A Mechanism of Weak-to-Strong Generalization: Feature Elicitation from Latent Knowledge

Ryoya Awano[1*] **Taiji Suzuki**[1,2]
[1]The University of Tokyo [2]Center for Advanced Intelligence Project, RIKEN
[*]awano-ryoya284@g.ecc.u-tokyo.ac.jp

Abstract

Weak-to-strong (W2S) generalization, in which a strong model is fine-tuned on outputs of a weaker, task-specialized model, has been proposed as an approach to aligning superhuman AI systems. Existing theoretical analyses either fix the student's representations or operate in restricted settings. Whether multi-step SGD can succeed in feature learning while preserving diverse pre-trained capabilities remains open. We study W2S in the setting of reward-model learning with two-layer neural networks. The strong model has pre-trained representations organized into low-dimensional subspaces V_k, and is fine-tuned under the supervision of a weak model specialized on task κ. We prove that the strong model efficiently learns task κ, eliciting its pre-trained knowledge while preserving pre-trained features for non-target tasks. This provides a provable account of W2S generalization in a feature-learning regime, in the sense that the strong model acquires the target feature direction through W2S training, rather than having it given a priori. Moreover, W2S preserves pre-trained off-target features, whereas the supervised fine-tuning procedure we analyze can cause catastrophic forgetting when off-target feature directions are correlated with the target's. Numerical experiments on synthetic data confirm our theoretical results.

1 Introduction

Burns et al. (2023) showed that fine-tuning a strong language model on the outputs of a weaker, task-specialized model can produce a student that surpasses its teacher on that task (e.g., GPT-4 (Achiam et al., 2023) fine-tuned on GPT-2 (Radford et al., 2019) labels outperforms GPT-2 across NLP benchmarks). This phenomenon, *weak-to-strong* (W2S) generalization, has become a workhorse for studying how imperfect supervision by a less capable model can nonetheless elicit useful behavior from a more capable one. It underlies the superalignment program, in which humans act as the weak supervisors of models beyond their own capabilities.

Empirically, Burns et al. (2023) demonstrate that W2S can successfully elicit capabilities even under linear probing, where the strong model's representations are fixed. This may partly explain why many early theoretical treatments of W2S have focused on the linear regime. However, they also show that full fine-tuning—allowing for feature learning—unlocks better performance. To explain this gain, they hypothesize that feature learning makes the concepts acquired during pre-training more "salient." W2S would thus elicit latent knowledge rather than instill entirely new capabilities.

The theoretical study of W2S has addressed either abstract frameworks that do not analyze gradient-based optimization (Lang et al., 2024; Shin et al., 2025; Charikar et al., 2024; Xue et al., 2025) or linear and random feature models (Wu and Sahai, 2025; Ildiz et al., 2025; Dong et al., 2025; Medvedev et al., 2025). Neither line derives the difference between weak and strong model features from the optimization process itself. Motivated by the empirical advantage of full fine-tuning, recent theoretical works (Moniri and Hassani, 2025; Oh et al., 2025) took the first steps toward W2S feature

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

learning, but their theoretical settings leave significant gaps regarding architecture, pre-training priors, and task specialization. Oh et al. (2025) analyze W2S by employing a linear CNN as the weak model and a two-layer ReLU CNN as the strong model, with convergence proved under full-batch gradient descent. However, this CNN-based architecture differs from the LLMs used in Burns et al. (2023), assumes no pre-training prior on the strong model, and does not specialize the weak model to a particular task. Moniri and Hassani (2025) use a two-component nonlinear model rather than a neural network (NN), assume the strong model already knows the exact direction of the off-target task at initialization, and their analysis tracks only a single gradient step.

A key open question is whether multi-step SGD can provably succeed at W2S feature learning while preserving diverse capabilities, in a more realistic setting. Answering this question would shed light on the mechanism by which W2S makes pre-trained latent knowledge salient. We study this question in the setting of reward-model learning, where, as in Burns et al. (2023), the strong model is trained on the weak model to estimate a reward directly.

Pre-trained models have been observed to localize individual tasks in distinct subsets of neurons (Dai et al., 2022; Panigrahi et al., 2023). We therefore adopt an additive reward $r^*(x) = \sum_{k=1}^{K} r_k^*(\theta_k^\top x)$, where each component r_k^* corresponds to one task. Furthermore, consistent with empirical findings that pre-trained models encode concepts in low-dimensional subspaces (Beaglehole et al., 2026), we assume the strong model has pre-training knowledge of the task subspaces V_k.

We focus on the setting where the weak model is a small two-layer NN specialized to task κ, trained to approximate r_κ^*. The strong model, a larger two-layer NN, queries the weak model based on its prior knowledge. It then trains by multi-step SGD on the weak model's outputs with nonlinear transformation.

Contributions

- **Feature elicitation via SGD analysis of W2S.** We study W2S in a setting closer to pre-trained LLMs, with an additive multi-feature reward and task subspaces as pre-training knowledge. To our knowledge, we give the first multi-step SGD convergence guarantee for W2S in the feature-learning regime, with sample complexity $\tilde{O}(s^{3/2})$ (Theorem 1). This sample complexity depends only on the subspace dimension s, not the input dimension d, and gives a provable account of how W2S elicits pre-trained latent features.
- **Feature preservation under W2S.** The strong model's pre-trained input distribution, concentrated on V_κ, and the nonlinear teacher transformation together act as an implicit regularizer within the task subspaces. We prove that the feature directions of off-target tasks are preserved, provided those tasks are allocated more neurons than the target task (Proposition 2). This preservation, combined with feature elicitation on task κ, provides a theoretical account of W2S generalization.
- **Comparison with SFT.** While Proposition 2 shows that W2S preserves off-target features, we also analyze the behavior of standard SFT on the same strong model, replacing the W2S algorithm with isotropic sampling and the direct teacher r_κ^*. Theorem 2 shows that off-target tasks with feature directions similar to θ_κ suffer catastrophic forgetting. Together with Theorem 1, this establishes advantages of W2S in sample complexity and feature preservation over SFT using online SGD with isotropic inputs and direct target labels.

Related work *Weak-to-strong generalization.* Theoretical analyses of W2S fall into two groups. One line characterizes when W2S succeeds through abstract frameworks without analyzing gradient-based optimization (Lang et al., 2024; Shin et al., 2025; Charikar et al., 2024; Xue et al., 2025). Another restricts to linear or random feature models (Wu and Sahai, 2025; Ildiz et al., 2025; Dong et al., 2025; Medvedev et al., 2025). Moving beyond these settings, Moniri and Hassani (2025); Oh et al. (2025) initiate the non-linear feature-learning analysis of W2S. Moniri and Hassani (2025) assume the strong model starts with pre-trained knowledge of the target direction and analyze one step of SGD, while Oh et al. (2025) use a CNN-based architecture without a pre-training prior. We analyze multi-step SGD in a setting closer to pre-trained LLMs, with an additive multi-feature reward and a more natural pre-training prior. Additional related work is discussed in Section B. We discuss the linear representation hypothesis motivating our subspace prior, and prior work on feature learning for single-index and additive models.

2 Problem setting

We formalize the W2S generalization framework for reward learning, modeling each prompt–response pair (u, v) as a vector $x \in \mathbb{R}^d$. The goal is to learn the true reward $r^*(x)$ from weak supervision.

Notations. We write $\|\cdot\|$ for the ℓ_2 norm (vectors) and operator norm (matrices). For $f\colon \mathbb{R} \to \mathbb{R}$, we denote f^i for the i-th power and $f^{(i)}$ for the i-th derivative. Let He_i be the i-th probabilist's Hermite polynomial. We write S^{d-1} for the unit sphere in $\mathbb{R}^d$, and $\tilde{\nabla}_w := (I - ww^\top)\nabla_w$ for the spherical gradient at $w \in S^{d-1}$. We use $\tilde{O}(\cdot), \tilde{\Omega}(\cdot), \tilde{\Theta}(\cdot)$ to suppress $\mathrm{polylog}(d)$ factors. An event holds *with high probability* if $\mathbb{P}(\cdot) \geq 1 - d^{-C}$ for an arbitrarily large constant $C > 0$. We write $\Delta^{K-1} = \{(\lambda_k)_{k=1}^K : \lambda_k \geq 0,\ \sum_k \lambda_k = 1\}$ for the probability simplex.

2.1 Sample complexity and the information exponent

We define the information (Dudeja and Hsu, 2018; Ben Arous et al., 2021) and generative (Damian et al., 2024b) exponents, which govern the sample complexity of learning a single-index model.

Definition 1 (Information and generative exponents)**.** Let $f\colon \mathbb{R} \to \mathbb{R}$ be square-integrable under $\mathcal{N}(0,1)$ with Hermite expansion $f(z) = \sum_{j\geq 0} \frac{\alpha_j}{\sqrt{j!}} \mathrm{He}_j(z)$. The *information exponent* of f is $\mathrm{IE}(f) = \min\{j > 0 : \alpha_j \neq 0\}$. The *generative exponent* of f is $\mathrm{GE}(f) = \inf_{\mathcal{T} \in L^2} \mathrm{IE}(\mathcal{T} \circ f)$.

For a polynomial f of degree q, we have $\mathrm{GE}(f) \leq \mathrm{IE}(f) \leq q$; moreover, $\mathrm{GE}(f) = 2$ if f is even, and $\mathrm{GE}(f) = 1$ otherwise (Lee et al., 2024). Intuitively, a larger $p = \mathrm{IE}(f)$ means the signal about the direction θ appears only in higher-order Hermite components of f, making it harder to detect from samples $x \sim \mathcal{N}(0, I_d)$. When $p > 2$, online SGD requires $\tilde{O}(d^{p-1})$ samples to find the hidden direction θ from observations $(x, f(\theta^\top x) + \zeta)$ of the single-index model, with $x \sim \mathcal{N}(0, I_d)$ (Ben Arous et al., 2021). However, by applying a nonlinear transformation, Chen and Meka (2020) achieve $\tilde{O}(d)$ sample complexity for polynomial link functions.

2.2 Additive reward model

As motivated in Section 1, we model the true reward as an additive combination of K single-index components. Following Oko et al. (2024a), we study the case where each link function is a polynomial. Each feature direction θ_k is in a pre-trained subspace $V_k \subseteq \mathbb{R}^d$, motivated by empirical evidence that LLMs encode concepts in low-dimensional subspaces (Beaglehole et al., 2026). A similar assumption appears in Oko et al. (2024b) and Nishikawa et al. (2025), where feature vectors are drawn from a low-dimensional subspace identified by pre-training.

Assumption 1 (Additive reward model with task subspaces)**.** The reward r^* decomposes additively across K tasks indexed by $k \in [K]$, with target task κ.

(i) **Additive reward.** The true reward is

$$r^*(x) = \sum_{k=1}^K r_k^*(\theta_k^\top x), \qquad r_k^* = \pi_k \sigma_k^*,$$

observed as $y = r^*(x) + \zeta$ with $x \sim \mathcal{N}(0, I_d)$, $\zeta \sim \mathcal{N}(0,1)$, and $\|\theta_k\| = 1$. Each link function $\sigma_k^* \in L^2(\mathcal{N}(0,1))$ is normalized so that $\mathbb{E}_{t\sim\mathcal{N}(0,1)}[\sigma_k^*(t)] = 0$ and $\mathbb{E}_{t\sim\mathcal{N}(0,1)}[(\sigma_k^*(t))^2] = 1$. The target link function σ_κ^* is a degree-q polynomial with $\mathrm{IE}(\sigma_\kappa^*) = p > 2$, as in Oko et al. (2024a); Simsek et al. (2025); Ren et al. (2025). The weights $\pi_1 \geq \cdots \geq \pi_K \geq 0$ satisfy $\sum_{k=1}^K \pi_k^2 = 1$.

(ii) **Task subspaces.** Each task k is associated with a subspace $V_k \subseteq \mathbb{R}^d$ of dimension $s = d^\alpha$ for some $\alpha \in (0,1)$, containing its feature direction $\theta_k \in V_k$. The subspaces V_k need not be mutually orthogonal; Proposition 2 exploits nonzero inner products $\theta_k^\top \theta_\kappa$ across tasks. We write Σ_k for the orthogonal projection onto V_k.

Remark. The Gaussian assumption $x \sim \mathcal{N}(0, I_d)$ is standard in the theoretical study of feature learning (Ben Arous et al., 2021; Damian et al., 2024b). Writing $\sigma_\kappa^*(t) = \sum_{i=p}^q \frac{\alpha_i}{\sqrt{i!}} \mathrm{He}_i(t)$, the normalization condition reads $\sum_{i=p}^q \alpha_i^2 = 1$. The subspaces V_k encode the pre-training prior that the strong model has identified the subspace V_k containing θ_k, but not θ_k itself, extending the setting of Moniri and Hassani (2025) where off-target feature vector is known exactly.

2.3 Weak model

We consider that the weak model is small and specialized to task κ. We model it as a two-layer network whose neurons are nearly aligned with θ_κ and whose output approximates r_κ^* in L^1.

Assumption 2 (Weak model)**.** The weak model, fine-tuned on task $\kappa \in [K]$, is a two-layer NN

$$r^w(x) = \frac{1}{N^w} \sum_{n=1}^{N^w} a_n^w \sigma_n^w(w_n^\top x + b_n^w),$$

where each activation σ_n^w is a degree-q polynomial with Hermite coefficients uniformly bounded in n, $\|w_n\| = 1$, $\|a^w\|_1 = \tilde{O}(\pi_\kappa N^w)$, and $|b_n^w| \le C_b = \tilde{O}(1)$. The weak model has learned feature κ in the following sense:

(i) **Feature alignment.** There exists $\tilde{\varepsilon}_w > 0$ such that $\theta_\kappa^\top w_n \ge 1 - \tilde{\varepsilon}_w$ for all $n \in [N^w]$.

(ii) L^1 **approximation.** There exists $\varepsilon_w > 0$ such that $\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\left|r^w(x) - r_\kappa^*(\theta_\kappa^\top x)\right|\right] \le \pi_\kappa \varepsilon_w$.

The next proposition shows that a weak model with pre-training knowledge of V_κ can satisfy Assumption 2 via fine-tuning on task κ, with sample complexity improved by the subspace prior (for the formal statement, see Proposition 1).

Proposition 1 (Weak model construction; informal)**.** Under Assumptions 1 and 4, online SGD (Algorithm 2) applied to the single-index model $r_\kappa^* = \pi_\kappa \sigma_\kappa^*$ produces a weak model satisfying Assumption 2 with $\varepsilon_w = \tilde{\Theta}(\tilde{\varepsilon}_w)$, using

$$T = \tilde{O}\Big(d^{p/2}\, s^{(p-2)/2} \vee d\, \tilde{\varepsilon}_w^{-2} \vee \tilde{\varepsilon}_w^{-3}\Big)$$

samples. This improves over the baseline $\tilde{O}(d^{p-1})$ required without subspace knowledge.

Remark. The analysis builds on the literature (Ben Arous et al., 2021; Damian et al., 2024b; Lee et al., 2024; Oko et al., 2024a). However, these results guarantee only (ii) in Assumption 2; satisfying (i) additionally requires PCA-like filtering of neurons without observing θ_κ (see Section D.6). If any neuron violates condition (i), the weak model's L^1 error under the subspace distribution is no longer controlled by its error under the d-dimensional isotropic distribution (see Lemma 21), and the bound in Theorem 1 fails. In addition, to incorporate the pre-trained knowledge of V_κ, we initialize w_n uniformly on $S^{d-1} \cap V_\kappa$ rather than isotropically on S^{d-1}. This achieves complexity $\tilde{O}(d^{p/2} s^{(p-2)/2})$ rather than the $\tilde{O}(d^{p-1})$ of Ben Arous et al. (2021). The number of samples required to drive the alignment $\theta_\kappa^\top w_n^t$ to $\Theta(1)$ scales as $(\theta_\kappa^\top w_n^0)^{-(p-2)}$, so the subspace prior reduces the factor $d^{(p-2)/2}$ to $s^{(p-2)/2}$. The remaining factor $d^{p/2}$, which arises from the input noise rather than initialization, remains unchanged.

2.4 Strong model

Task localization (Dai et al., 2022; Panigrahi et al., 2023) in pre-trained models motivates partitioning the strong model's neurons by task. We model the strong model as a pre-trained two-layer network with K groups of neurons, each group corresponding to one component r_k^* of the additive reward:

$$r(x) = \sum_{k=1}^{K} r_k(x), \quad r_k(x) = \frac{1}{N_k} \sum_{n=1}^{N_k} a_{k,n}\, \sigma_{k,n}(w_{k,n}^\top x + b_{k,n}),$$

where $\|w_{k,n}\| = 1$. Unlike Oko et al. (2024a), where task localization emerges during training from isotropic initialization, we assume that the strong model is pre-trained and already localized into task-specific groups. For task κ, the target neurons are initialized with $\Sigma_\kappa w_{\kappa,n}^0$ uniform on $S^{d-1} \cap V_\kappa$ and perpendicular component satisfying $\|\Sigma_\kappa^\perp w_{\kappa,n}^0\|^2 = o(s^{-1/2})$ (see Assumption 5). For $k \ne \kappa$, the projection $\|\Sigma_\kappa w_{k,n}^0\|^2$ of a task-k neuron onto V_κ determines how much the task-κ gradient update affects task-k neurons. Small values of this quantity preserve pre-trained features for $k \ne \kappa$, as formalized in Proposition 2.

The strong model generates responses from its pre-trained distribution, in line with recent W2S literature (Ji et al., 2024; Tao and Li, 2025) rather than the original setup of Burns et al. (2023). We model the resulting embedding distribution as μ concentrating on V_κ, as an idealization of the pre-training prior that the strong model has internalized task-κ knowledge.

