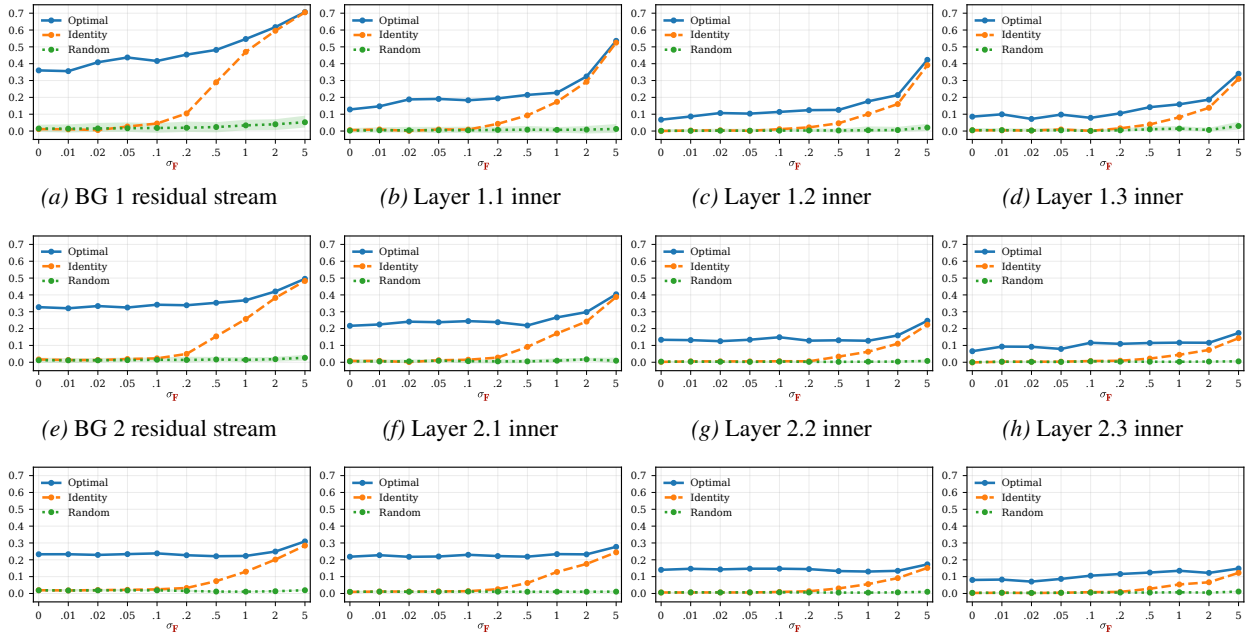
Figure 8. Layerwise activation matching objectives over epochs, separated by layer and by different symmetry-breaking architectures (including the σ-MLP variant, see Lim et al. (2024b)). We plot optimal, identity, and random permutations.

Further, in Figs. 9 and 10, we plot layerwise versions of the $\sigma_{\mathbf{F}}$ sweep for both MLPs on MNIST and ResNets on CIFAR-10 respectively. We also sweep the sparsity (proportion of fixed weights) in Fig. 11 to measure its effect on symmetry breaking, and find that sparsity alone without fixed weights does not yield strong neuron identifiability.



(a) After layer 1 *(b)* After layer 2 *(c)* After layer 3

Figure 9. Activation matching objectives per layer ($\boldsymbol{W}$-MLP on MNIST), sweeping over fixed weight scale $\sigma_{\mathbf{F}}$.

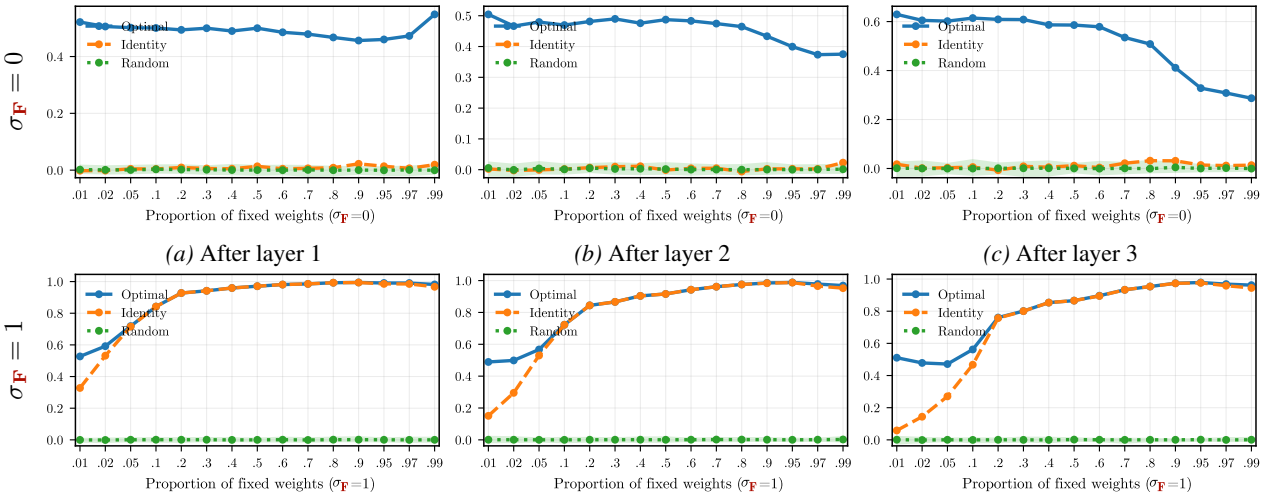


(a) BG 1 residual stream *(b)* Layer 1.1 inner *(c)* Layer 1.2 inner *(d)* Layer 1.3 inner