Algorithm 1: W2S feature learning via online SGD

Input: Weak model r^w; initialized strong model $r_{(\Theta_k)_{k=1}^K}$ with $\Theta_k = (a_{k,n}, b_{k,n}, w_{k,n}^0)_{n=1}^{N_k}$ (Assumption 5); learning rate schedule $(\eta^t)_{t\geq 0}$; temperature $\rho > 0$; iterations T.

for $t = 0, 1, \ldots, T-1$ **do**

  Sample $x^t \sim \mu$ (Assumption 3).

  $\bar{r}^w(x^t) \leftarrow \mathrm{clip}\big(\rho^{-1} r^w(x^t),\ \pm 1/\log d\big)$.

  $y^t \leftarrow \bar{r}^w(x^t)\ \exp\big(\bar{r}^w(x^t)\big)$. `// nonlinear transformation`

  for $k \in [K]$, $n \in [N_k]$ **do**

    $w_{k,n}^{t+1} \leftarrow w_{k,n}^t + \eta^t\, y^t\, \tilde{\nabla}_{w_{k,n}} r_\Theta(x^t)$. `// spherical gradient`

    $w_{k,n}^{t+1} \leftarrow w_{k,n}^{t+1} / \big\| w_{k,n}^{t+1} \big\|$.

Output: $\hat{w}_{k,n} \leftarrow w_{k,n}^T$ for all $k \in [K]$, $n \in [N_k]$.

Assumption 3 (Generative distribution)**.** When the strong model is queried on task κ, we model its input distribution μ as the following mixture.

$$\mu = \sum_{k=1}^{K} \lambda_k\, \mu_k, \quad (\lambda_k)_{k=1}^K \in \Delta^{K-1}, \quad 1 - \lambda_\kappa = o(s^{-1/2}),$$

where $\mu_\kappa = \mathcal{N}(0, \Sigma_\kappa)$ is the Gaussian supported on the target subspace V_κ, and for $k \neq \kappa$ each μ_k is a sub-Gaussian distribution supported on V_k with identity covariance on V_k.

With $1 - \lambda_\kappa = o(s^{-1/2})$, nearly all sampled embeddings lie in V_κ, reducing the effective dimension from d to s in the sample complexity of Theorem 1 and suppressing gradient updates to off-target neurons in Proposition 2.

3 W2S learning

Algorithm 1 trains the first-layer weights of the strong model via online SGD similar to Ben Arous et al. (2021); Damian et al. (2024a); Lee et al. (2024); Oko et al. (2024a), using a nonlinear transformation of the weak model output as supervision and inputs drawn from the generative distribution μ (Assumption 3). Each update takes a gradient step on the correlation loss $-y^t r_\Theta(x^t)$ with respect to $w_{k,n}$. The transformation on Algorithm 1 scales the weak model output by ρ^{-1} before clipping it to $\pm 1/\log d$ and applying $z \mapsto z\exp(z)$ analogous to Nishikawa et al. (2025), placing the gradient update outside the correlational statistical query (CSQ) framework (Bshouty and Feldman, 2002) and exploiting $\mathrm{GE}(\sigma_\kappa^*)$. When σ_κ^* is even, this reduces the effective exponent from p to 2. Together, this exponent reduction and the s-dimensional support of μ_κ (Assumption 3) yield the $\tilde{O}(s^{3/2})$ complexity in Theorem 1, which is lower than the $\tilde{O}(d^{p/2}\, s^{(p-2)/2})$ cost of the weak model (Proposition 1).

3.1 Main results

Theorem 1 (W2S feature alignment; informal)**.** Assume Assumptions 1 to 3 and 5. Let $\mathcal{E}_s = \varepsilon_w \vee \tilde{\varepsilon}_w \vee (1 - \lambda_\kappa) \lesssim s^{-1/2}$, where ε_w and $\tilde{\varepsilon}_w$ are the L^1 approximation error and the alignment error of the weak model (Assumption 2). Fix a target accuracy $\tilde{\varepsilon} = \tilde{\Omega}(\mathcal{E}_s)$. Then, for a suitable choice of the temperature ρ, Algorithm 1 with

$$T = \tilde{O}\Big(s^{3/2} \vee s\,\tilde{\varepsilon}^{-1} \log \tilde{\varepsilon}^{-1} \vee \tilde{\varepsilon}^{-2} \log \tilde{\varepsilon}^{-1}\Big)$$

iterations produces weights satisfying $\theta_\kappa^\top \hat{w}_{\kappa,n} \geq 1 - \tilde{\varepsilon}$ for $\Theta(N_\kappa)$ neurons, with high probability. Under the scaling $\tilde{\varepsilon} = \tilde{\Theta}(s^{-1/2})$, this reduces to $T = \tilde{O}(s^{3/2})$.

Remark. The condition $\tilde{\varepsilon} = \tilde{\Omega}(\mathcal{E}_s)$ means the strong model cannot achieve alignment error below $\mathcal{E}_s$. The three components of $\mathcal{E}_s$ reflect three sources of error: the L^1 approximation error ε_w and alignment error $\tilde{\varepsilon}_w$ of the weak model (Assumption 2), and the mixture weight $1 - \lambda_\kappa$ on non-target components of μ (Assumption 3). Our guarantees apply in the regime $\mathcal{E}_s \lesssim s^{-1/2}$. Assumption 5 (i)

assumes σ_κ^* is even, giving $\mathrm{GE}(\sigma_\kappa^*) = 2$; this is essential for feature preservation, as discussed in Remark 2. Under Assumption 3, inputs concentrate on the s-dimensional subspace V_κ, so the sample complexity $\tilde{O}(s^{3/2})$ depends only on s, not on d, which is lower than the $\tilde{O}(d^{p/2}\, s^{(p-2)/2})$ cost of the weak model (Proposition 1). When $s < d^{2/3}$, this complexity $\tilde{O}(s^{3/2})$ improves on the information-theoretic lower bound $\tilde{\Omega}(d)$ for Gaussian single-index models (Dudeja and Hsu, 2024; Damian et al., 2024b). The same holds unconditionally against the CSQ lower bound $\tilde{\Omega}(d^{p/2})$ (Damian et al., 2022), since $s \le d$ and $p > 2$. The gap between $\tilde{O}(s^{3/2})$ and the information-theoretic limit $\tilde{\Omega}(s)$ in $\mathbb{R}^s$ reflects the cost of imperfect concentration of μ on V_κ ($\lambda_\kappa < 1$); see Remark 1.

Burns et al. (2023) find that early stopping is necessary to prevent overfitting to the weak teacher, motivating our analysis of feature preservation. Medvedev et al. (2025) provide a theoretical account of this phenomenon in the kernel regime. The following proposition shows that Algorithm 1 prevents forgetting of off-target features at termination, providing a theoretical justification for early stopping.

Proposition 2 (Preservation of pre-trained features; informal)**.** Under the conditions of Theorem 1, let $\chi_k = N_\kappa \pi_k / (N_k \pi_\kappa)$ for $k \neq \kappa$. If $\chi_k \le 1/\mathrm{polylog}(d)$, then at termination of Algorithm 1, with high probability, for every $k \neq \kappa$ and every $n \in [N_k]$,

$$\theta_k^\top \hat{w}_{k,n} \ge \theta_k^\top w_{k,n}^0 - \tilde{O}\Big(\underbrace{\|\Sigma_\kappa w_{k,n}^0\|^2}_{\text{init. overlap with } V_\kappa} + \tilde{\varepsilon} + \underbrace{\chi_k s^{-1/2}}_{\text{SGD noise}} \Big).$$

where $\tilde{\varepsilon} \gtrsim \mathcal{E}_s$ is as in Theorem 1. The full bound is given in Section E.

The factor χ_k measures the localization of task k relative to task κ and controls the magnitude of the gradient which is induced on off-target neurons. The condition $\chi_k \lesssim 1/\mathrm{polylog}(d)$ is satisfied when task k occupies more neurons than task κ, carries smaller weight π_k in the reward, or both.

W2S Generalization. Theorem 1 and Proposition 2 together establish that Algorithm 1 achieves feature elicitation on task κ and preservation of pre-trained features for $k \neq \kappa$ simultaneously. The strong model surpasses the weak model in that it acquires θ_κ via W2S training while retaining its initial knowledge of off-target tasks $k \neq \kappa$, which the weak model, specialized to task κ alone, does not possess. For instance, task κ may correspond to a specialized alignment criterion, while off-target tasks $k \neq \kappa$ represent broader capabilities related to κ acquired during pre-training. We quantify W2S generalization in terms of the feature alignments $\theta_k^\top \hat{w}_{k,n}$ for $k \in [K]$. Our results are complementary to analyses that assume fixed strong-model representations (Charikar et al., 2024; Xue et al., 2025; Dong et al., 2025; Wu and Sahai, 2025; Medvedev et al., 2025), in that we establish that pre-trained features are elicited through W2S training, whereas these analyses treat feature quality as given. Unlike Moniri and Hassani (2025) and Oh et al. (2025), our analysis derives how a subspace prior enables feature elicitation under weak supervision from gradient-based optimization, tracking multi-step dynamics in a multi-neuron network and yielding quantitative bounds on both the gain in target alignment and the change in off-target feature directions. The bound on forgetting is enabled by the localization parameter χ_k, which emerges from the multi-neuron architecture and has no counterpart in prior works.

Remark. Empirically, Burns et al. (2023) find that generative finetuning, which trains the strong model on task-related text via a language modeling objective without human labels, prior to W2S training, improves W2S generalization. This can be interpreted as concentrating the strong model's generative distribution on task-κ inputs, corresponding to the concentration $1 - \lambda_\kappa = o(s^{-1/2})$, which limits the achievable alignment accuracy in Theorem 1. Generative finetuning may, however, pull off-target neurons toward V_κ, increasing $\|\Sigma_\kappa w_{k,n}^0\|^2$ and worsening the bound in Proposition 2.

3.2 Proof sketch of Theorem 1

Intuitively, the gradient dynamics keep $w_{\kappa,n}^t$ approximately within V_κ throughout training, so θ_κ is learned without leakage outside the subspace. Two quantities govern the training dynamics, the alignment $z^t := \theta_\kappa^\top w_{\kappa,n}^t$ and the complement-subspace deviation $D^t := \|\Sigma_\kappa^\perp w_{\kappa,n}^t\|^2$. The following lemma gives a decomposition of the negative gradient g_κ^t with respect to $w_{\kappa,n}^t$.

Lemma (Gradient decomposition; informal, Lemma 22)**.** $g_\kappa^t = y^t \tilde{\nabla}_{w_\kappa^t} r_\Theta(x^t)$ decomposes as

$$g_\kappa^t = \bar{\alpha}_2 \beta_{\kappa,n,2} \lambda_\kappa z^t\, \theta_\kappa + Z^t + R^t,$$

where $\bar{\alpha}_2, \beta_{\kappa,n,2}$ are the second Hermite coefficients of $\bar{r}^w \exp(\bar{r}^w)$ and $a_{\kappa,n}\sigma_\kappa(\cdot + b_{\kappa,n})$, Z^t is mean-zero, and $\|R^t\| = \tilde{O}((\mathcal{E}_s \vee D^t)\, s^{1/2})$.

Since $\mathrm{GE}(\sigma_\kappa^*) = 2$, Lemma 18 gives $\bar{\alpha}_2 \neq 0$, so the alignment satisfies $z^{t+1} \approx (1+\eta^t \bar{\alpha}_2 \beta_{\kappa,n,2} \lambda_\kappa)\, z^t$. By contrast, online SGD on r_κ^* has drift $(z^t)^{p-1}$ (Ben Arous et al., 2021). The W2S signal reduces this to a linear drift, which accounts for the improved sample complexity. The following lemma bounds D^t over each sub-interval of training.

Lemma (Subspace deviation bound; informal, Lemma 24)**.** For any step t with $z^t \geq 0$, $D^{t+1} - D^t = \tilde{O}(\eta^t)$. Moreover, over any sub-interval $[\tau, \tau+\tau']$ with $z^t \geq 0$ for all $t \in [\tau, \tau+\tau']$, with probability $1-\delta'$, $D^{\tau+t} \leq Q^t$ for all $t = 0, \ldots, \tau'$, where

$$Q^0 = \underbrace{D^\tau}_{\text{initial}} + \underbrace{\tilde{O}(\eta\, \mathcal{E}_s\, \tau')}_{\text{weak-model error}} + \underbrace{\Theta((1-\lambda_\kappa)^{1/2}\delta'^{-1/2}\eta\tau'^{1/2} + (1-\lambda_\kappa)\delta'^{-1}\eta^2 s\tau')}_{\text{off-subspace perturbation}},$$
$$Q^t = Q^{t-1} + \tilde{O}(\eta (Q^{t-1})^2).$$

Here $\delta' = 1/\operatorname{polylog} d$.

Controlling this bound requires Doob's maximal inequality rather than standard concentration, which forces a smaller learning rate and drives the overall complexity to $\tilde{O}(s^{3/2})$; see Remark 1 for details. A large D^t reduces the effective drift on z^t to $\bar{\alpha}_2\beta_{\kappa,n,2}\lambda_\kappa z^t - O(D^t)$ (Lemma 22), so the condition $D^t = O(z^t)$ must be maintained throughout training. Since the per-step increment of D^t is $O(\eta(D^t)^2)$ (Lemma 24) while z^t grows at rate $\Omega(\eta z^t)$ (Lemmas 25 and 27), the ratio D^t/z^t is non-increasing during the initial phase where $z^t = o(1)$. Lemmas 25 and 27 shows that the condition $D^t = o(z^t)$ is maintained throughout each sub-interval with probability $1 - o(1)$, using the bound on D^t established in Lemma 24.

Alignment dynamics. Three stages complete the argument (Lemmas 25, 27 and 29 to 31), following Lee et al. (2024); Oko et al. (2024a). In the first two stages the learning rate is $\eta_1 \lesssim s^{-3/2}$. In the third it is reduced to $\eta_2 \lesssim \tilde{\varepsilon} s^{-1} \wedge \tilde{\varepsilon}^2$. In weak recovery (Lemma 25), starting from $z^0 \geq s^{-1/2}$ (Assumption 5), each interval of $\tau_1 = O(s^{-1/2}\eta_1^{-1})$ steps multiplies z^t by $1 + \frac{1}{2}\bar{\alpha}_2\beta_{\kappa,n,2}\eta_1\tau_1 = 1 + \Theta(1)$, so $O(s^{1/2}\log s)$ intervals suffice to reach $z^t = 1/\operatorname{polylog} d$, totaling $T_1 = \tilde{O}(\eta_1^{-1})$ steps. In alignment amplification (Lemma 27), z^t is driven from $1/\operatorname{polylog} d$ to $1 - o(1)$ in $T_2 = \tilde{O}(\eta_1^{-1})$ steps. In strong recovery (Lemmas 29 and 32), the Hermite structure condition in Assumption 5 ensures that g_κ^t has a component along $-\Sigma_\kappa^\perp w_{\kappa,n}^t$ of magnitude $\Omega(z^t)$, so the gradient update shrinks $\|\Sigma_\kappa^\perp w_{\kappa,n}^{t+1}\|$ by a factor $1 - \Omega(\eta_2)$ when $z^t \geq 1/2$. The following lemma quantifies this decay.

Lemma (Subspace deviation decay; informal, Lemma 29)**.** Suppose $\theta_\kappa^\top w_{\kappa,n}^t \geq 1/2$ and $\eta_2 \lesssim \tilde{\varepsilon} s^{-1} \wedge \tilde{\varepsilon}^2$. Then

$$D^{t+1} \leq (1 - o(\eta_2))\, D^t + o(\tilde{\varepsilon}).$$

Consequently, D^t decays exponentially to $O(\tilde{\varepsilon})$ within $\tilde{O}(\eta_2^{-1}\log\tilde{\varepsilon}^{-1})$ steps (Lemmas 30 and 31). With $D^t = O(\tilde{\varepsilon})$ maintained, z^t converges to $1 - O(\tilde{\varepsilon})$ within T_3 steps (Lemmas 30 and 31). The total cost $T_1 + T_2 + T_3 = \tilde{O}(s^{3/2} \vee s\tilde{\varepsilon}\log\tilde{\varepsilon}^{-1} \vee \tilde{\varepsilon}^{-2}\log\tilde{\varepsilon}^{-1})$ gives Theorem 1.

Remark 1 (Why $s^{3/2}$ rather than s). For online SGD on a single-index target with $\mathrm{IE} = 2$, the standard complexity is $\tilde{O}(s)$ at learning rate $\eta \lesssim s^{-1}$ (Ben Arous et al., 2021). The difficulty stems from the heavy-tailed structure of $\Sigma_\kappa^\perp Z^t$. When $x^t \sim \mu_k$ for $k \neq \kappa$, the sample falls outside V_κ and causes an $O(1)$ perturbation, even though this event has probability $1 - \lambda_\kappa = O(s^{-1/2})$. Concentration inequalities control sums of martingale differences only through their per-step maximum, so the rare $O(1)$ jumps dominate and the resulting bound is too loose. Doob's maximal inequality, applied to a supermartingale in Lemma 24, instead exploits the low probability of the large-increment event and yields a tight bound. Since Doob's inequality bounds the running maximum of each martingale by its cumulative variance, the bound becomes vacuous if the interval is too long. The horizon is therefore partitioned into $\tilde{O}(s^{1/2})$ sub-intervals and a union bound is taken over all of them. This forces $\eta \lesssim s^{-3/2}$ rather than s^{-1}, yielding $\tilde{O}(s^{3/2})$ steps.