(e) BG 2 residual stream *(f)* Layer 2.1 inner *(g)* Layer 2.2 inner *(h)* Layer 2.3 inner

(i) BG 3 residual stream *(j)* Layer 3.1 inner *(k)* Layer 3.2 inner *(l)* Layer 3.3 inner

Figure 10. Activation matching objectives separated by layer/activation matching point ($\boldsymbol{W}$-ResNet on CIFAR-10). In our $\boldsymbol{W}$-ResNets, activation matching points lie within each of the three block groups' (BG) residual streams (used to estimate a global permutation for the residual stream that multiple layers write into), and between the two lin. layers of the three inner two-layer MLPs within each block group.



(a) After layer 1 *(b)* After layer 2 *(c)* After layer 3

(d) After layer 1 *(e)* After layer 2 *(f)* After layer 3

Figure 11. Activation matching objectives per layer ($\boldsymbol{W}$-asymmetric MLPs on MNIST), sweeping over sparsity parameter (proportion of fixed weights) for both $\sigma_{\mathbf{F}} = 0$ (standard sparse training) and $\sigma_{\mathbf{F}} = 1$.

H.2. Empirical Validation of Center Separation Rate

We also examine how the minimum distance between projected neuron centers changes with layer width m. For randomly sampled $\mathbf{F}$ we observe the expected scaling behavior from Thm. C.12. To this end, we sweep the hidden dimension m and intrinsic dimension $k \in \{2, 8, 32\}$ on the Gaussian mixture dataset. The constant c_k in Fig. 12 is chosen per k such that the rate is anchored at the respective γ_{out} for $m = 16$. Thm. C.12 predicts a rate of $\Theta(\sigma_{\mathbf{F}}\sqrt{k}m^{-2/k})$, where $\sigma_{\mathbf{F}}\sqrt{k}$ and further constants hidden by Θ are taken care of by the anchoring. This helps explain that neuron swaps are only plausible when projected centers are close, which is mainly the case for small intrinsic dimension or extremely wide layers. For realistic values of k, layers would need to be astronomically wide for such close centers to become likely.

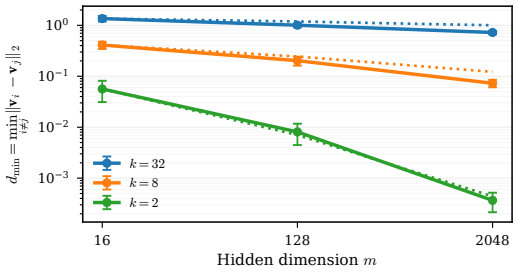


Figure 12. **Minimum pairwise projected center distance** γ_{out} (solid, mean $\pm$ std) vs. **predicted asymptotic rate** $c_k m^{-2/k}$ (dotted).

H.3. Neuron Swap Costs via Ridge Regression and Additional Pairwise Neuron Swap Costs

In § 5, we estimated the realization cost, and derived from that the neuron swap cost, via the Mahalanobis distance (5) in a PCA space with 90% explained variance. Appendix § E.1 suggests a different way to estimate realization costs, via a ridge regression objective Def. E.3. We compare the two methods in the following figures in more detail, separated by architecture, layer, and estimation method: We provide distribution box plots, comparing neuron swap costs for the four tested MLP-based architectures (MNIST) / zero fixed weights vs. center-dominated regime $\boldsymbol{W}$-asym. MLPs (Gaussian mixture data), and detailed pairwise neuron swap cost plots showing large/small/negative costs.

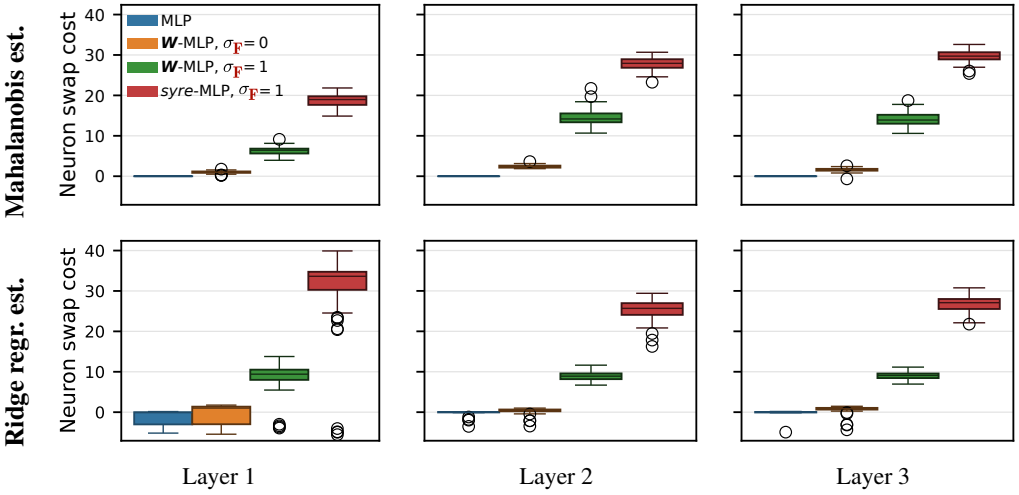


Figure 13. Distributions of neuron swap costs by architecture on **MNIST**. Plotted are signed sqrt. transformed $\Delta^{\text{out}}_{(ij)}$ for disjoint consecutive pairs (i, j) of neurons, estimated via Mahalanobis distance (5) and ridge regression (Def. E.3, $\beta = 0.01$).

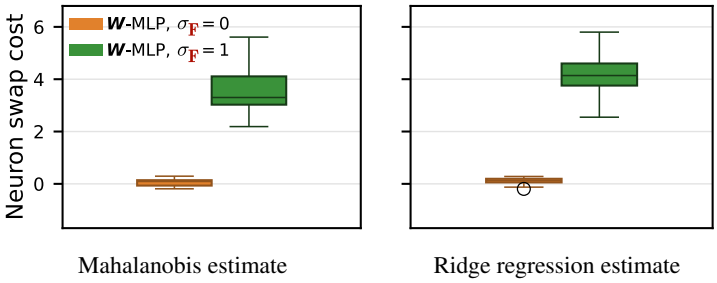


Figure 14. Neuron swap costs for $\boldsymbol{W}$-MLPs with $\sigma_{\mathbf{F}} \in \{0, 1\}$ on **Gaussian mixture data**. Plotted are signed sqrt.transformed $\Delta^{\text{out}}_{(ij)}$ for disj. consecutive pairs (i, j) of neurons, estimated via Mahalanobis dist. (5) and ridge regression (Def. E.3, $\beta = 0.01$).

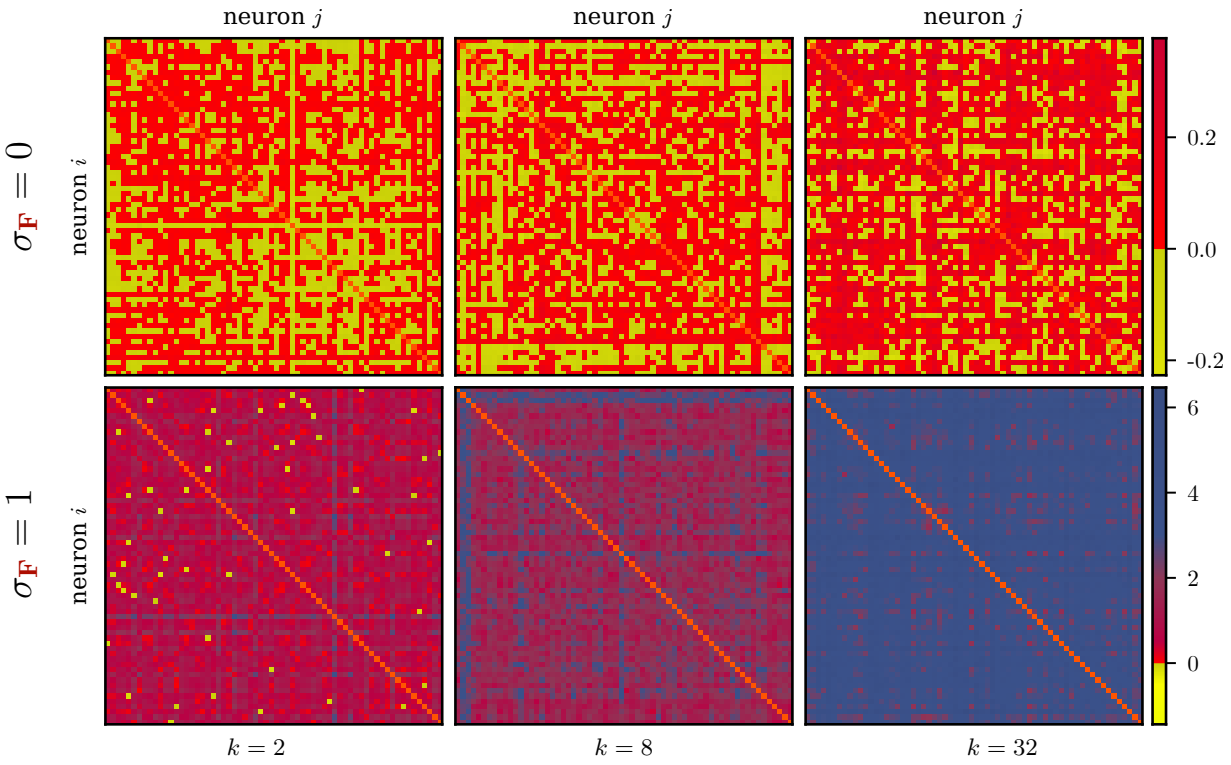


Figure 15. Neuron swap costs $\Delta^{\text{out}}_{(ij)}$ (signed square-root transformed) estimated via **Mahalanobis distance** on **Gaussian mixture data**, for varying *intrinsic dimension* k and *fixed weight scale* $\sigma_{\mathbf{F}}$ (same as Fig. 5).