4 SFT forgetting

While Proposition 2 shows that W2S preserves off-target features, we show that SFT does not. This is consistent with empirical observations of catastrophic forgetting in large language models during fine-tuning (Wang et al., 2024; Kotha et al., 2024; Luo et al., 2025). In our setting, SFT on task κ applies online SGD to the same strong model (Section 2.4) with isotropic inputs $x^t \sim \mathcal{N}(0, I_d)$ and direct teacher labels $y^t = r_\kappa^*(\theta_\kappa^\top x^t) + \zeta^t$ as in Ben Arous et al. (2021). The first-layer update is

$$w_{k,n}^{t+1} \leftarrow w_{k,n}^t + \eta^t \, y^t \, \tilde{\nabla}_{w_{k,n}} r_\Theta(x^t), \quad w_{k,n}^{t+1} \leftarrow w_{k,n}^{t+1} / \|w_{k,n}^{t+1}\|.$$

Theorem 2 confirms that neurons initialized toward θ_k are driven to align with θ_κ whenever $\theta_k^\top \theta_\kappa$ is non-negligible, so the pre-trained feature θ_k is lost before task κ is learned. The role of feature similarity in causing forgetting is also examined by Hiratani (2024) in the continual learning setting.

Theorem 2 (SFT forgetting; informal)**.** Assume Assumptions 1 and 5. Let $t_{1,\kappa}$ be the first time at which $\theta_\kappa^\top w_{\kappa,n}^t \geq o(1)$ for some $n \in [N_\kappa]$. Under Algorithm 3, with high probability $\theta_\kappa^\top w_{k,n}^t \geq 1 - o(1)$ for every neuron $n \in [N_k]$ satisfying the sign condition, all $t \geq t_{1,\kappa}$, and every $k \neq \kappa$ satisfying Assumption 6 (see Section F for full conditions).

The off-target neurons thus reach θ_κ no later than the task-κ neurons, so the alignment $\theta_k^\top w_{k,n}^t$ with θ_k is lost once task κ is learned. Under the random initialization in Section E.2, $\Theta(N_k)$ neurons satisfy the sign condition, so a constant fraction of task-k neurons is affected. Since r^w carries no signal for $k \neq \kappa$, this loss of alignment is irrecoverable, affecting the approximation quality of r_k^*.

Proof idea. The base online SGD analysis follows the same framework as Proposition 1 and Theorem 1 and prior work (Ben Arous et al., 2021; Damian et al., 2024a; Lee et al., 2024; Oko et al., 2024a). Under isotropic sampling the gradient propagates to every neuron, and the Hermite expansion of the direct teacher has $\mathrm{IE}(\sigma_\kappa^*) = p$, so the drift on $\theta_\kappa^\top w_{k,n}^t$ scales as $\chi_k(\theta_\kappa^\top w_{k,n}^t)^{p-1}$. By Assumption 6, $\theta_\kappa^\top w_{k,n}^0$ exceeds the $\Theta(s^{-1/2})$ alignment of the target neurons $w_{\kappa,n}^0 \in V_\kappa$. A Bihari–LaSalle ODE comparison (formalized in Section F) shows that the larger initial condition forces $\theta_\kappa^\top w_{\kappa,n}^t$ to reach $1 - o(1)$ before $\theta_\kappa^\top w_{\kappa,n}^t$ does, so the alignment $\theta_k^\top w_{k,n}^t$ with the pre-trained feature is lost.

W2S versus SFT. For the same model architecture and pretrained initialization, W2S offers advantages in sample complexity and feature preservation over SFT using online SGD with isotropic inputs and direct target labels. Unlike SFT, which trains on a fixed labeled dataset, W2S treats the weak model as an oracle and queries it on inputs drawn from the strong model's generative distribution μ. W2S further applies a nonlinear transformation to the weak model output, achieving feature elicitation with $\tilde{O}(s^{3/2})$ samples. Under isotropic sampling, SFT with the same pretrained initialization learns the target feature in $\tilde{O}(d^{p/2} s^{(p-2)/2})$ steps (Ben Arous et al., 2021). Under the intended setting where task-k neurons satisfy $\theta_k^\top w_{k,n}^0 \approx 1$, Assumption 6 reduces to $\theta_k^\top \theta_\kappa \gtrsim s^{-1/2}$, meaning forgetting occurs when tasks k and κ are more similar than two random directions in the s-dimensional subspace. Even in this setting, Proposition 2 guarantees preservation of θ_k, since $\|\Sigma_\kappa w_{k,n}^0\|^2 \gtrsim (\theta_\kappa^\top \theta_k)^2 \gtrsim s^{-1}$, in contrast to SFT, under which the alignment $\theta_k^\top w_{k,n}^t$ with the pre-trained feature is lost.

Remark. Even if a nonlinear teacher transformation enables SFT under isotropic sampling to attain the $p = 2$ scaling of $\tilde{O}(d^{p/2} s^{(p-2)/2})$, namely $\tilde{O}(d)$, the $\tilde{O}(s^{3/2})$ bound for Algorithm 1 is smaller when $s < d^{2/3}$. This comparison illustrates the additional benefit of sampling from the target-concentrated distribution μ.

5 Numerical experiments

We verify Theorems 1 and 2 and Proposition 2 numerically with input dimension $d = 1024$, subspace dimension $s = 128$, number of tasks $K = 2$ with target task $\kappa = 1$, neuron counts $N_1 = 64$ and $N_2 = 512$, link function $\sigma_k^* = \mathrm{He}_4/\sqrt{4!}$ (IE $= 4$, GE $= 2$), feature correlation $\theta_1^\top \theta_2 = 0.3$, and mixture weight $\lambda_\kappa = 0.9$, so that $1 - \lambda_\kappa = 0.1 \lesssim s^{-1/2}$ as required by Assumption 3. Our results hold for general K. We use $K = 2$ to illustrate learning of θ_1 and preservation of alignment with θ_2. The weak model is a two-layer network with $N^w = 3$ neurons whose coefficients a_n and biases

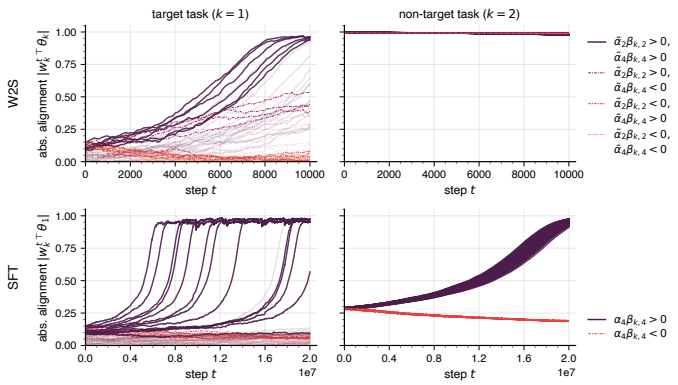


Figure 1: Per-neuron alignment magnitude $|\theta_k^\top w_{k,n}^t|$ during training ($d = 1024$, $s = 128$, $K = 2$, $\sigma_k^* = \mathrm{He}_4/\sqrt{4!}$, $\theta_1^\top \theta_2 = 0.3$). Line colors distinguish neuron types by the signs of $\tilde{\alpha}_2\beta_{k,2}$ and $\tilde{\alpha}_4\beta_{k,4}$ (W2S) or $\tilde{\alpha}_4\beta_{k,4}$ alone (SFT); neurons with initial alignment magnitude $|\theta_k^\top w_{k,n}^0| < s^{-1/2}$ are shown semi-transparent. **Top (W2S, $\eta = 0.2$, $T = 10000$):** Absolute alignment $|\theta_1^\top w_{1,n}^t|$ with the target feature (left) and $|\theta_2^\top w_{2,n}^t|$ with the off-target feature (right). **Bottom (SFT, $\eta = 7.5 \times 10^{-5}$, $T = 2 \times 10^7$):** Absolute alignment $|\theta_1^\top w_{k,n}^t|$ for both task-1 (left) and task-2 (right) neurons.

b_n are set to the true values of r_1^* plus independent $\mathcal{N}(0, 0.1^2)$ noise, and whose directions w_n are similarly perturbed and then divided by $\sqrt{d}$; this construction is intended to satisfy Assumption 2. The strong model is a two-layer network whose coefficients $a_{k,n}$ are initialized to ± 1 and whose biases $b_{k,n}$ are initialized to 0. Its directions $w_{k,n}$ are sampled uniformly on V_k and then perturbed by a small additive term. For $k = 1$, the perturbation is drawn from $\mathcal{N}(0, 0.1^2\Sigma_1^\perp)$ and normalized by $\sqrt{d - s}$ before being added. For $k = 2$, it is drawn from $\mathcal{N}(0, I_d)$ and normalized by $\sqrt{d}$. Neuron activations are drawn from $\{\pm\mathrm{He}_2 \pm \mathrm{He}_4\}$, chosen so that Assumption 5 is satisfied. We set the temperature parameter to $\rho = 1$ in the W2S experiment. The learning rate is $\eta = 0.2 \sim s^{-3/2}$ for W2S and $\eta = 7.5 \times 10^{-5} \sim d^{-p/2}$ for SFT, matching the theoretical scalings. Fig. 1 plots the per-neuron alignment magnitude $|\theta_k^\top w_{k,n}^t|$ over training steps.

Fig. 1 (top) verifies Theorem 1 and Proposition 2 under W2S. The left panel shows that task-1 neurons ($\kappa = 1$, target task) satisfying the sign condition of Proposition 1 with initial alignment $|\theta_1^\top w_{1,n}^0| \geq s^{-1/2}$ converge to alignment ≈ 0.95 with θ_1, in agreement with Theorem 1. The right panel shows that task-2 neurons (non-target) maintain alignment ≈ 1.0 with θ_2 throughout training without drifting toward θ_1, in agreement with Proposition 2. Fig. 1 (bottom) verifies Theorem 2 under SFT, with both panels tracking alignment with θ_1. The left panel shows that task-1 (target) neurons have not yet converged at $T = 2 \times 10^7$, consistent with the $\tilde{O}(d^2 s) \approx 10^8$ sample complexity predicted by Proposition 1. The right panel shows that task-2 neurons reach alignment ≈ 0.8 with θ_1 before the target neurons have fully converged, so forgetting of the pre-trained feature θ_2 precedes completion of the target task.

6 Discussion

Limitations. Learning the second layer of the strong model is not formally analyzed. Our guarantee for feature preservation concerns first-layer W2S with $a_{k,n}$ and $b_{k,n}$ held fixed, and does not establish preservation of r_k after refitting the second layer. This limitation is shared with Moniri and Hassani (2025), who similarly establishes preservation at the feature level. In an additional experiment, standard ridge refitting on target weak labels increased the MSE on task 2 more than first-layer W2S. Penalizing changes in $a_{k,n}$ reduced but did not eliminate this degradation (Section C.2). Theorem 2 is algorithm-specific rather than an information-theoretic lower bound, so forgetting can likely be avoided by other algorithms with explicit regularization. The evenness assumption on σ_κ^* is essential to our preservation analysis because a first-order Hermite component in the transformed teacher signal can drive off-target neurons toward θ_κ.

Future work. Preserving off-target rewards r_k when training the second layer remains an open problem. Following Oko et al. (2024a), the coefficients $a_{k,n}$ can be fitted by ridge regression. Since $r^w \approx r^*_\kappa$, fitting these coefficients to weak labels does not directly constrain preservation of r_k for $k \neq \kappa$. Resetting $b_{k,n}$ can also alter a previously learned r_k that was approximately preserved during W2S. Our additional experiment (Section C.2) motivates establishing preservation guarantees for training the second layer with penalties on changes in $a_{k,n}$. The analysis of deeper architectures such as Transformers is also a natural future direction.

Acknowledgments and Disclosure of Funding

RA was partially supported by JSPS KAKENHI (25H01107). TS was partially supported by JSPS KAKENHI (24K02905) and JST CREST (JPMJCR2115). This work was partially supported by JST ERATO (JPMJER2601). This research is supported by the National Research Foundation, Singapore and the Ministry of Digital Development and Information under the AI Visiting Professorship Programme (award number AIVP-2024-004). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and the Ministry of Digital Development and Information.

References

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *Conference on Learning Theory (COLT)*, volume 195, pages 2552–2623. PMLR, 2023.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report, 2023. arXiv:2303.08774.

Luca Arnaboldi, Yatin Dandi, Florent Krzakała, Luca Pesce, and Ludovic Stephan. Repetita iuvant: Data repetition allows SGD to learn high-dimensional multi-index functions. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 37932–37946, 2022.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: A spiked random matrix perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Daniel Beaglehole, Adityanarayanan Radhakrishnan, Enric Boix-Adserà, and Mikhail Belkin. Toward universal steering and monitoring of ai models. *Science*, 391(6787):787–792, 2026.

Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22 (106):1–51, 2021.

Gérard Ben Arous, Murat A Erdogdu, Nuri Mert Vural, and Denny Wu. Learning quadratic neural networks in high dimensions: SGD dynamics and scaling laws. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *Foundations of Computational Mathematics*, pages 1–84, 2024.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Nader H Bshouty and Vitaly Feldman. On using extended statistical queries to avoid membership queries. *Journal of Machine Learning Research*, 2(Feb):359–395, 2002.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-Strong Generalization: Eliciting strong capabilities with weak supervision, 2023. arXiv:2312.09390.

Seok-Ho Chang, Pamela C Cosman, and Laurence B Milstein. Chernoff-type bounds for the Gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944, 2011.

Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. Quantifying the gain in Weak-to-Strong Generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 126474–126499, 2024.

Sitan Chen and Raghu Meka. Learning polynomials in few relevant dimensions. In *Conference on Learning Theory (COLT)*, volume 125, pages 1161–1227. PMLR, 2020.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for SGD: Optimal sample complexity for learning single index models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2024a.

Alex Damian, Loucas Pillaud-Vivien, Jason Lee, and Joan Bruna. Computational-statistical gaps in Gaussian single-index models (extended abstract). In *Conference on Learning Theory (COLT)*, volume 247 of *Proceedings of Machine Learning Research*, pages 1262–1262, 30 Jun–03 Jul 2024b. Full version available at arXiv:2403.05529.

Alexandru Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory (COLT)*, volume 178, pages 5413–5452. PMLR, 2022.

Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborová, and Florent Krzakała. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. In *International Conference on Machine Learning (ICML)*, 2024.

Yijun Dong, Yicheng Li, Yunai Li, Jason D. Lee, and Qi Lei. Discrepancies are virtue: Weak-to-Strong Generalization through lens of intrinsic dimension. In *International Conference on Machine Learning (ICML)*, 2025.

Rishabh Dudeja and Daniel Hsu. Learning single-index models in Gaussian space. In *Conference on Learning Theory (COLT)*, volume 75, pages 1887–1930, 2018.

Rishabh Dudeja and Daniel Hsu. Statistical-computational trade-offs in tensor PCA and related problems via communication complexity. *The Annals of Statistics*, 52(1):131–156, 2024.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. arXiv:2209.10652.

Margalit Glasgow, Denny Wu, and Joan Bruna. Propagation of chaos in one-hidden-layer neural networks beyond logarithmic time, 2025. arXiv:2504.13110.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations (ICLR)*, 2024.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Naoki Hiratani. Disentangling and mitigating the impact of task similarity for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Muhammed Emrullah Ildiz, Halil Alperen Gozeten, Ege Onur Taga, Marco Mondelli, and Samet Oymak. High-dimensional analysis of knowledge distillation: Weak-to-Strong Generalization and scaling laws. In *International Conference on Learning Representations (ICLR)*, 2025.

Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Tianyi Qiu, Juntao Dai, and Yaodong Yang. Aligner: Efficient alignment by learning to correct. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Nirmit Joshi, Theodor Misiakiewicz, and Nathan Srebro. On the complexity of learning sparse functions with statistical and gradient queries. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *International Conference on Learning Representations (ICLR)*, 2024.

Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. Theoretical analysis of Weak-to-Strong Generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 46837–46880, 2024.

Jason D. Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with SGD near the information-theoretic limit. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 58716–58756, 2024. doi: 10.52202/079017-1872.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025.

Arvind Mahankali, Haochen Zhang, Kefan Dong, Margalit Glasgow, and Tengyu Ma. Beyond NTK with vanilla gradient descent: A mean-field analysis of neural networks with polynomial width, samples, and time. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Marko Medvedev, Kaifeng Lyu, Dingli Yu, Sanjeev Arora, Zhiyuan Li, and Nathan Srebro. Weak-to-Strong Generalization even in random feature networks, provably. In *International Conference on Machine Learning (ICML)*, 2025.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Language models implement simple word2vec-style vector arithmetic. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

Tomás Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2013.

Behrad Moniri and Hamed Hassani. On the mechanisms of Weak-to-Strong Generalization: A theoretical perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. In *International Conference on Machine Learning (ICML)*, volume 235, pages 36106–36159. PMLR, 2024.

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. In *International Conference on Learning Representations (ICLR)*, 2022.

Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A. Erdogdu. Gradient-based feature learning under structured data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *BlackboxNLP Workshop at Empirical Methods in Natural Language Processing (BlackboxNLP@EMNLP)*, 2023.

Naoki Nishikawa, Yujin Song, Kazusato Oko, Denny Wu, and Taiji Suzuki. Nonlinear transformers can perform inference-time feature learning. In *International Conference on Machine Learning (ICML)*, 2025.

Junsoo Oh, Jerry Song, and Chulhee Yun. From linear to nonlinear: Provable Weak-to-Strong Generalization through feature learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Learning sum of diverse features: computational hardness and efficient gradient-based training for ridge combinations. In *Conference on Learning Theory (COLT)*, volume 247, pages 4009–4081, 2024a.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Pretrained transformer efficiently learns low-dimensional target functions in-context. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. Task-specific skill localization in fine-tuned language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. URL `https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf`.

Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D. Lee. Emergence and scaling laws in SGD learning of shallow neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Changho Shin, John Cooper, and Frederic Sala. Weak-to-Strong Generalization through the data-centric lens. In *International Conference on Learning Representations (ICLR)*, 2025.

Berfin Simsek, Amire Bendjeddou, and Daniel Hsu. Learning Gaussian multi-index models with gradient flow: Time complexity and directional convergence. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Leitian Tao and Yixuan Li. Your weak LLM is secretly a strong teacher for alignment. In *International Conference on Learning Representations (ICLR)*, 2025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. arXiv:2308.10248.

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

Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix X. Yu, Cho-Jui Hsieh, Inderjit S. Dhillon, and Sanjiv Kumar. Two-stage LLM fine-tuning with less specialization and more generalization. In *International Conference on Learning Representations (ICLR)*, 2024.

David Xing Wu and Anant Sahai. Provable Weak-to-Strong Generalization via benign overfitting. In *International Conference on Learning Representations (ICLR)*, 2025.

Yihao Xue, Jiping Li, and Baharan Mirzasoleiman. Representations shape Weak-to-Strong Generalization: Theoretical insights and empirical predictions. In *International Conference on Machine Learning (ICML)*, 2025.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the contributions and scope. The main claims are formalized in Theorems 1 and 2 and Propositions 1 and 2.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in the Limitations paragraph of Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All assumptions are stated in Sections 2 to 4 and referenced in each theorem. Full proofs are given in Sections D to F, with proof sketches in the main body.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 specifies all hyperparameters and initializations needed to reproduce the numerical experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The experiments are small-scale toy simulations that do not require code beyond standard numerical computation.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 specifies all architecture choices, initialization schemes, and hyperparameters necessary to understand and reproduce the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The experiments are numerical verifications of theoretical predictions. No statistical significance testing is required as the results directly illustrate the learning dynamics predicted by the theorems.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section C in the appendix describes the hardware and wall-clock times for all experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is primarily theoretical and no immediate societal impact is expected.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release models, code, or datasets, and poses no misuse risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing code, datasets, or pretrained models.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new code, datasets, or models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only for writing assistance and did not contribute to the core methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Contents

1 Introduction **1**

2 Problem setting **3**

2.1 Sample complexity and the information exponent 3

2.2 Additive reward model 3

2.3 Weak model 4

2.4 Strong model 4

3 W2S learning **5**

3.1 Main results 5

3.2 Proof sketch of Theorem 1 6

4 SFT forgetting **8**

5 Numerical experiments **8**

6 Discussion **9**

A Notations **23**

B Additional related work **24**

C Experimental details and additional results **24**

C.1 Experimental setup and computational resources 24

C.2 Second-layer refitting and off-target reward error 24

D Full proof of Proposition 1: weak model construction **25**

D.1 Initialization 27

D.2 Gradient decomposition 28

D.3 Weak alignment 29

D.4 Amplification of alignment 31

D.5 Strong alignment and localization 32

D.6 Neuron filtering 33

D.7 Second-layer training 34

E Full proof of Theorem 1: W2S feature learning **35**

E.1 Nonlinear Transformation of the Teacher Signal 37

E.2 Initialization 38

E.3 Error of the Teacher Signal in the Subspace 38

E.4 Gradient Decomposition 39

E.5 Upper Bound on Complement-Subspace Deviation 43

E.6 Weak Alignment 44

E.7 Amplification of Alignment 46
E.8 Strong Alignment 47
E.9 Non-Forgetting of Other Tasks 49

F Full proof of Theorem 2: SFT forgetting **51**

G Technical lemmas **55**

A Notations

Symbol	Meaning
Geometry and tasks	
d, K, κ	Ambient input dimension, number of tasks, and target task index.
$s = d^\alpha$	Dimension of each task subspace V_k, with $\alpha \in (0, 1)$.
θ_k	True hidden feature direction for task k; $\theta_k \in V_k \cap S^{d-1}$.
Σ_k	Orthogonal projection onto V_k.
$\Sigma_k^\perp$	Projection onto the orthogonal complement, $\Sigma_k^\perp = I - \Sigma_k$.
Reward model	
π_k	Reward weight for task k.
σ_k^*	True link function for task k; a degree-q polynomial.
p, q	Information exponent $p = \mathrm{IE}(\sigma_\kappa^*)$ and common degree of σ_k^*, σ_n^w, $\sigma_{k,n}$.
$r^*(x)$	True total reward, $r^*(x) = \sum_{k=1}^K \pi_k \sigma_k^*(\theta_k^\top x)$.
Weak and strong models	
$r^w(x)$, $r(x)$	Outputs of the weak model and the strong model, respectively.
N^w, N_k	Width of the weak model and of the k-th neuron group of the strong model.
σ_n^w, $\sigma_{k,n}$	Activation of neuron n in the weak model and in the k-th group of the strong model.
w_n^w, a_n^w, b_n^w	First-layer weight, second-layer coefficient, and bias of neuron n in the weak model.
$w_{k,n}^t, a_{k,n}, b_{k,n}$	First-layer weight, second-layer coefficient, and bias of neuron n in group k of the strong model at step t.
$\bar{r}_\kappa^*$, ρ	Clipped teacher signal $\bar{r}_\kappa^* = \mathrm{clip}(\rho^{-1} r_\kappa^*; \pm 1/\log d)$ and its temperature $\rho = \Theta(\log^{C_\rho} d)$.
$\alpha_i, \bar{\alpha}_i$	i-th Hermite coefficients of r_κ^* and $\bar{r}_\kappa^* \exp(\bar{r}_\kappa^*)$.
$\beta_{n,i}$	i-th Hermite coefficient of $a_n \sigma_n(\cdot + b_n)$.
$\tilde{\beta}_{k,n,i}$	i-th Hermite coefficient of $a_{k,n}\sigma_{k,n}(\cdot + b_{k,n})$ (before rescaling).
$\beta_{k,n,i}$	Rescaled coefficient, $\beta_{k,n,i} = \pi_k^{-1} \tilde{\beta}_{k,n,i}$.
Training dynamics	
η^t, T	Learning rate at step t and total iteration count.
$\tilde{\nabla}_w$	Riemannian gradient on S^{d-1} at w, defined by $\tilde{\nabla}_w f = (I - ww^\top)\nabla_w f$.
g_k^t	Rescaled negative gradient with respect to w_k^t at step t.
χ_k	Cross-task gradient ratio $\chi_k = N_\kappa \pi_k / (N_k \pi_\kappa)$; governs how task-$k$ neurons respond to target-task supervision.
Error and constants	
ε_w, $\tilde{\varepsilon}_w$	Weak-model L^1 error and alignment slack (Assumption 2).
μ, μ_k, λ_k	Oracle input distribution $\mu = \sum_k \lambda_k \mu_k$, where μ_k is the task-k probability measure on V_k and $(\lambda_k)_{k=1}^K \in \Delta^{K-1}$.
C_a, C_b	Universal bounds on neuron parameters, $O(1)$ and $\tilde{O}(1)$ respectively.
c_β, C_β	Constants bounding the Hermite coefficients of neuron activations, $c_\beta^{-1}, C_\beta = \mathrm{polylog}(d)$; see Assumptions 4 and 5.

B Additional related work

Linear representation hypothesis. The linear representation hypothesis posits that high-level concepts are encoded as linear subspaces in the representation space of neural networks (Mikolov et al., 2013; Arora et al., 2016; Elhage et al., 2022; Park et al., 2024). This structure has been observed in Transformer LLMs across diverse domains, including relational knowledge (Merullo et al., 2024; Hernandez et al., 2024) and spatial and temporal concepts (Gurnee and Tegmark, 2024). A related body of work shows that these representations can be linearly probed and used to steer model behavior (Nanda et al., 2023; Turner et al., 2023). At scale, Beaglehole et al. (2026) extract linear subspace representations for over 500 concepts in large-scale AI models, providing direct evidence for our assumption that the strong model's pre-training knowledge is organized into task subspaces V_k.

Feature learning for single-index and additive models. Online-SGD analyses of single-index models (Ben Arous et al., 2021; Ba et al., 2022; Mousavi-Hosseini et al., 2022; Ba et al., 2023; Mousavi-Hosseini et al., 2023; Moniri et al., 2024; Mahankali et al., 2023; Berthier et al., 2024; Damian et al., 2024a; Glasgow et al., 2025) establish the information-exponent sample complexity $\tilde{O}(d^{p-1})$ for $p > 2$, with analogous guarantees in the misspecified setting (Oko et al., 2024a). A line of recent work (Chen and Meka, 2020; Dandi et al., 2024; Lee et al., 2024; Arnaboldi et al., 2024; Joshi et al., 2024; Damian et al., 2024b) shows that nonlinear transformations of the teacher signal replace the information exponent with the generative exponent, circumventing the $\Omega(d^{p/2})$ correlational statistical query (CSQ) (Bshouty and Feldman, 2002) lower bound of Damian et al. (2022); Algorithm 1 exploits this mechanism. In particular, Chen and Meka (2020) achieve $\tilde{O}(d)$ sample complexity for polynomial link functions, near the information-theoretic limit $\tilde{\Omega}(d)$ (Dudeja and Hsu, 2024; Damian et al., 2024b). Additive-model extensions, which underlie our reward model, appear in Oko et al. (2024a); Simsek et al. (2025); Ren et al. (2025); Ben Arous et al. (2025).

C Experimental details and additional results

C.1 Experimental setup and computational resources

Both experiments were run on a shared laboratory cluster equipped with an Intel Xeon E5-2680 v4 processor (28 cores, 2.40 GHz) and 503 GB RAM, using CPU only. The W2S experiment ($T = 10{,}000$ steps) completed in a few minutes and the SFT experiment ($T = 2 \times 10^7$ steps) required approximately 10 hours. No preliminary or failed runs required substantially more compute than the reported results.

C.2 Second-layer refitting and off-target reward error

To use the learned target features for reward prediction, we consider fitting the second-layer coefficients to target-task weak labels. Our theory establishes target-feature learning and off-target feature preservation during first-layer W2S with fixed second-layer coefficients and biases (Theorem 1 and Proposition 2). It does not analyze second-layer learning, which can alter the coefficients associated with off-target features even when their directions are preserved. We therefore examine numerically how first-layer W2S and subsequent second-layer refitting affect target and off-target reward errors, and whether penalizing coefficient changes mitigates off-target degradation.

We use two tasks with target task $\kappa = 1$, normalized link functions $\sigma_1^* = \sigma_2^* = \mathrm{He}_4/\sqrt{4!}$, and reward weights $\pi_1 = \pi_2 = 1$. The dimensions, neuron counts, and W2S training parameters follow Section 5 and Fig. 1. Task-2 first-layer weights are initialized near θ_2 by adding isotropic Gaussian noise with expected squared norm 0.1^2 and then normalizing each weight to unit norm. We first fit only the task-2 second-layer coefficients to $r_2^*(\theta_2^\top x)$ by ridge regression with coefficient 10^{-3}, so that the model starts with low task-2 error. We then train the first layer by W2S on task 1 for $10{,}000$ iterations, keeping the second-layer coefficients fixed. Finally, we freeze the first layer and jointly refit the second-layer coefficients of both tasks to the untransformed weak output $r^w(x)$, using the total model output $r_\Theta(x)$. The strong-model biases remain fixed at zero throughout the experiment. Each regression uses a separately drawn set of $4{,}096$ independent inputs from $\mathcal{N}(0, I_d)$.

Let a denote the vector of second-layer coefficients and a^{old} its value immediately before the final refitting step. Both refitting methods minimize the sum of squared errors $\sum_{t=1}^{4096}(r_\Theta(x^t) - r^w(x^t))^2$

plus a regularization penalty. Standard ridge uses $\lambda\|a\|_2^2$, while the alternative replaces this penalty with $\lambda\|a - a^{\mathrm{old}}\|_2^2$.

For task k, the mean squared error (MSE) is the empirical mean of $(r_k(x) - r_k^*(\theta_k^\top x))^2$, where r_k is the output of the task-k neuron group in Section 2.4. For each seed, we draw $8{,}192$ held-out inputs from $\mathcal{N}(0, I_d)$ and reuse this evaluation set at all stages. Table 1 summarizes the results.

Table 1: Task errors before and after second-layer refitting, reported as mean $\pm$ standard deviation over three seeds.

Stage or refitting method	Penalty strength λ	Task 1 (target) MSE	Task 2 (off-target) MSE
After initial task-2 ridge	—	1.1560 ± 0.1214	0.0001 ± 0.0000
After first-layer W2S	—	$\mathbf{1.0301} \pm 0.0968$	$\mathbf{0.1479} \pm 0.0104$
Standard ridge	0.001	0.5589 ± 0.0652	0.9310 ± 0.1100
	0.01	0.5593 ± 0.0643	0.9199 ± 0.1115
	0.1	0.5739 ± 0.0554	0.9207 ± 0.1140
	1	0.6464 ± 0.0442	0.9196 ± 0.1172
Deviation penalty	2	0.7222 ± 0.0465	0.4914 ± 0.0810
	3	0.7643 ± 0.0513	0.3975 ± 0.0665
	5	$\mathbf{0.8200} \pm 0.0597$	$\mathbf{0.2944} \pm 0.0466$
	7	0.8557 ± 0.0658	0.2423 ± 0.0345

First-layer W2S increases the task-2 MSE from 0.0001 to 0.1479, while standard ridge refitting increases it further to 0.9196–0.9310 across all tested coefficients. Thus, in this setting, the larger increase in off-target error occurs during second-layer refitting. Increasing the penalty on coefficient changes reduces the task-2 error, but also reduces the improvement on task 1. All tested deviation penalties improve the target MSE relative to its value before refitting, while none restores the off-target MSE to its value before refitting.

These results show that preservation of first-layer features alone does not ensure preservation of off-target reward prediction after second-layer learning, and motivate additional control of the second-layer coefficients.

D Full proof of Proposition 1: weak model construction

We prove Proposition 1. The arguments in Section D.1, Section D.2, Section D.4, Section D.5, and Section D.7 build on the discussion in Oko et al. (2024a).

Assumption 4. When training the weak model, we impose the following conditions.

(i) The weak model learns the single-index model $r_\kappa^* = \pi_\kappa \sigma_\kappa^*$ from observations

$$x \sim \mathcal{N}(0, I_d), \quad y = r_\kappa^*(x) + \zeta_\kappa, \quad \zeta_\kappa \sim \mathcal{N}(0, \pi_\kappa^2),$$

where r_κ^* is the reward for task $\kappa \in [K]$ defined in Assumption 1.

(ii) The weak model $r_{\tilde{\Theta}}^w$ is initialized with $\tilde{N}^w \gtrsim N^w$ neurons:

$$\frac{1}{\tilde{N}^w} \sum_{n=1}^{\tilde{N}^w} a_n^0 \sigma_n^w(w_n^{0\top} x + b_n^0).$$

Each first-layer weight satisfies $\|w_n^0\| = 1$ and is drawn uniformly from the subspace V_κ, i.e., $w_n^0 \sim \mathrm{Unif}(S^{d-1} \cap V_\kappa)$. Writing $r_\kappa^*(\cdot) = \sum_{i=p}^q \frac{\alpha_i}{\sqrt{i!}} \mathrm{He}_i(\cdot)$ and $a_n^0 \sigma_n^w(\cdot + b_n) = \sum_{i=0}^q \frac{\beta_{n,i}}{\sqrt{i!}} \mathrm{He}_i(\cdot)$, the following hold for $\Theta(\tilde{N}^w)$ neurons:

$$\begin{aligned} &\pi_\kappa c_\beta \le |\beta_{n,p}| \le \pi_\kappa C_\beta, \quad |\beta_{n,i}| \le \pi_\kappa C_\beta \text{ for } i \ne p, \\ &\alpha_i \beta_{n,i} > 0 \text{ for } i = p, \quad \alpha_i \beta_{n,i} \ge 0 \text{ for } p < i \le q. \end{aligned} \tag{1}$$

The most distinctive condition in Assumption 4 is $w_n^0 \sim \mathrm{Unif}(S^{d-1} \cap V_\kappa)$, which reflects the knowledge acquired through pre-training. The remaining conditions are standard in the misspecified

Algorithm 2: Weak model training by online SGD

Input: Initialized weak model r^w_Θ with $\Theta = (a_n, b_n, w_n^0)_{n=1}^{\tilde{N}^w}$; learning rate schedule η^t; regularization parameter λ; sample sizes T_1, T_2; bias scale C_b; filtering threshold $\tilde{\varepsilon}_w$.

Phase I: first-layer training

for $t = 0, 1, \ldots, T_1 - 1$ **do**

Draw $x^t \sim \mathcal{N}(0, I_d)$, $y^t = r^*_\kappa(\theta_\kappa^\top x^t) + \zeta$.

$w_n^{t+1} \leftarrow w_n^t + \eta^t y^t \tilde{\nabla}_w r^w_{(a_n, b_n, w_n^t)_{n=1}^{\tilde{N}^w}}(x^t)$

$w_n^{t+1} \leftarrow w_n^{t+1} / \|w_n^{t+1}\|, \quad (n = 1, 2, \ldots, \tilde{N}^w)$

$\hat{w}_n \leftarrow w_n^{T_1}$

Phase II: neuron filtering

$W_{\tilde{N}^w} \leftarrow \frac{1}{\tilde{N}^w} \sum_{n=1}^{\tilde{N}^w} \hat{w}_n \hat{w}_n^\top$

$\hat{\theta} \leftarrow$ leading unit eigenvector of $W_{\tilde{N}^w}$.