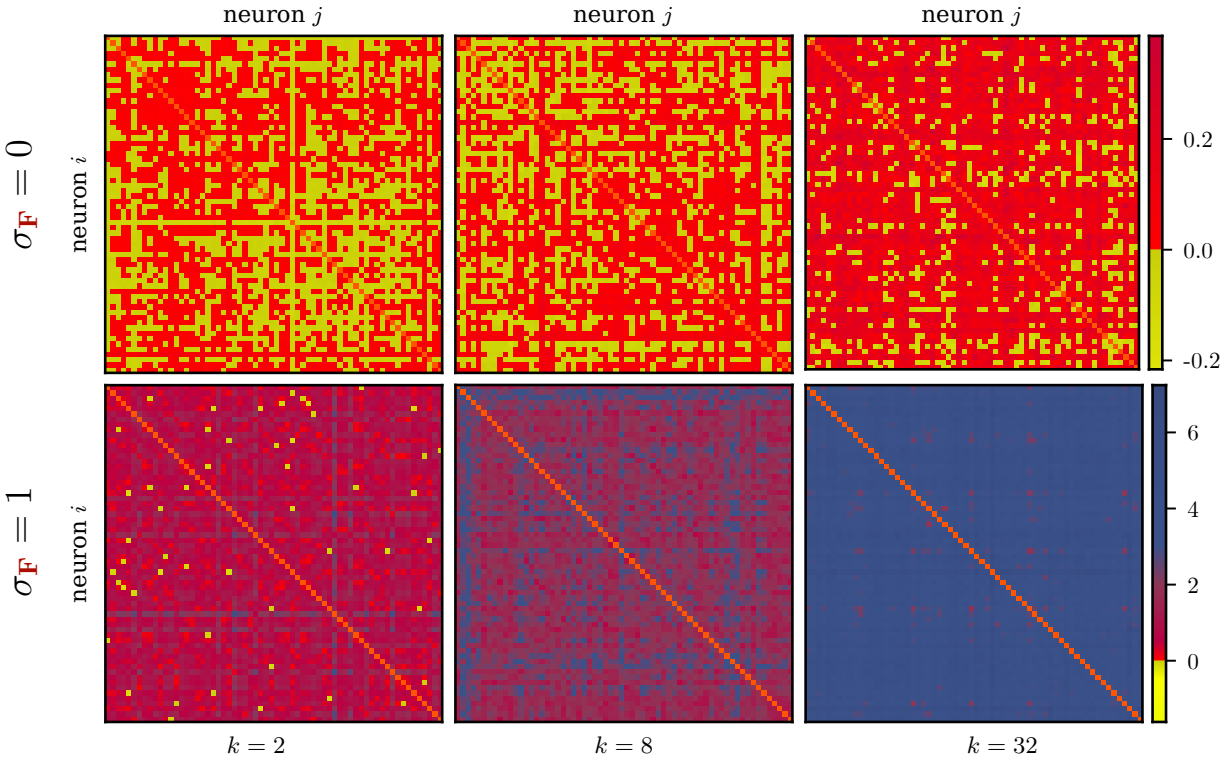


Figure 16. Neuron swap costs $\Delta^{\text{out}}_{(ij)}$ (signed square-root transformed) estimated via **ridge regression**, $\beta = 0.01$ (see Def. E.3) on **Gaussian mixture data**, for varying *intrinsic dimension* k and *fixed weight scale* $\sigma_{\mathbf{F}}$.

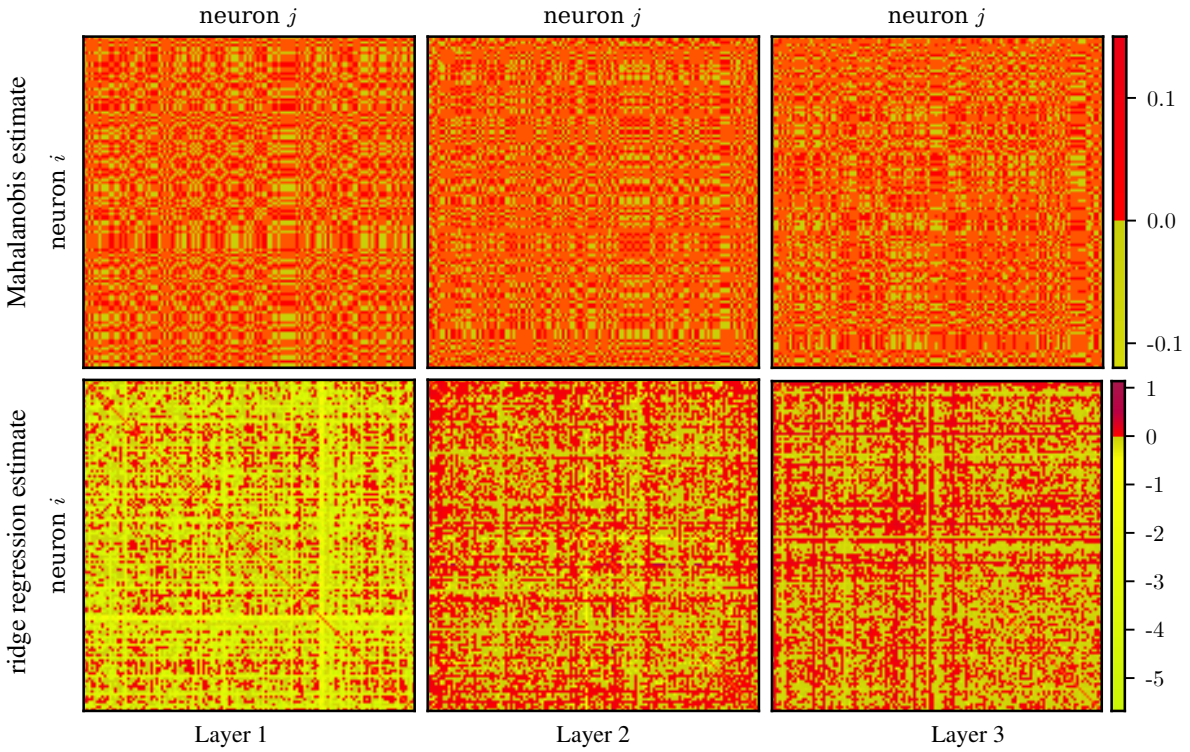


Figure 17. Neuron swap costs $\Delta^{\text{out}}_{(ij)}$ (signed square-root transformed) *per layer* for **standard MLP** trained on **MNIST**, estimated via *Mahalanobis distance* (5) vs. *ridge regression* (Def. E.3, $\beta = 0.01$). Layers 1 – 3 have hidden dimension 512, shown are costs for first 128×128 neuron pairs.