$\mathcal{N}^w \leftarrow \{n \in [\tilde{N}^w] \mid |\hat{\theta}^\top \hat{w}_n| \geq 1 - 2\tilde{\varepsilon}_w\}$

Re-initialize Drop neurons not in $\mathcal{N}^w$; set $\Theta = (a_n, b_n, \hat{w}_n)_{n=1}^{N^w}$ with $N^w = |\mathcal{N}^w|$. Resample $b_n \sim \mathrm{Unif}([-C_b, C_b])$.

Phase III: second-layer training

Draw $x^t \sim \mathcal{N}(0, I_d)$, $y^t = r^*_\kappa(\theta_\kappa^\top x^t)$, for $t = T_1, \ldots, T_1 + T_2 - 1$.

$\hat{a} \leftarrow \mathrm{argmin}_{a \in \mathbb{R}^{N^w}} \frac{1}{T_2} \sum_{t=T_1}^{T_1+T_2-1} \big(r_{(a_n, b_n, \hat{w}_n)_{n=1}^{N^w}}(x^t) - y^t\big)^2 + \lambda \|a\|^2$

Output: $r^w_{\hat{\Theta}}$ with $\hat{\Theta} = (\hat{a}_n, b_n, \hat{w}_n)_{n=1}^{N^w}$.

setting (Mousavi-Hosseini et al., 2023; Oko et al., 2024a; Lee et al., 2024). Condition (1) is needed to ensure the expected correlation loss is monotone in the alignment $\theta_\kappa^\top w_n$ (see Oko et al., 2024a, Appendix B.2), and is automatically satisfied in the well-specified case $\sigma^*_\kappa = \sigma^w_n$ (Ben Arous et al., 2021). In Section D.1, following Oko et al. (2024a); Lee et al. (2024), we show how to construct an initialization satisfying these conditions.

Training Algorithm Algorithm 2 describes the weak model training procedure. It follows the layer-wise training paradigm commonly used in feature learning theory (Damian et al., 2022; Ba et al., 2022; Bietti et al., 2022; Abbe et al., 2023; Mousavi-Hosseini et al., 2023).

In Phase I, the first-layer weights w_n are trained by online SGD to minimize the correlation loss $\mathcal{L} = -y r^w_\Theta(x)$ (Oko et al., 2024a; Lee et al., 2024), which matches the behavior of squared loss when the learning rate is sufficiently small (Lee et al., 2024). The update uses the spherical gradient $\tilde{\nabla}_w r(x) = (I - ww^\top)\nabla_w r(x)$, where ∇_w denotes the Euclidean gradient.

After Phase I, sufficiently many neurons satisfy $|\theta_\kappa^\top w_n^{T_1}| \geq 1 - \tilde{\varepsilon}_w$, but some do not. Phase II filters these out by computing the empirical covariance matrix of $(\hat{w}_n)_n$, extracting its leading eigenvector $\hat{\theta}$, and retaining only neurons with $|\hat{\theta}^\top \hat{w}_n| \geq 1 - 2\tilde{\varepsilon}_w$, which guarantees all remaining neurons satisfy Assumption 2 (i). The biases b_n are then re-randomized before Phase III, in which the second-layer coefficients a_n are learned by ℓ^2-regularized least squares regression.

Proposition 1 (Weak model construction; formal)**.** Under Assumptions 1 and 4, fix $\tilde{\varepsilon}_w \in (0, 3c_1)$. Algorithm 2 with the two-phase learning rate schedule

$$\eta^t = \begin{cases} \eta_1 \leq c_\eta\, d^{-p/2} & (0 \leq t \leq T_1 - 1), \\ \eta_2 \leq c_\eta\, \tilde{\varepsilon}_w\, d^{-1} \wedge c_\eta\, \tilde{\varepsilon}_w^2 & (T_1 \leq t \leq T_1 + T_2 - 1), \end{cases}$$

produces a weak model satisfying Assumption 2 with $\varepsilon_w = \tilde{\Theta}(\tilde{\varepsilon}_w)$, using a total of

$$T = \tilde{O}\Big(d^{p/2}\, s^{(p-2)/2} \vee d\, \tilde{\varepsilon}_w^{-2} \vee \tilde{\varepsilon}_w^{-3}\Big)$$

samples.

Proof. Setting $\tilde{\varepsilon} = \tilde{\varepsilon}_w/6$ in Lemma 11, Phase I produces neurons satisfying $\theta_\kappa^\top w^{T_1} > 1 - \tilde{\varepsilon}_w/2$. Setting $\bar{\varepsilon} = \tilde{\varepsilon}_w/2$ in Lemma 13, the neuron filtering step (Section D.6) retains neurons satisfying

$\theta_\kappa^\top \hat{w}_n \geq 1 - \tilde{\varepsilon}_w$, establishing Assumption 2(i). By Lemmas 14 to 16, second-layer ridge regression in Phase III produces $\hat{a}$ satisfying $\mathbb{E}_x[|r_{\hat{a}}(x) - r_\kappa^*(\theta_\kappa^\top x)|] \leq \pi_\kappa \varepsilon_w$, establishing Assumption 2(ii). Phase I uses $T_1 = \tilde{O}(d^{p/2} s^{(p-2)/2} \vee d\tilde{\varepsilon}_s^{-2} \vee \tilde{\varepsilon}_w^{-3})$ samples (Lemma 7, Section D.4, and Lemma 11) and Phase III uses $T_2 = \tilde{O}(\tilde{\varepsilon}_w^{-2})$ samples (Lemma 16). The total is $T = \tilde{O}(d^{p/2} s^{(p-2)/2} \vee d\tilde{\varepsilon}_s^{-2} \vee \tilde{\varepsilon}_w^{-3})$. □

Notations. Throughout this section, the task κ is fixed. Accordingly, we write $\theta = \theta_\kappa$ and $\alpha_i = \alpha_{\kappa,i}$ for all i. Moreover, during the training of the first layer, the gradient updating each w_n^t does not depend on the other neurons. Hence, in what follows we focus on a single neuron n, suppress the index n, and write $\sigma = \sigma_n^w$, $a = a_n^w$, $b = b_n^w$, $w^t = w_n^t$, and $\beta_i = \beta_{n,i}$ for all i.

The subsequent arguments are carried out after rescaling the learning rate in the stochastic gradient descent update rule appropriately. Specifically, by setting $\eta^t \leftarrow N^w \pi_\kappa^{-1} \eta^t$, the gradient $\eta^t y^t \tilde{\nabla}_w r_\Theta^w(x^t)$ is expressed as $\pi_\kappa^{-1} \eta^t \sum_{n=1}^{N^w} (\sigma_\kappa^*(\theta^\top x^t) + \zeta) a_n \sigma_n^w(w_n^{t\top} x^t + b_n)$ with $\zeta \sim \mathcal{N}(0,1)$.

We take constants of order $\mathrm{polylog}\, d$ satisfying the following ordering:

$$C_1 \lesssim c_1^{-1} \lesssim C_2 \lesssim c_\eta^{-1} = \tilde{O}(1).$$

The precise ordering of these constants will be determined so as to keep the subsequent proofs consistent. For example, an estimate such as $C_1^2 \leq C_2$ is allowed. These constants are only valid within the present section, and denote constants different from those defined analogously in Sections E and F. Furthermore, we assume d is sufficiently large, that c_1, c_η are smaller than any of the finitely many $\Theta(1)$ constants appearing in the subsequent proofs, while C_1 and C_2 are larger than any such constant.

D.1 Initialization

We verify that random initialization satisfies Assumption 4 with probability $\Omega(1)$. Following Oko et al. (2024a); Lee et al. (2024), we show that the initial alignment of w_n^0 is $s^{-1/2}$ with probability $\Theta(1)$, and that the sign condition (1) holds with probability $\Omega(1)$.

Lemma 1. Let $w^0 \sim \mathrm{Unif}(S^{s-1})$. Then,

$$\mathbb{P}\left[e_1^\top w^0 \geq s^{-\frac{1}{2}}\right] = \mathbb{P}\left[e_1^\top w^0 \leq -s^{-\frac{1}{2}}\right] = \Theta(1).$$

Lemma 2 (Chang et al. (2011), Theorem 2)**.** Let $\beta > 1$ and $a \in \mathbb{R}$ be arbitrary. Then,

$$\frac{\sqrt{2e(\beta-1)}}{2\beta\sqrt{\pi}} e^{-\frac{\beta a^2}{2}} \leq \int_a^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

Proof of Lemma 1. Since $e_1^\top w^0 \overset{d}{=} e_1^\top \frac{g}{\|g\|}$ with $g \sim \mathcal{N}(0, I_s)$, we have

$$\begin{aligned}
\mathbb{P}\left[e_1^\top w^0 \geq s^{-\frac{1}{2}}\right] &\geq \mathbb{P}_g\left[e_1^\top g \geq 1 \wedge \|g\| \leq s^{\frac{1}{2}}\right] \\
&\geq \mathbb{P}_g\left[e_1^\top g \geq 1\right] - \mathbb{P}_g\left[\|g\| \geq s^{\frac{1}{2}}\right] \\
&\gtrsim \frac{\sqrt{2e(\beta-1)}}{2\beta\sqrt{\pi}} e^{-\frac{\beta}{2}} - e^{-\Omega(s)} = \Theta(1).
\end{aligned}$$

□

Lemma 3. For $n = 1, 2, \ldots N^w$, let $a_n \sim \mathrm{Unif}\{\pm 1\}$ and $\xi_{n,i} \sim \mathrm{Unif}\{\pm 1\}$, and let $1 \leq p \leq q$. Then, with probability $\Omega(1)$, $a_n \sigma_n = \sum_{i=1}^q a_n \xi_{n,i} \mathrm{He}_i$ has its i-th Hermite coefficient ($p \leq i \leq q$) taking the desired sign for every i. Therefore, by taking these a_n, σ_n, and $b_n = 0$ as the initial values of a_n^w, σ_n^w, and b_n^w respectively, condition (1) of Assumption 4 is satisfied.

Proof. By construction, the probability that $a_n \xi_{n,i}$ has the desired sign is 2^{-q}. □

D.2 Gradient decomposition

Lemma 4. The gradient term decomposes as follows:

$$\nabla_w y^t a\sigma(w^{t\top}x^t+b)=\sum_{i=p}^{q}\left[i\alpha_i\beta_i(\theta^\top w^t)^{i-1}\theta+\sqrt{(i+2)(i+1)}\alpha_i\beta_{i+2}(\theta^\top w^t)^i w^t\right]+Z^t.$$

Here, the random variable Z^t has mean 0 and satisfies $\|Z^t\|=\tilde{O}(d^{\frac{1}{2}})$ with high probability, and for every $v\in\mathbb{R}^d$, $|v^\top Z^t|=\tilde{O}(1)$ with high probability. Furthermore, $\|\nabla_w y^t a\sigma(w^{t\top}x^t+b)\|=\tilde{O}(d^{\frac{1}{2}})$ with high probability, and for every $v\in\mathbb{R}^d$ with $\|v\|=O(1)$, $(\nabla_w y^t a\sigma(w^{t\top}x^t+b))^\top v=\tilde{O}(1)$ holds with high probability.

Proof.

$$\begin{aligned}
&\nabla_w\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[y^t a\sigma(w^{t\top}x+b)\right]\\
&=\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\sigma_\kappa^*(\theta^\top x)a\sigma'(w^{t\top}x+b)x\right]\\
&=\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\left(\sum_{i=p}^{q}\frac{\alpha_i}{\sqrt{i!}}\mathrm{He}_i(\theta^\top x)\right)\left(\sum_{i=1}^{q}i\frac{\beta_i}{\sqrt{i!}}\mathrm{He}_{i-1}(w^{t\top}x)\right)x\right]\\
&=\sum_{i=p}^{q}\sum_{j=1}^{q}\frac{j}{\sqrt{i!j!}}\alpha_i\beta_j\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[i\theta\mathrm{He}_{i-1}(\theta^\top x)\mathrm{He}_{j-1}(w^{t\top}x)+(j-1)w^t\mathrm{He}_i(\theta^\top x)\mathrm{He}_{j-2}(w^{t\top}x)\right]\\
&=\sum_{i=p}^{q}\left[i\alpha_i\beta_i(\theta^\top w^t)^{i-1}\theta+\sqrt{(i+2)(i+1)}\alpha_i\beta_{i+2}(\theta^\top w^t)^i w^t\right].
\end{aligned}$$

Define $Z^t=\nabla_w y^t a\sigma(w^{t\top}x^t+b)-\nabla_w\mathbb{E}[y^t a\sigma(w^{t\top}x+b)]$. Then $\mathbb{E}[Z^t]=0$, and

$$\|Z^t\|\le|y^t||a||\sigma'(w^{t\top}x^t+b)|\|x^t\|+\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[|y(x)||a||\sigma'(w^{t\top}x+b)|\|x\|\right]=\tilde{O}(d^{\frac{1}{2}})$$

with high probability. Moreover, for every $v\in\mathbb{R}^d$, $|v^\top Z^t|=\tilde{O}(1)$ with high probability. Similarly, $\|\nabla_w y^t a\sigma(w^{t\top}x^t+b)\|=\tilde{O}(d^{\frac{1}{2}})$ with high probability, and for every $v\in\mathbb{R}^d$ with $\|v\|=O(1)$, $(\nabla_w y^t a\sigma(w^{t\top}x^t+b))^\top v=\tilde{O}(1)$ with high probability. □

Lemma 5. Let $\eta=\eta^t\le c_\eta d^{-1}$ and suppose $\theta^\top w^t\ge\frac{1}{2}s^{-\frac{1}{2}}$. Then,

$$\begin{aligned}
&\theta^\top w^t+\eta\sum_{i=p}^{q}\left[i\alpha_i\beta_i(\theta^\top w^t)^{i-1}(1-(\theta^\top w^t)^2)\right]-\eta^2C_1^2(\theta^\top w^t)d+\eta\theta^\top P_{w^t}^\perp Z^t\\
&\le\theta^\top w^{t+1}\\
&\le\theta^\top w^t+\eta\sum_{i=p}^{q}\left[i\alpha_i\beta_i(\theta^\top w^t)^{i-1}(1-(\theta^\top w^t)^2)\right]+\eta\theta^\top P_{w^t}^\perp Z^t.
\end{aligned}$$

Here, Z^t is a mean-zero random variable satisfying $\|Z^t\|=\tilde{O}(d^{\frac{1}{2}})$ with high probability, and for every $v\in\mathbb{R}^d$ with $\|v\|=O(1)$, $|v^\top Z^t|=\tilde{O}(1)$ with high probability. Furthermore, $|\theta^\top w^t-\theta^\top w^{t+1}|=\tilde{O}(\eta)$ with high probability.

Proof. Let $g=\nabla_w y^t a^t\sigma(w^{t\top}x^t+b)$. Then,

$$\begin{aligned}
\|w^t-\eta P_{w^t}^\perp(-g)\|^{-1}&=(\|w^t\|^2+\eta^2\|P_{w^t}^\perp g\|^2)^{-\frac{1}{2}}\\
&\ge1-\frac{1}{2}\eta^2\|P_{w^t}^\perp g\|^2\ge1-\frac{1}{2}\eta^2\|g\|^2.
\end{aligned}$$

Therefore, with high probability,

$$\theta^\top w^{t+1}=\theta^\top\frac{w^t-\eta P_{w^t}^\perp(-g)}{\|w^t-\eta P_{w^t}^\perp(-g)\|}$$

$$
\begin{aligned}
&\geq (\theta^\top w^t + \eta\theta^\top P^\perp_{w^t} g)(1 - \frac{1}{2}\eta^2\|g\|^2) \\
&= \theta^\top w^t + \eta\theta^\top P^\perp_{w^t} g - \frac{1}{2}(\theta^\top w^t)\eta^2\|g\|^2 - \frac{1}{2}\eta^3|\theta P^\perp_{w^t} g|\|g\|^2 \\
&\geq \theta^\top w^t + \eta\theta^\top P^\perp_{w^t}\left(\sum_{i=p}^{q}\left[i\alpha_i\beta_i(\theta^\top w^t)^{i-1}\theta + \sqrt{(i+2)(i+1)}\alpha_i\beta_{i+2}(\theta^\top w^t)^i w^t\right] + Z^t\right) \\
&\quad - \frac{1}{2}(\theta^\top w^t)\eta^2\|g\|^2 - \frac{1}{2}\eta^3\|g\|^3 \\
&\geq \theta^\top w^t + \eta\sum_{i=p}^{q}\left[i\alpha_i\beta_i(\theta^\top w^t)^{i-1}\right]\theta^\top P^\perp_{w^t}\theta + \eta\theta^\top P^\perp_{w^t} Z^t - \frac{1}{2}(\theta^\top w^t)\eta^2 C_1^2 d - \frac{1}{2}\eta^3 C_1^3 d^{\frac{3}{2}} \\
&\geq \theta^\top w^t + \eta\sum_{i=p}^{q}\left[i\alpha_i\beta_i(\theta^\top w^t)^{i-1}\right](1 - (\theta^\top w^t)^2) - (\theta^\top w^t)\eta^2 C_1^2 d + \eta\theta^\top P^\perp_{w^t} Z^t
\end{aligned}
$$

Here, the second and third inequalities use Lemma 4, and the fourth uses $\eta \leq c_\eta s^{-1}$ and $\theta^\top w^t \geq \frac{1}{2}s^{-\frac{1}{2}}$. On the other hand,

$$
\begin{aligned}
\theta^\top w^{t+1} &\leq \theta^\top(w^t - \eta P^\perp_{w^t}(-g)) \\
&= \theta^\top w^t + \eta\theta^\top P^\perp_{w^t} g = \theta^\top w^t + \eta\sum_{i=p}^{q}\left[i\alpha_i\beta_i(\theta^\top w^t)^{i-1}\right](1 - (\theta^\top w^t)^2) + \eta\theta^\top P^\perp_{w^t} Z^t.
\end{aligned}
$$

Combining these gives the desired inequality. The properties of Z^t follow from Lemma 4. We now show $|\theta^\top w^{t+1} - \theta^\top w^t| = \tilde{O}(\eta)$. From the discussion above,

$$
|\theta^\top w^{t+1} - \theta^\top w^t| \leq \eta|\theta^\top P^\perp_{w^t} g| + \frac{1}{2}|\theta^\top w^t|\eta^2\|g\|^2 + \frac{1}{2}\eta^3|\theta P^\perp_{w^t} g|\|g\|^2
$$

and since $\eta \leq c_\eta d^{-1}$, the right-hand side is $\tilde{O}(\eta)$. □

D.3 Weak alignment

Lemma 6. Suppose $\theta^\top w^0 \geq s^{-\frac{1}{2}}$ and $\theta^\top w^t \leq c_1$ for all $t \leq \tau$. Then, setting $\eta^t = \eta \leq c_\eta d^{-\frac{p}{2}}$, for every $t \leq \tau$ we have

$$
\theta^\top w^{t+1} \geq (1 - c_1)(\theta^\top w^0) + \eta(1 - c_1)\sum_{t'=0}^{t} p\alpha_p\beta_p(\theta^\top w^{t'})^{p-1} \tag{2}
$$

with high probability. Furthermore, defining $(P^t)_{t=0}^{\tau+1}$ by

$$
P^0 = (1 - c_1)(\theta^\top w^0),
$$

$$
P^{t+1} = P^t + \eta(1 - c_1)p\alpha_p\beta_p(P^t)^{p-1} \quad (0 < t \leq \tau),
$$

we have $\theta^\top w^{t+1} \geq P^{t+1}$ with high probability for all $t \leq \tau$.