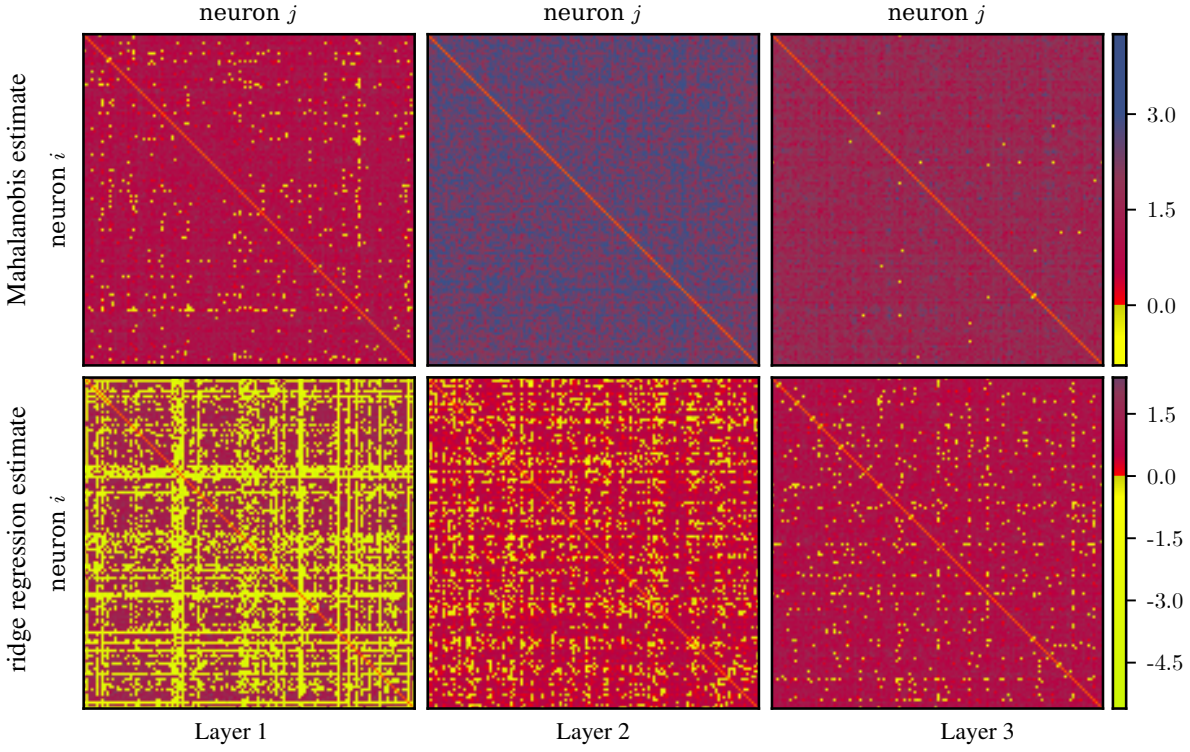


Figure 18. Neuron swap costs $\Delta^{\text{out}}_{(ij)}$ (signed square-root transformed) *per layer* for ***W*-asymmetric MLP ($\sigma_{\mathbf{F}} = 0$)** trained on **MNIST**, estimated via *Mahalanobis distance* (5) and *ridge regression* (Def. E.3, $\beta = 0.01$). Layers 1 – 3 have hidden dimension 512, shown are costs for first 128×128 neuron pairs.

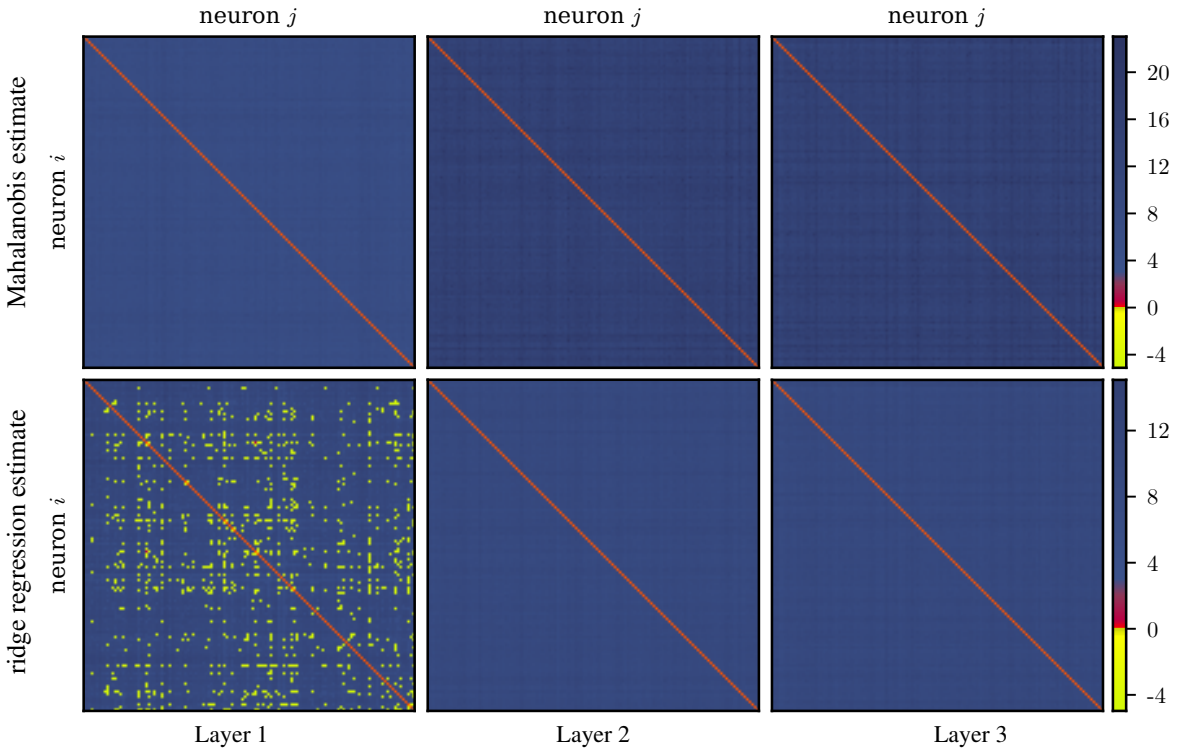


Figure 19. Neuron swap costs $\Delta^{\text{out}}_{(ij)}$ (signed square-root transformed) *per layer* for ***W*-asymmetric MLP ($\sigma_{\mathbf{F}} = 1$)** trained on **MNIST**, estimated via *Mahalanobis distance* (5) and *ridge regression* (Def. E.3, $\beta = 0.01$). Layers 1 – 3 have hidden dimension 512, shown are costs for first 128×128 neuron pairs.

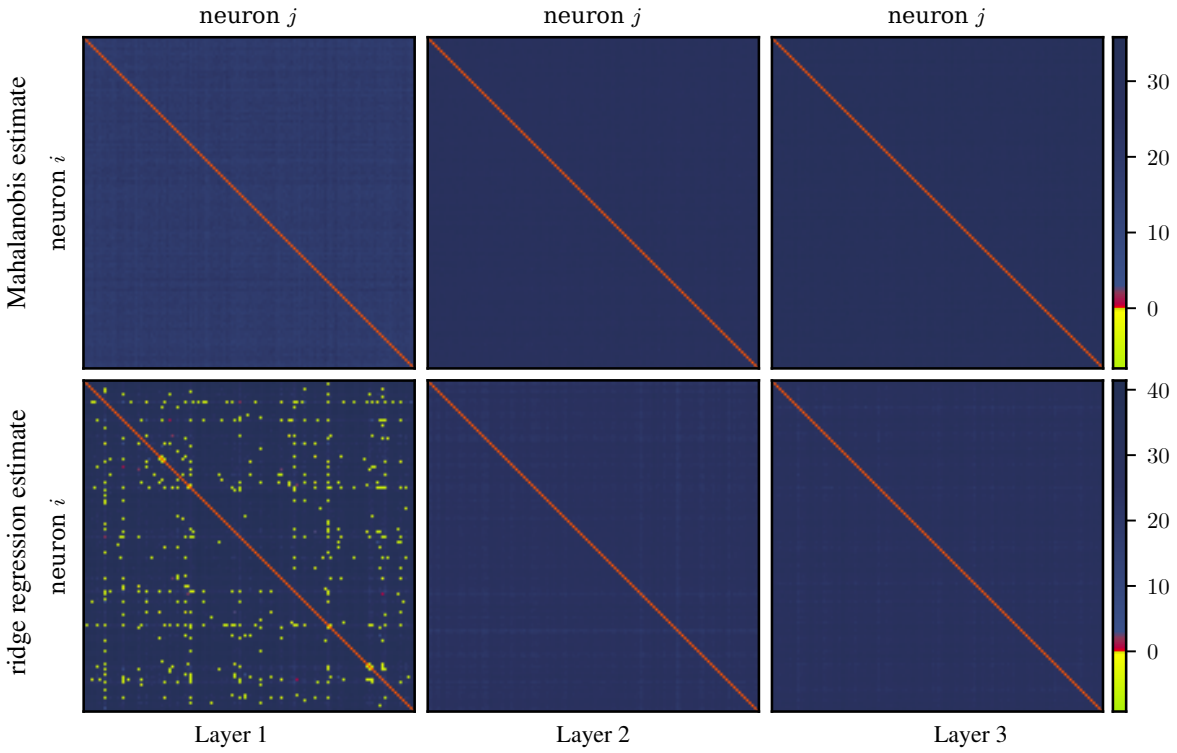


Figure 20. Neuron swap costs $\Delta^{\text{out}}_{(ij)}$ (signed square-root transformed) *per layer* for ***syre*-MLP ($\sigma_{\mathbf{F}} = 1$)** trained on **MNIST**, estimated via *Mahalanobis distance* (5) and *ridge regression* (Def. E.3, $\beta = 0.01$). Layers 1 – 3 have hidden dimension 512, shown are costs for first 128×128 neuron pairs.

H.4. Transformers

§ 5 studies the effect of neural parameter symmetries on optimization and linear mode connectivity for MLPs and ResNets. We additionally examine Vision Transformers (ViTs) (Dosovitskiy et al., 2021) on CIFAR-10 to test whether the same symmetry breaking mechanism extends to attention-based architectures. A transformer is composed of linear maps, including the query, key, value, and output projections in multi-head self-attention, the two linear layers in each feed-forward block, the patch embedding layer, and the final classification head. Transformers are an especially interesting extension because attention admits richer $\mathrm{GL}(\mathbb{R}^n)$ symmetries beyond the permutation setting in § 5. We note that our goal is not to introduce a highly tuned state-of-the-art ViT, but to use a trainable transformer baseline whose training accuracy reached high performance. The goal is to investigate whether the same parameter-level intervention can reduce the interpolation barrier.

The ViT setting is more challenging for several reasons. First, transformer symmetries are not limited to hidden-unit or channel permutations. Multi-head attention introduces head-permutation symmetries, while the feed-forward blocks contain MLP-like hidden-unit symmetries. Second, LayerNorm, residual connections, positional embeddings, and the class token may interact with these symmetries in ways not captured by a simple linear-map masking scheme. Third, our implementation applies the $\boldsymbol{W}$-asymmetric parameterization directly to the transformer's linear maps. This is a natural extension of the original construction, but it may not be the optimal transformer-specific way to break symmetries. For example, breaking individual entries of $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, $\boldsymbol{W}_V$, $\boldsymbol{W}_O$ may not fully eliminate head-level equivalences unless the asymmetry is coordinated with the internal head structure. Similarly, asymmetrizing all linear maps may introduce additional optimization difficulty because fixed entries constrain the parameterization throughout training.

We use a CIFAR-10 ViT with patch size 4, embedding dimension 384, depth 8, 6 attention heads, and MLP ratio 4. Following the experimental setup in § 5, we apply a random 80%/10%/10% train/val/test split. For each linear map, we replace the ordinary trainable weight matrix $\boldsymbol{W}$ with a $\boldsymbol{W}$-asymmetric parameterization in the form of (1). The fixed entries are sampled once at initialization, and the number of fixed entries is 8. This construction is a natural extension of the $\boldsymbol{W}$-asymmetric parameterization used for MLPs and ResNets, but it is not necessarily the optimal way to break transformer symmetries. For example, a more tailored approach might use head-aware masks or a design that explicitly respects the coupled structure of $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, $\boldsymbol{W}_V$, $\boldsymbol{W}_O$. We leave such transformer-specific variants for future work. We train the model using AdamW (Loshchilov & Hutter, 2019) with cosine learning-rate decay for 200 epochs using a batch size of 128. In our preliminary tuning, we found that larger fixed-weight scales can make transformer optimization more difficult. In particular, settings such as $\sigma_{\mathbf{F}} = 2$ were harder to train. When their scale is too large, they can dominate the trainable entries, alter the scale of attention logits and MLP activations, and reduce the flexibility of the model during optimization. Therefore, unlike in an idealized symmetry breaking setting, increasing $\sigma_{\mathbf{F}}$ should not be interpreted as a monotonic improvement in practice.