Proof. If $\theta^\top w^t \geq \frac{1}{2}s^{-\frac{1}{2}}$, then by Lemma 5

$$
\begin{aligned}
\theta^\top w^{t+1} &\geq \theta^\top w^t + \eta\sum_{i=p}^{q}\left[i\alpha_i\beta_i(\theta^\top w^t)^{i-1}(1 - (\theta^\top w^t)^2)\right] - (\theta^\top w^t)\eta^2 C_1^2 d + \eta\theta^\top P^\perp_{w^t} Z^t \\
&\geq \theta^\top w^t + \eta p\alpha_p\beta_p(\theta^\top w^t)^{p-1}(1 - (\theta^\top w^t)^2) - (\theta^\top w^t)\eta^2 C_1^2 d + \eta\theta^\top P^\perp_{w^t} Z^t \\
&\geq \theta^\top w^t + \eta p\alpha_p\beta_p(\theta^\top w^t)^{p-1}\left(1 - \frac{1}{3}c_1\right) - \eta(\theta^\top w^t)c_\eta d^{-\frac{p-2}{2}} C_1^2 + \eta\theta^\top P^\perp_{w^t} Z^t.
\end{aligned}
$$

In the third inequality we used $(\theta^\top w^t)^2 \le c_1^2 \le \frac{1}{3}c_1$. Also, since $\theta^\top w^t \ge \frac{1}{2}s^{-\frac{1}{2}} \ge \frac{1}{2}d^{-\frac{1}{2}}$,

$$\eta(\theta^\top w^t)c_\eta d^{-\frac{p-2}{2}}C_1^2 \le \frac{1}{3}c_1\eta p\alpha_p\beta_p(\theta^\top w^t)^{p-1}$$

and hence

$$\theta^\top w^{t+1} \ge \theta^\top w^t + \eta p\alpha_p\beta_p(\theta^\top w^t)^{p-1}\left(1-\frac{2}{3}c_1\right) + \eta\theta^\top P_{w^t}^\perp Z^t. \tag{3}$$

We prove the claim by induction. Take $t' \le \tau$. Assume that $\theta^\top w^t \ge (1-c_1)(\theta^\top w^0)$ and $\theta^\top w^t \ge \frac{1}{2}s^{-\frac{1}{2}}$ hold for $t = 0, 1, \dots, t'$. Then, by (3),

$$\begin{aligned}\theta^\top w^{t'+1} &\ge \theta^\top w^{t'} + \left(1-\frac{2}{3}c_1\right)\eta p\alpha_p\beta_p(\theta^\top w^{t'})^{p-1} + \eta\theta^\top P_{w^{t'}}^\perp Z^{t'} \\ &\ge \theta^\top w^0 + \sum_{t=0}^{t'}\left(1-\frac{2}{3}c_1\right)\eta p\alpha_p\beta_p(\theta^\top w^t)^{p-1} + \sum_{t=0}^{t'}\eta\theta^\top P_{w^t}^\perp Z^t. \end{aligned}\tag{4}$$

If $t' \le C_2(\theta^\top w^0)^{2-2p}$, then

$$\begin{aligned}-\sum_{t=0}^{t'}\eta\theta^\top P_{w^t}^\perp Z^t &\le \eta C_1\sqrt{t'} \le c_\eta s^{-\frac{p}{2}}C_1\sqrt{C_2}(\theta^\top w^0)^{1-p} \\ &\le 2^p c_\eta C_1\sqrt{C_2}(\theta^\top w^0) \le c_1(\theta^\top w^0)\end{aligned}$$

with high probability. On the other hand, when $t' > C_2(\theta^\top w^0)^{2-2p}$,

$$\begin{aligned}-\sum_{t=0}^{t'}\eta\theta^\top P_{w^t}^\perp Z^t &\le \eta C_1\sqrt{t'} < \eta C_1 t' C_2^{-\frac{1}{2}}(\theta^\top w^0)^{p-1} \\ &\le \frac{1}{3}c_1\eta t' p\alpha_p\beta_p((1-c_1)(\theta^\top w^0))^{p-1} \le \frac{1}{3}\sum_{t=0}^{t'}c_1\eta p\alpha_p\beta_p(\theta^\top w^t)^{p-1}\end{aligned}$$

with high probability. Substituting these into (4),

$$\theta^\top w^{t'+1} \ge (1-c_1)(\theta^\top w^0) + \sum_{t=0}^{t'}(1-c_1)\eta p\alpha_p\beta_p(\theta^\top w^t)^{p-1}$$

with high probability. Hence (2) follows for $t = t'+1$, and $\theta^\top w^{t'+1} \ge (1-c_1)(\theta^\top w^0)$ and $\theta^\top w^{t'+1} \ge \frac{1}{2}s^{-\frac{1}{2}}$ are satisfied. Moreover, for $(P^{t'})_{t'=0}^{\tau}$ we have

$$P^{t'+1} = P^0 + \sum_{t=0}^{t'}\eta(1-c_1)p\alpha_p\beta_p(P^t)^{p-1},$$

so comparing with (2) yields $\theta^\top w^t \ge P^t$ for all $t \le \tau+1$. □

Lemma 7. Let $\eta^t = \eta \le c_\eta d^{-\frac{p}{2}}$. Then there exists $t_1 \le T_{1,1} = \Theta(\eta^{-1}s^{\frac{p-2}{2}})$ such that $\theta^\top w^{t_1} > c_1$ with high probability.

Proof. Suppose $\theta^\top w^t \le c_1$ for all $t \le T_{1,1}$. Setting $c = \eta(1-c_1)p\alpha_p\beta_p$, define

$$T_{1,1} = \left\lfloor c^{-1}(1+c)^{p-1}(p-2)^{-1}(P^0)^{-(p-2)}\right\rfloor.$$

Then, by Lemma 6 and the Bihari–LaSalle inequality (Lemma 38),

$$\theta^\top w^t \ge P^t \ge \frac{P^0}{\left(1-c(1+c)^{-(p-1)}(p-2)(P^0)^{p-2}t\right)^{\frac{1}{p-2}}}$$

and at $t = T_{1,1}$, using the fact that $1 \le a(\lfloor (a)^{-1}\rfloor + 1)$ for every $a > 0$,

$$\theta^\top w^{T_{1,1}} \ge \frac{P^0}{\left(c(1+c)^{-(p-1)}(p-2)(P^0)^{p-2}\right)^{\frac{1}{p-2}}} = \frac{1}{\left(\eta(1-c_1)(1+c)^{-(p-1)}p(p-2)\alpha_p\beta_p\right)^{\frac{1}{p-2}}} > 1.$$

This contradicts $\theta^\top w^{T_{1,1}} \le 1$. Therefore, there exists $t_1 \le T_{1,1}$ such that $\theta^\top w^t \le c_1$ for all $t \le t_1 - 1$ and $\theta^\top w^{t_1} > c_1$. □

D.4 Amplification of alignment

For the remainder of this section, we shift the time index so that $t_1 = 0$, i.e., weak alignment is achieved at $t = 0$.

Lemma 8. Suppose $\theta^\top w^0 \geq c_1$, $\eta = \eta_t \leq c_\eta d^{-\frac{p}{2}}$, and $\theta^\top w^t \leq 1 - c_1$ for all $t \leq \tau$. Moreover, define $(P^t)_{t=0}^{\tau+1}$ by $P^0 = (1 - c_1)(\theta^\top w^0)$ and $P^{t+1} = P^t + c_1 \eta p \alpha_p \beta_p (P^t)^{p-1}$. Then, for every $t \leq \tau + 1$, $\theta^\top w^t \geq P^t$ with high probability.

Proof. By an argument analogous to the proof of Lemma 6, if $\theta^\top w^t \geq \frac{1}{2} s^{-\frac{1}{2}}$ then

$$\begin{aligned}\theta^\top w^{t+1} &\geq \theta^\top w^t + \eta p \alpha_p \beta_p (\theta^\top w^t)^{p-1}(1 - (\theta^\top w^t)^2) - \eta^2 C_1^2 (\theta^\top w^t) d + \eta \theta^\top P_{w^t}^\perp Z^t \\ &\geq \theta^\top w^t + \eta p \alpha_p \beta_p (\theta^\top w^t)^{p-1}(1 - (1 - c_1)^2) - \eta c_\eta C_1^2 (\theta^\top w^t) d^{-\frac{p-2}{2}} + \eta \theta^\top P_{w^t}^\perp Z^t.\end{aligned}$$

As in Lemma 6, take $t' \leq \tau$ and assume that $\theta^\top w^t \geq (1 - c_1)(\theta^\top w^0)$ and $\theta^\top w^t \geq \frac{1}{2} s^{-\frac{1}{2}}$ for $t = 0, 1, \ldots, t'$. Since $\theta^\top w^t \geq \frac{1}{2} s^{-\frac{1}{2}} \geq d^{-\frac{1}{2}}$, we have $\eta c_\eta C_1^2 (\theta^\top w^t) d^{-\frac{p-2}{2}} \leq \frac{1}{3} c_1 \eta p \alpha_p \beta_p (\theta^\top w^t)^{p-1}$. Furthermore, since $c_1^2 \leq \frac{1}{3} c_1$, $(1 - (1 - c_1)^2) = 2c_1 - c_1^2 \geq \frac{5}{3} c_1$. Thus,

$$\begin{aligned}\theta^\top w^{t'+1} &\geq \theta^\top w^{t'} + \frac{4}{3} c_1 \eta p \alpha_p \beta_p (\theta^\top w^{t'})^{p-1} + \eta \theta^\top P_{w^{t'}}^\perp Z^{t'} \\ &\geq \theta^\top w^0 + \frac{4}{3} \sum_{t=0}^{t'} c_1 \eta p \alpha_p \beta_p (\theta^\top w^t)^{p-1} + \sum_{t=0}^{t'} \eta \theta^\top P_{w^t}^\perp Z^t.\end{aligned}$$

Furthermore,

$$-\sum_{t=0}^{t'} \eta \theta^\top P_{w^t}^\perp Z^t \leq c_1 (\theta^\top w^0) + \frac{1}{3} \sum_{t=0}^{t'} c_1 \eta p \alpha_p \beta_p (\theta^\top w^t)^{p-1}$$

with high probability. Therefore,

$$\theta^\top w^{t'+1} \geq (1 - c_1)(\theta^\top w^0) + \sum_{t=0}^{t'} c_1 \eta p \alpha_p \beta_p (\theta^\top w^t)^{p-1} \tag{5}$$

with high probability. Hence $\theta^\top w^{t'+1} \geq (1 - c_1)(\theta^\top w^0)$ and $\theta^\top w^{t'+1} \geq \frac{1}{2} s^{-\frac{1}{2}}$. Therefore, for every $t' \leq \tau$, $\theta^\top w^{t'} \geq (1 - c_1)(\theta^\top w^0)$ and $\theta^\top w^{t'} \geq \frac{1}{2} s^{-\frac{1}{2}}$ hold, and (5) is satisfied by induction. Comparing with the update equation for $P^{t'}$, we obtain $\theta^\top w^{t'} \geq P^{t'}$ with high probability for every $t' \leq \tau + 1$. □

Lemma 9. Let $\eta = \eta_t \leq c_\eta d^{-\frac{p}{2}}$. For neurons satisfying Lemma 5, there exists $t_2 \leq T_{1,2} = \tilde{\Theta}(\eta^{-1})$ such that $\theta^\top w^{t_2} > 1 - c_1$ with high probability.

Proof. Suppose $\theta^\top w^t \leq 1 - c_1$ for all $t \leq T_{1,2}$. Setting $c = \eta c_1 p \alpha_p \beta_p$, define

$$T_{1,2} = \lfloor c^{-1}(1 + c)^{p-1}(p - 2)^{-1}(P^0)^{-(p-2)} \rfloor.$$

Then, by Lemma 8 and the Bihari–LaSalle inequality Lemma 38, at $t = T_{1,2}$

$$\begin{aligned}\theta^\top w^{T_{1,2}} \geq P^{T_{1,2}} &\geq \frac{P^0}{\left(1 - c(1 + c)^{-(p-1)}(p - 2) T_{1,2}\right)^{\frac{1}{p-2}}} \\ &\geq \frac{P^0}{\left(c(1 + c)^{-(p-1)}(p - 2)(P^0)^{p-2}\right)^{\frac{1}{p-2}}} = \frac{1}{\left(\eta c_1 (1 + c)^{-(p-1)} p (p - 2) \alpha_p \beta_p\right)^{\frac{1}{p-2}}} > 1.\end{aligned}$$

This contradicts $\theta^\top w^{T_{1,2}} \leq 1$. Therefore, there exists $t_2 \leq T_{1,2}$ such that $\theta^\top w^t \leq 1 - c_1$ for all $t \leq t_2 - 1$ and $\theta^\top w^{t_2} > 1 - c_1$. □

D.5 Strong alignment and localization

For the remainder of this section, we shift the time index so that $t_1 + t_2 = 0$, i.e., amplification of alignment is achieved at $t = 0$.

Lemma 10. Let $0 < \bar{\varepsilon} < c_1$. If $\eta^t = \eta \le c_\eta \bar{\varepsilon} d^{-1} \wedge c_\eta \bar{\varepsilon}^2$, and if $\theta^\top w^0 \ge 1 - 2c_1$ and $\theta^\top w^t \le 1 - \bar{\varepsilon}$ for all $t \le \tau$, then for every $t \le \tau + 1$,

$$\theta^\top w^t \ge \theta^\top w^0 - c_1 \bar{\varepsilon} + t \bar{\varepsilon} \eta p \alpha_p \beta_p$$

with high probability.

Proof. If $\theta^\top w^t \ge \frac{1}{2} s^{-\frac{1}{2}}$, then by Lemma 5,

$$\begin{aligned}
\theta^\top w^{t+1} &\ge \theta^\top w^t + \eta \sum_{i=p}^{q} \left[i \alpha_i \beta_i (\theta^\top w^t)^{i-1} (1 - (\theta^\top w^t)^2) \right] - (\theta^\top w^t) \eta^2 C_1^2 d + \eta \theta^\top P_{w^t}^\perp Z^t \\
&\ge \theta^\top w^t + \eta p \alpha_p \beta_p (\theta^\top w^t)^{p-1} (1 - (\theta^\top w^t)^2) - (\theta^\top w^t) \eta^2 C_1^2 d + \eta \theta^\top P_{w^t}^\perp Z^t.
\end{aligned}$$

If $\theta^\top w^t \ge 1 - 3c_1$, then

$$(\theta^\top w^t)^{p-1} (1 - (\theta^\top w^t)^2) \ge (1 - 3c_1)^{p-1} (1 + (1 - 3c_1))(1 - \theta^\top w^t) \ge \frac{5}{3} (1 - \theta^\top w^t). \quad (6)$$

Moreover, if $\theta^\top w^t \le 1 - \bar{\varepsilon}$ and $\eta \le c_\eta \bar{\varepsilon} d^{-1}$, then

$$(\theta^\top w^t) \eta^2 C_1^2 d \le \eta c_\eta C_1^2 (1 - \theta^\top w^t) \le \frac{1}{3} \eta p \alpha_p \beta_p (1 - \theta^\top w^t). \quad (7)$$

Take an arbitrary $t' \le \tau$. If $1 - 3c_1 \le (\theta^\top w^t) \le 1 - \bar{\varepsilon}$ for all $t \le t'$, then by (6) and (7),

$$\begin{aligned}
\theta^\top w^{t'+1} &\ge \theta^\top w^{t'} + \frac{4}{3} \eta p \alpha_p \beta_p (1 - \theta^\top w^{t'}) + \eta \theta^\top P_{w^{t'}}^\perp Z^{t'} \\
&\ge \theta^\top w^0 + \frac{4}{3} \sum_{t=0}^{t'} \eta p \alpha_p \beta_p (1 - \theta^\top w^t) + \sum_{t=0}^{t'} \eta \theta^\top P_{w^t}^\perp Z^t. \quad (8)
\end{aligned}$$

Furthermore, since $\eta \le c_\eta \bar{\varepsilon}^2$,

$$\begin{aligned}
-\sum_{t=0}^{t'} \eta \theta^\top P_{w^t}^\perp Z^t &\le \eta C_1 \sqrt{t'} \\
&\le \begin{cases} \eta C_1 \sqrt{C_2} \bar{\varepsilon}^{-1} \le c_1 \bar{\varepsilon} & (t' \le C_2 \bar{\varepsilon}^{-2}) \\ t' \eta C_1 t'^{-\frac{1}{2}} < t' \eta C_2^{-\frac{1}{2}} \bar{\varepsilon} \le \frac{1}{3} t' \eta p \alpha_p \beta_p (1 - \theta^\top w^t) & (t' > C_2 \bar{\varepsilon}^{-2}) \end{cases}
\end{aligned}$$

with high probability. Substituting the above into (8),

$$\begin{aligned}
\theta^\top w^{t'+1} &\ge \theta^\top w^0 - c_1 \bar{\varepsilon} + \sum_{t=0}^{t'} \eta p \alpha_p \beta_p (1 - \theta^\top w^t) \\
&\ge \theta^\top w^0 - c_1 \bar{\varepsilon} + t' \eta p \alpha_p \beta_p \bar{\varepsilon} \quad (9)
\end{aligned}$$

with high probability. Hence $1 - 3c_1 \le \theta^\top w^t$ follows at $t = t' + 1$. By induction, $1 - 3c_1 \le \theta^\top w^t$ holds for every $t \le \tau + 1$, and (9) follows for every $t' \le \tau$. □

Lemma 11. Let $0 < \tilde{\varepsilon} < c_1$. Let

$$\eta^t = \begin{cases} \eta_1 \le c_\eta d^{-\frac{p}{2}} & (0 \le t \le \Delta T_1 + \Delta T_2 - 1) \\ \eta_2 \le c_\eta \tilde{\varepsilon} d^{-1} \wedge c_\eta \tilde{\varepsilon}^2. & (\Delta T_1 + \Delta T_2 \le t \le \Delta T_1 + \Delta T_2 + T_{1,3} - 1) \end{cases}$$

Here, $\Delta T_1 = T_{1,1} - t_1$, $\Delta T_2 = T_{1,2} - t_2$, and $T_{1,3} = \tilde{\Theta}(\tilde{\varepsilon}^{-1} \eta_2^{-1})$. Then, for every neuron satisfying $\theta^\top w^0 \ge 1 - c_1$, $\theta^\top w^{\Delta T_1 + \Delta T_2 + T_{1,3}} > 1 - 3\tilde{\varepsilon}$ with high probability.