Fig. 21 reports the training accuracy along the linear interpolation path between independently trained transformer endpoints. We see that the standard transformer exhibits a pronounced interpolation barrier and the $\boldsymbol{W}$-Transformer with $\sigma_{\mathbf{F}} = 0$ closely matches this behavior, suggesting that masking alone, when the fixed entries are zero, is not sufficient to meaningfully improve unaligned LMC. In contrast, the $\boldsymbol{W}$-Transformer with $\sigma_{\mathbf{F}} = 1$ yields a noticeably flatter interpolation curve and substantially higher midpoint accuracy. However, the interpolation curve still remains well below the endpoint accuracy. The improvement at $\sigma_{\mathbf{F}} = 1$ suggests that nonzero $\boldsymbol{W}$-asymmetric fixed entries can reduce the unaligned interpolation barrier in transformers, qualitatively matching the mechanism observed for MLPs and ResNets in § 5. At the same time, the remaining barrier indicates that the current entrywise $\boldsymbol{W}$-asymmetric construction does not fully resolve the symmetries or optimization effects present in transformer architectures. This may be due to residual attention-head symmetries, richer $\mathrm{GL}(\mathbb{R}^n)$ invariances, interactions with LayerNorm and residual connections, or the fact that our masks are inherited from the MLP and ResNet setting rather than designed specifically for attention. Thus, this provides evidence that the mechanism identified by our theory carries over qualitatively to transformers, while also motivating future work on transformer-specific symmetry breaking schemes.

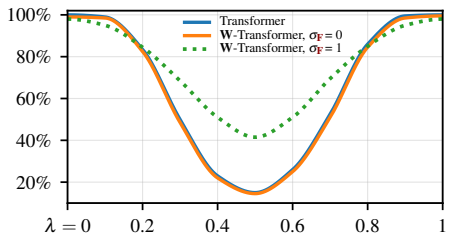
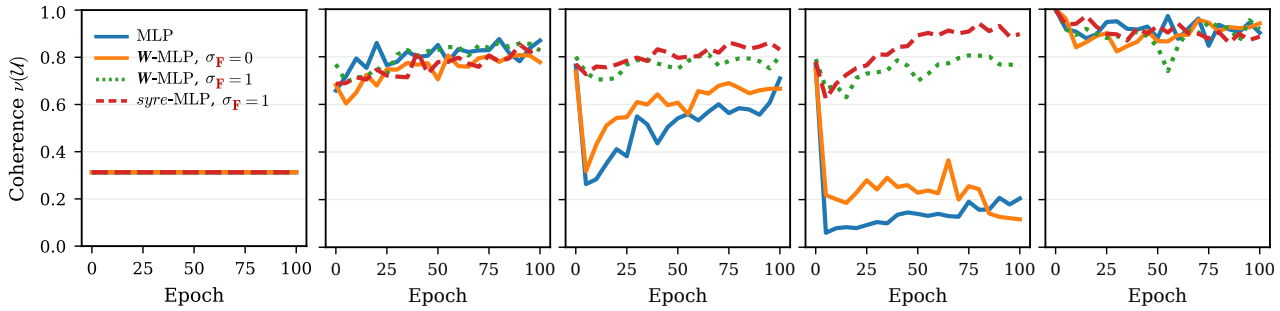


Figure 21. LMC of Transformer and $\boldsymbol{W}$-Transformer on CIFAR-10, measured by training accuracy along the interpolation path.

H.5. Subspace Coherence of MNIST Inputs and Hidden Representations

We empirically measure the subspace coherence $\nu(\mathcal{U})$ of model inputs, hidden representations and the final output for standard and asymmetric MLPs on MNIST data variants over training. Fig. 22 shows the results: in particular, variants



(a) Model inputs *(b)* Layer 2 inputs *(c)* Layer 3 inputs *(d)* Layer 4 inputs *(e)* Model outputs

Figure 22. Subspace coherences $\nu(\mathcal{U})$ computed for model inputs, outputs, and hidden representations, on MNIST data, for standard and asymmetric MLP variants over training.

without effective symmetry breaking (MLP and $\boldsymbol{W}$-asymmetric MLP with zero fixed weights) exhibit low subspace coherence, in particular in later layers, while $\boldsymbol{W}$-asymmetric MLPs and *syre*-MLPs with large fixed weights both tend to exhibit higher subspace coherence values.