Proof. First, we show that $\theta^\top w^t \geq 1 - 2c_1$ for every $0 \leq t \leq \Delta T_1 + \Delta T_2$. Since $\eta_1 \leq c_\eta d^{-\frac{p}{2}}$, setting $\bar{\varepsilon} = d^{-\frac{1}{2}}$, the assumption $\eta \leq c_\eta \bar{\varepsilon} d^{-1} \wedge c_\eta \bar{\varepsilon}^2$ of Lemma 10 is satisfied. By Lemma 10, until the first $\tau < \Delta T_1 + \Delta T_2$ at which $\theta^\top w^\tau > 1 - \bar{\varepsilon}$, we have $\theta^\top w^t \geq \theta^\top w^0 - \bar{\varepsilon} \geq 1 - 2c_1$ for every $t < \tau$. If there exists $t > \tau$ with $\theta^\top w^t < 1 - \bar{\varepsilon}$, let τ' be the smallest such t. By Lemma 5, $|\theta^\top w^{\tau'} - \theta^\top w^{\tau'-1}| \leq C_1\eta_1 \leq \bar{\varepsilon}$, and $\theta^\top w^{\tau'} \geq \theta^\top w^{\tau'-1} - \bar{\varepsilon} \geq 1 - 2\bar{\varepsilon} \geq 1 - c_1$. Then, by Lemma 10, until $\theta^\top w^t$ exceeds $1 - \bar{\varepsilon}$, we have $\theta^\top w^t \geq \theta^\top w^{\tau'} - c_1 \geq 1 - 2c_1$. Applying this argument repeatedly yields the desired result. In particular, $\theta^\top w^{\Delta T_1 + \Delta T_2} \geq 1 - 2c_1$.

Next, let $\Delta T_1 + \Delta T_2 \leq t \leq \Delta T_1 + \Delta T_2 + T_{1,3}$. Setting $\bar{\varepsilon} = \tilde{\varepsilon}$, the assumption $\eta_2 \leq c_\eta \bar{\varepsilon} d^{-1} \wedge c_\eta \bar{\varepsilon}^2$ of Lemma 10 is satisfied. Define

$$T_{1,3} = \left\lceil 3c_1(\eta_2 \tilde{\varepsilon} p \alpha_p \beta_p)^{-1} \right\rceil.$$

If $\theta^\top w^t \leq 1 - \tilde{\varepsilon}$ for every $t = \Delta T_1 + \Delta T_2, \Delta T_1 + \Delta T_2 + 1, \ldots, \Delta T_1 + \Delta T_2 + T_{1,3}$, then by Lemma 10

$$\begin{aligned}\theta^\top w^t &\geq 1 - 2c_1 - c_1\tilde{\varepsilon} + (t - \Delta T_1 - \Delta T_2)\tilde{\varepsilon}\eta p \alpha_p \beta_p \\ &\geq 1 - 3c_1 + (t - \Delta T_1 - \Delta T_2)\tilde{\varepsilon}\eta p \alpha_p \beta_p\end{aligned}$$

with high probability, but at $t = \Delta T_1 + \Delta T_2 + T_{1,3}$ we have $(\mathrm{RHS}) \geq 1$, a contradiction. Therefore, with high probability, there exists $t_3 \leq \Delta T_1 + \Delta T_2 + T_{1,3}$ such that $\theta^\top w^{t_3} > 1 - \tilde{\varepsilon}$.

Finally, we show $\theta^\top w^{\Delta T_1 + \Delta T_2 + T_{1,3}} \geq 1 - 3\tilde{\varepsilon}$. If there exists $t > t_3$ with $\theta^\top w^t < 1 - \tilde{\varepsilon}$, let τ be the smallest such t. By Lemma 5, $|\theta^\top w^\tau - \theta^\top w^{\tau-1}| \leq C_1\eta_2$, and by the same argument as before, $\theta^\top w^\tau \geq 1 - 2\tilde{\varepsilon}$. Then, by Lemma 10, until $\theta^\top w^t > 1 - \tilde{\varepsilon}$, we have $\theta^\top w^t \leq 1 - 3\tilde{\varepsilon}$. Repeating this argument, we obtain $\theta^\top w^t \geq 1 - 3\tilde{\varepsilon}$ with high probability for every $t_3 \leq t \leq \Delta T_1 + \Delta T_2 + T_{1,3}$. □

D.6 Neuron filtering

Lemma 12. Let $V \subset \mathbb{R}^d$ be an $s = d^\alpha$-dimensional subspace containing $\theta \in V$, and let $W = V^\perp$ be its orthogonal complement. Let x be a random variable on S^{d-1} whose probability density function is symmetric under any rotation within V about the axis θ, and invariant under any rotation within W. Let $\Sigma = \mathbb{E}[xx^\top]$, and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ be its eigenvalues. If, for $\bar{\varepsilon} = o(1)$, $p = \mathbb{P}[|x^\top\theta| \geq 1 - \bar{\varepsilon}] = \Theta(1)$, then

$$\lambda_1 = \Theta(1), \quad \lambda_2 = \lambda_3 = \cdots = \lambda_s = O(d^{-\alpha}), \quad \lambda_{s+1} = \lambda_{s+2} = \cdots = \lambda_d = O(d^{-1}),$$

and $\sum_{i=1}^d \lambda_i = 1$.

Proof. W.l.o.g., let $\theta = e_1$ and $V = \mathrm{span}(e_1, e_2, \ldots, e_s)$. By the rotational symmetry of x about the axis θ, for $i, j \leq s$ with $i \neq j$, $\Sigma_{ij} = 0$. By the symmetry of x under any rotation within W, for $i > s$ and $j \neq i$, $\Sigma_{ij} = 0$. Therefore, there exist $\lambda_1', \lambda_2', \ldots, \lambda_d' \geq 0$ such that $\Sigma = \mathrm{diag}(\lambda_1', \lambda_2', \ldots, \lambda_d')$. Here, $\lambda_1' = \mathbb{E}[(x^\top\theta)^2] \geq p(1 - \bar{\varepsilon})^2 = \Theta(1)$. Furthermore, since $\sum_{i=1}^d \lambda_i' = \mathrm{Tr}(\mathbb{E}[xx^\top]) = 1$, we have $\lambda_2' = \lambda_3' = \ldots \lambda_s' = O(s^{-1})$ and $\lambda_{s+1}' = \lambda_{s+2}' = \cdots = \lambda_d' = O(d^{-1})$. Therefore, λ_1' is the largest eigenvalue, and $\lambda_1 = \lambda_1'$ follows. □

Lemma 13. Suppose $\bar{\varepsilon} = o(1)$, and let $x_1, x_2, \ldots, x_N$ be i.i.d. random variables on S^{d-1} satisfying the symmetry assumption of Lemma 12. Define $X_N = \frac{1}{N}\sum_{n=1}^N x_n x_n^\top$ and let $\hat{\theta}$ be its leading unit eigenvector. If we take $N \gtrsim \bar{\varepsilon}^{-2}\log d$, then $|\hat{\theta}^\top\theta| \geq 1 - \bar{\varepsilon}$ with high probability. Furthermore, if $y \in S^{d-1}$ satisfies $|\theta^\top y| \geq 1 - \bar{\varepsilon}$, then $|\hat{\theta}^\top y| \geq 1 - 2\bar{\varepsilon}$ with high probability.

Proof. Setting $X = \mathbb{E}[x_1 x_1^\top]$, by Lemma 39, for $t \geq 0$,

$$\mathbb{P}\left[\|X_N - X\| \geq t\right] \leq 2d\exp\left(-\frac{t^2N^2/2}{N\|X\| + 2tN/3}\right).$$

Here we used $\|\sum_{n=1}^N \mathbb{E}[(x_n x_n^\top)^2]\| = \|\sum_{n=1}^N \mathbb{E}[x_n x_n^\top]\| = N\|X\|$. In particular, substituting $t = \sqrt{\frac{C\log d}{N}} = o(1)$ gives $\mathbb{P}[\|X_N - X\| \geq t] \lesssim \exp(-C\log d)$. That is, $\|X_N - X\| = O(\sqrt{\log d/N})$ with high probability. By Lemma 12 and Lemma 40,

$$\sin\Theta(\hat{\theta}, \theta) = O\left(\sqrt{\log d/N}\right).$$

Furthermore, $1 - |\hat{\theta}^\top \theta| = O(\sqrt{\log d / N})$. Therefore, by taking $N \gtrsim \bar{\varepsilon}^{-2} \log d$, $|\hat{\theta}^\top \theta| \geq 1 - \bar{\varepsilon}$ with high probability. If $y \in S^{d-1}$ satisfies $|\theta^\top y| \geq 1 - \bar{\varepsilon}$, then

$$|\hat{\theta}^\top y| \geq |\hat{\theta}^\top \theta||\theta^\top y| \geq 1 - 2\bar{\varepsilon}.$$

□

D.7 Second-layer training

Let σ be a polynomial of degree q.

Lemma 14 (Oko et al. (2024a), Lemma 29)**.** Let $b_j \sim \mathrm{Unif}([-C_b, C_b])$ with $C_b = \tilde{O}(1)$, and let $h(s)$ be a polynomial of degree q. Then, for $v \in \mathbb{S}^{d-1}$, there exist $a_1, a_2, \ldots, a_N$ such that

$$\sup_{t=T_1+1,\ldots,T_1+T_2} \left| \frac{1}{N} \sum_{j=1}^{N} a_j \sigma(v^\top x^t + b_j) - h(v^\top x^t) \right| = \tilde{O}(N^{-1})$$

with high probability. Here, $\sum_{j=1}^{N} a_j^2 = \tilde{O}(N)$ and $\sum_{j=1}^{N} |a_j| = \tilde{O}(N)$.

Define $r_a(x) = \frac{1}{N^w} \sum_{m=1}^{N^w} a_n \sigma(\hat{w}_n^\top x + b_n)$.

Lemma 15. Let $\tilde{N}^w \gtrsim N^w \log d$. Then, there exists $a^* = (a_n^*)_{n=1}^{N^w} \in \mathbb{R}^{N^w}$ such that

$$\frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} \left(r_{a^*}(x^t) - \pi_\kappa \sigma_\kappa^*(\theta_\kappa^\top x^t) \right)^2 \leq C_1 \pi_\kappa^2 (|N^w|^{-2} + \bar{\varepsilon}^2).$$

Here, $\|a^*\|_2^2 = \tilde{O}(\pi_\kappa^2 (\tilde{N}^w)^2 |N^w|^{-1})$ and $\|a^*\|_1 = \tilde{O}(\pi_\kappa \tilde{N}^w)$.

Proof. Obtained from Oko et al. (2024a), Lemma 31 by setting $M = 1$ and multiplying the objective function by π_κ. □

Let $\hat{a}$ be the ridge-regularized empirical risk minimizer

$$\hat{a} = \operatorname*{argmin}_{a \in \mathbb{R}^{N^w}} \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} \left(y^t - \frac{1}{N^w} \sum_{n=1}^{N^w} a_n \sigma_n(\hat{w}_n^\top x^t + b_n) \right)^2 + \frac{\lambda}{2} \|a\|_2^2.$$

The regularization parameter $\lambda > 0$ can be chosen so that

$$\hat{\mathcal{L}}(\hat{a}) \leq \hat{\mathcal{L}}(a^*), \quad \|\hat{a}\|_2 \leq \|a^*\|_2$$

hold (Oko et al., 2024a).

Lemma 16. Let $\tilde{N}^w = \Theta(N_{\min}^w \log d)$ and let σ be a polynomial of degree q. Then, there exists $\lambda > 0$ such that the ridge estimator $\hat{a}$ satisfies, with probability $1 - o_d(1)$,

$$\mathbb{E}_x[|r_{\hat{a}}(x) - r^*(x)|] \lesssim \pi_\kappa(|N^w|^{-1} + \tilde{\varepsilon}) + \pi_\kappa \sqrt{\frac{\log d}{T_2}}.$$

In particular, setting $T_2 = \tilde{\Theta}(\varepsilon^{-2})$, $\tilde{\varepsilon} = \tilde{\Theta}(\varepsilon)$, and $N^w = \tilde{\Theta}(\varepsilon^{-1})$, we have $\mathbb{E}_x[|r_{\hat{a}}(x) - r^*(x)|] \lesssim \pi_\kappa \varepsilon$.

Proof. Obtained from Oko et al. (2024a), Lemma 14 by setting $M = 1$ and multiplying the objective function by π_κ. □

E Full proof of Theorem 1: W2S feature learning

Notations. Throughout this section, we consider multiple tasks $k = 1, 2, \ldots, K$. We write θ_k for the true feature vector corresponding to task k, and define $\Sigma_k^\perp = I - \Sigma_k$ for each k. Write the Hermite expansions of the transformed teacher signal and the activation functions of strong-model neurons as

$$\bar{r}_\kappa^* \exp(\bar{r}_\kappa^*) = \sum_{i=1}^{\infty} \frac{\bar{\alpha}_i}{\sqrt{i!}} \mathrm{He}_i, \qquad a_{k,n}\sigma_{k,n}^s(\cdot + b_{k,n}) = \sum_{i=1}^{q} \frac{\tilde{\beta}_{k,n,i}}{\sqrt{i!}} \mathrm{He}_i.$$

In the training of the strong model (Algorithm 1), the parameter updates of each neuron can be analyzed independently. When focusing on a specific neuron $n \in [N_k]$ belonging to task $k \in [K]$, we sometimes suppress indices and write

$$w_k^t = w_{k,n}^t, \quad a_k = a_{k,n}, \quad b_k = b_{k,n}, \quad \tilde{\beta}_{k,i} = \tilde{\beta}_{k,n,i}.$$

The subsequent arguments are carried out after rescaling the learning rate in the SGD update rule. Specifically, we set $\eta \leftarrow N_\kappa \pi_\kappa^{-1} \eta$, scaling by the task-$\kappa$ signal, so that the weight update for task k in Algorithm 1 becomes

$$w_k^{t+1} \leftarrow w_k^t + \eta g_k^t, \qquad g_k^t = \frac{N_\kappa}{N_k} \pi_\kappa^{-1} \nabla_{w_k} \bar{y}^t a_k \sigma_k(w_k^{t\top} x + b_k).$$

Define the scaled expansion coefficients $\beta_{k,i} = \pi_k^{-1} \tilde{\beta}_{k,i}$. Then $c_\beta < \beta_{k,i} < C_\beta$ for all $k \in [K]$ and $i \le q$.

Set $c_w = 2\bar{\alpha}_2 \beta_{\kappa,2} + \sqrt{12}\bar{\alpha}_2 \beta_{\kappa,4}$. By Assumption 5 (ii), $2\bar{\alpha}_2 \tilde{\beta}_{\kappa,2} + \sqrt{12}\bar{\alpha}_2 \tilde{\beta}_{\kappa,4} > 0$, so $c_w > 0$. By Lemma 18, $|\bar{\alpha}_2| = \tilde{\Theta}(1)$, and hence $c_w = \tilde{\Theta}(1)$.

We take constants of order $\mathrm{polylog}(d)$ satisfying the following ordering:

$$C_1 \lesssim c_1^{-1} \lesssim c_2^{-1} \lesssim C_2 \lesssim c_r^{-1} \lesssim c_3^{-1} \lesssim \begin{cases} \delta^{-1} \lesssim \begin{cases} c_\varepsilon^{-1} \\ c_\eta^{-1} \end{cases} \\ c_\chi^{-1} \end{cases} = \tilde{O}(1),$$

$$c_w^{-1} \lesssim C_1 \lesssim C_2 \lesssim C_3 \lesssim c_3^{-1},$$

$$\tilde{\delta}^{-1} \lesssim \delta^{-1}.$$

The precise ordering of these constants will be determined so as to keep the subsequent proofs consistent. These constants are defined independently of those in Sections D and F.

The formal versions of Theorem 1 and Proposition 2 rely on the following assumption.

Assumption 5 (Strong-model initialization and Hermite conditions)**.** The following conditions hold for the strong model training in Algorithm 1.

(i) **Generative exponent.** The link function σ_κ^* of the task-κ reward is an even polynomial, so that $\mathrm{GE}(\sigma_\kappa^*) = 2$.

(ii) **Initialization and Hermite structure.** The strong model is initialized with $N^s := \sum_{k=1}^K N_k$ neurons and parameters $\Theta_k = (a_{k,n}^0, b_{k,n}^0, w_{k,n}^0)_{n=1}^{N_k}$. For $k = \kappa$, $\|\Sigma_\kappa^\perp w_{\kappa,n}^0\|^2 \le c_r s^{-1/2}$ and $\Sigma_\kappa w_{\kappa,n}^0 / \|\Sigma_\kappa w_{\kappa,n}^0\| \sim \mathrm{Unif}(S^{d-1} \cap V_\kappa)$. Writing $\bar{r}_\kappa^* \exp(\bar{r}_\kappa^*) = \sum_{i \ge 0} \frac{\bar{\alpha}_i}{\sqrt{i!}} \mathrm{He}_i$ and $a_{k,n}^0 \sigma_{k,n}(\cdot + b_{k,n}^0) = \sum_{i=0}^q \frac{\tilde{\beta}_{k,n,i}}{\sqrt{i!}} \mathrm{He}_i$, the constants c_β, C_β satisfy $c_\beta \pi_k \le |\tilde{\beta}_{k,n,i}| \le C_\beta \pi_k$ for all $k, n, i \le q$. Moreover, $\Theta(N_\kappa)$ of the task-κ neurons satisfy

$$\bar{\alpha}_i \tilde{\beta}_{\kappa,n,i} > 0 \ \ (i = p), \qquad \bar{\alpha}_i \tilde{\beta}_{\kappa,n,i} \ge 0 \ \ (p < i \le q),$$

$$2\bar{\alpha}_2 \tilde{\beta}_{\kappa,n,2} + \sqrt{12}\, \bar{\alpha}_2 \tilde{\beta}_{\kappa,n,4} > 0, \qquad i\bar{\alpha}_i \tilde{\beta}_{\kappa,n,i} + \sqrt{(i+2)(i+1)}\, \bar{\alpha}_i \tilde{\beta}_{\kappa,n,i+2} \ge 0 \ \ (3 \le i \le q),$$

together with $1 - (2\bar{\alpha}_2 \tilde{\beta}_{\kappa,n,2})^{-1} \sqrt{12}\, \bar{\alpha}_2 \tilde{\beta}_{\kappa,n,4} = \tilde{\Omega}(1)$.

A concrete initialization satisfying Assumption 5 (ii) is provided in Section E.2 (Lemma 20).

Remark 2. The assumption $\mathrm{GE}(\sigma_\kappa^*) = 2$ is essential for feature preservation (Proposition 2). If $\mathrm{GE}(\sigma_\kappa^*) = 1$, the transformed teacher signal retains a nonzero $i = 1$ Hermite component $\bar{\alpha}_1$, contributing a constant gradient term $\bar{\alpha}_1 \beta_{k,1} \theta_\kappa$ to off-target neurons regardless of their alignment $\theta_\kappa^\top \tilde{w}_k^t$. Since this term does not decay with $\theta_\kappa^\top \tilde{w}_k^t$, it cannot be absorbed into the remainder R^t in Lemma 22 (ii), and off-target neurons accumulate alignment with θ_κ over iterations, causing forgetting of pre-trained features for $k \neq \kappa$.

Theorem 1 (W2S feature alignment; formal version of Theorem 1)**.** Assume Assumptions 1 to 3 and 5. Let $\mathcal{E}_s := \varepsilon_w \vee \tilde{\varepsilon}_w \vee (1 - \lambda_\kappa)$, and suppose $\mathcal{E}_s \leq c_\varepsilon s^{-1/2}$. Fix $\tilde{\varepsilon}$ with $\tilde{\varepsilon} = \tilde{\Omega}(\mathcal{E}_s)$ and set

$$T = \tilde{\Theta}\Big(s^{3/2} \ \vee \ s\,\tilde{\varepsilon}^{-1} \log \tilde{\varepsilon}^{-1} \ \vee \ \tilde{\varepsilon}^{-2} \log \tilde{\varepsilon}^{-1}\Big).$$

Then, for a suitable choice of the learning-rate schedule $(\eta^t)_{t \geq 0}$ and the temperature ρ, with probability $1 - 6\tilde{\delta}$, Algorithm 1 outputs weights such that at least $\Theta(N_\kappa)$ of the task-κ neurons satisfy

$$\theta_\kappa^\top \hat{w}_{\kappa,n} \geq 1 - \tilde{\varepsilon}.$$

Under the scaling $\tilde{\varepsilon} = \tilde{\Theta}(s^{-1/2})$, this reduces to $T = \tilde{O}(s^{3/2})$.

Proof. Fix a neuron $n \in [N_\kappa]$. By Assumption 5 (ii) and Lemma 19, $\theta_\kappa^\top w_{\kappa,n}^0 \geq s^{-1/2}$ with probability $\Theta(1)$, and $\|\Sigma_\kappa^\perp w_{\kappa,n}^0\|^2 \leq c_r s^{-1/2}$. Applying Lemma 26 with $\eta_1 \leq c_\eta s^{-3/2}$, there exists $t_1 \leq T_{1,1} = \tilde{\Theta}(\eta_1^{-1})$ such that $\theta_\kappa^\top w_{\kappa,n}^{t_1} > c_1$ and $\|\Sigma_\kappa^\perp w_{\kappa,n}^t\|^2 \leq 3c_r s^{-1/2}$ for all $t \leq t_1$, with probability $1 - \tilde{\delta} - \delta$. Since $3c_r s^{-1/2} \leq c_r c_1$, the conditions of Lemma 28 hold at t_1, and there exists $t_2 \leq T_{1,2} = \tilde{\Theta}(\eta_1^{-1})$ such that $\theta_\kappa^\top w_{\kappa,n}^{t_1+t_2} > 1 - c_1$ and $\|\Sigma_\kappa^\perp w_{\kappa,n}^{t_1+t_2}\|^2 \leq 2c_r$, with probability $1 - \tilde{\delta}$. The conditions of Lemma 32 then hold at $t_1 + t_2$. Switching to $\eta_2 \leq c_\eta C_2^{-1} \tilde{\varepsilon} s^{-1} \wedge c_\eta C_2^{-2} \tilde{\varepsilon}^2$ and running for $T_{1,3} + T_{1,4} = \tilde{\Theta}(\eta_2^{-1} \log \tilde{\varepsilon}^{-1})$ additional steps gives $\theta_\kappa^\top \hat{w}_{\kappa,n} \geq 1 - \tilde{\varepsilon}$ with high probability. Let $N'_\kappa = \Theta(N_\kappa)$ denote the number of well-initialized neurons (those with $\theta_\kappa^\top w_{\kappa,n}^0 \geq s^{-1/2}$), and let $F = \#\{n \in [N'_\kappa] : \theta_\kappa^\top \hat{w}_{\kappa,n} < 1 - \tilde{\varepsilon}\}$. A union bound and linearity of expectation give $\mathbb{E}[F] \leq 3N'_\kappa \tilde{\delta}$, and Markov's inequality gives $\mathbb{P}(F \geq N'_\kappa/2) \leq 6\tilde{\delta}$, so at least $N'_\kappa/2 = \Theta(N_\kappa)$ neurons satisfy the bound with probability $1 - 6\tilde{\delta}$. The total sample count is $T_{1,1} + T_{1,2} + T_{1,3} + T_{1,4} = \tilde{\Theta}\big(s^{3/2} \vee (s\,\tilde{\varepsilon}^{-1} \vee \tilde{\varepsilon}^{-2}) \log \tilde{\varepsilon}^{-1}\big)$, which reduces to $\tilde{O}(s^{3/2})$ under $\tilde{\varepsilon} = \tilde{\Theta}(s^{-1/2})$. □

Proposition 2 (Preservation of pre-trained features; formal version of Proposition 2)**.** Under the conditions of Theorem 1, let $\chi_k := N_\kappa \pi_k / (N_k \pi_\kappa)$ for $k \neq \kappa$, and assume $\chi_k \leq c_\chi \wedge c_2 (\log \tilde{\varepsilon}^{-1})^{-1}$. Then, at the output of Algorithm 1, with high probability, for every $k \neq \kappa$ and every $n \in [N_k]$,

$$\theta_k^\top \hat{w}_{k,n} \ \geq \ \theta_k^\top w_{k,n}^0 \ - \ c_2 C_3 \big(\|\Sigma_\kappa w_{k,n}^0\|^2 + \tilde{\varepsilon}\big) \ - \ c_2 \chi_k s^{-1/2}.$$

In particular, the drop in alignment is $\tilde{O}(\|\Sigma_\kappa w_{k,n}^0\|^2 + \tilde{\varepsilon} + \chi_k s^{-1/2})$.

Proof of Proposition 2. Fix $k \neq \kappa$ and $n \in [N_k]$, and set $\bar{\varepsilon} = C_2^{-1} \tilde{\varepsilon}$. Since $C_2 \mathcal{E}_s \leq \tilde{\varepsilon}$ by the hypothesis of Lemma 32, we have $\mathcal{E}_s \leq \bar{\varepsilon}$. Since $\chi_k \leq c_\chi$, Lemma 34(ii) gives

$$\theta_k^\top \hat{w}_{k,n} \geq \theta_k^\top w_{k,n}^0 - \chi_k (C_3 \vee 2C_2 \log \bar{\varepsilon}^{-1}) \big(\|\Sigma_\kappa w_{k,n}^0\|^2 + \mathcal{E}_s\big) - \chi_k c_3 \bar{\varepsilon} \log \bar{\varepsilon}^{-1} - c_2 (\chi_k s^{-3/4} + \chi_k^2 s^{-1/2}).$$

Using $\chi_k \leq c_2 (\log \tilde{\varepsilon}^{-1})^{-1}$ and $\log \bar{\varepsilon}^{-1} = \log(C_2 \tilde{\varepsilon}^{-1}) \leq 2 \log \tilde{\varepsilon}^{-1}$ (for $\tilde{\varepsilon}$ small), the log factors cancel: $\chi_k (C_3 \vee 2C_2 \log \bar{\varepsilon}^{-1}) \leq c_2 C_3$ and $\chi_k c_3 \bar{\varepsilon} \log \bar{\varepsilon}^{-1} \leq c_2 c_3 \bar{\varepsilon} \leq c_2 c_3 \tilde{\varepsilon}$. Since $\mathcal{E}_s \leq \bar{\varepsilon} \leq \tilde{\varepsilon}$ and $s^{-3/4} \leq s^{-1/2}$ and $\chi_k^2 s^{-1/2} \leq \chi_k s^{-1/2}$, substituting yields the stated bound. □

The remainder of this section proves Theorem 1 and Proposition 2.

E.1 Nonlinear Transformation of the Teacher Signal

We show that transforming the weak model's output reduces its information exponent. Let r^* be a polynomial of degree q with $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[r^*(z)] = 0$ and $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[(r^*(z))^2] = 1$. Let $\rho = \Theta(\log^{C_\rho} d)$ be a temperature parameter with $C_\rho \geq q/2 + 1$, and define $\bar{r}^*(z) = \mathrm{clip}(\rho^{-1}r^*; \pm 1/\log d)$. By Lemma 17, $\bar{r}^* = \rho^{-1}r^*$ with high probability.

Lemma 17 (Oko et al. (2024b), Corollary 17)**.** Let $\theta \in S^{d-1}$ and $x \sim \mathcal{N}(0, I_d)$. Then $|r^*(\theta^\top x)| \lesssim (\log d)^{q/2}$ with high probability.

Write the Hermite expansion of $\bar{r}^* \exp(\bar{r}^*)$ as $\bar{r}^* \exp(\bar{r}^*) = \sum_{i=1}^\infty \frac{r_i}{i!}\mathrm{He}_i$.

Lemma 18 (Clip version of Nishikawa et al. (2025), Lemma 9)**.** Let $p \geq \mathrm{GE}(r^*)$ and let $e_p = \min\{i \geq 1 \mid \mathrm{IE}((r^*)^i) \leq p\}$, where we assume $e_p < \infty$. Then $r_p = \Theta((\log d)^{-C_\rho e_p})$.

On the other hand, when $\mathrm{GE}(r^*) > p \geq 1$, $r_p = O(d^{-C})$ for some sufficiently large constant C.

Proof. The proof follows Nishikawa et al. (2025), Lemma 9. We first treat the case $p \geq \mathrm{GE}(r^*)$. By Taylor expansion,

$$
\begin{aligned}
r_p &= \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\bar{r}^*(z)\exp(\bar{r}^*(z))\mathrm{He}_p(z)] = \mathbb{E}\left[\sum_{i=1}^\infty \frac{1}{(i-1)!}(\bar{r}^*(z))^i \mathrm{He}_p(z)\right] \\
&= \frac{1}{(e_p-1)!}\mathbb{E}[(\bar{r}^*(z))^{e_p}\mathrm{He}_p(z)] + \sum_{i=e_p}^{C_\rho e_p - 1}\frac{1}{i!}\mathbb{E}[(\bar{r}^*(z))^{i+1}\mathrm{He}_p(z)] + \sum_{i\geq C_\rho e_p}\frac{1}{i!}\mathbb{E}[(\bar{r}^*(z))^{i+1}\mathrm{He}_p(z)] \\
&=: I_1 + I_2 + I_3.
\end{aligned}
$$

For I_1, we write

$$
\begin{aligned}
&\mathbb{E}[(\bar{r}^*(z))^{e_p}\mathrm{He}_p(z)] \\
&= \rho^{-e_p}\mathbb{E}[(r^*(z))^{e_p}\mathrm{He}_p(z)] - \rho^{-e_p}\mathbb{E}\left[\left((r^*(z))^{e_p} - \mathrm{sign}(r^*(z))\left(\tfrac{\rho}{\log d}\right)^{e_p}\right)\mathrm{He}_p(z)\mathbf{1}(|r^*(z)| \geq \rho/\log d)\right].
\end{aligned}
$$

By the Cauchy–Schwarz inequality,

$$
\mathbb{E}[|(r^*(z))^{e_p}\mathrm{He}_p(z)\mathbf{1}(|r^*(z)| \geq \rho/\log d)|] \leq \mathbb{E}[|(r^*(z))^{e_p}\mathrm{He}_p(z)|^2]^{1/2}\mathbb{P}[|r^*(z)| \geq \rho/\log d]^{1/2},
$$

$$
\mathbb{E}\left[\left|\mathrm{sign}(r^*(z))\left(\rho(\log d)^{-1}\right)^{e_p}\mathrm{He}_p(z)\mathbf{1}(|r^*(z)| \geq \rho/\log d)\right|\right] \leq \left(\rho(\log d)^{-1}\right)^{e_p}\mathbb{E}[|\mathrm{He}_p(z)|^2]^{1/2}\mathbb{P}[|r^*(z)| \geq \rho/\log d]^{1/2}.
$$

By Lemma 17, $\mathbb{P}[|r^*(z)| \geq \rho/\log d] = O(d^{-1})$, so $I_1 = \Theta((\log d)^{-C_\rho e_p})$.

For I_2, similarly to I_1, the leading term of $\mathbb{E}[(\bar{r}^*(z))^{i+1}\mathrm{He}_p(z)]$ for $i = e_p, \ldots, C_\rho e_p - 1$ is $\rho^{-(i+1)}\mathbb{E}[(r^*(z))^{i+1}\mathrm{He}_p(z)]$. By Cauchy–Schwarz inequality, $|\rho^{-(i+1)}\mathbb{E}[(r^*(z))^{i+1}\mathrm{He}_p(z)]| \lesssim \rho^{-(i+1)}\sqrt{\mathbb{E}[(r^*(z))^{2(i+1)}]}$, so

$$
|I_2| \lesssim \sum_{i=e_p}^{C_\rho e_p - 1}\rho^{-(i+1)}\sqrt{\mathbb{E}[(r^*(z))^{2(i+1)}]} \lesssim \rho^{-(e_p+1)} = O((\log d)^{-C_\rho(e_p+1)}).
$$

For I_3, by the boundedness of $\bar{r}^*$ and Cauchy–Schwarz, for $i \geq C_\rho e_p$,

$$
\left|\frac{1}{i!}\mathbb{E}[(\bar{r}^*(z))^{i+1}\mathrm{He}_p(z)]\right| \leq \frac{1}{i!}\mathbb{E}[\bar{r}^*(z)^{2(i+1)}]^{1/2}\mathbb{E}[\mathrm{He}_p(z)^2]^{1/2} \lesssim \frac{1}{i!}(\log d)^{-(i+1)},
$$

giving $|I_3| \lesssim \sum_{i\geq C_\rho e_p}\frac{1}{i!}(\log d)^{-(i+1)} \lesssim (\log d)^{-(C_\rho e_p+1)}$. Combining, $I_1 = \Theta((\log d)^{-C_\rho e_p})$ dominates I_2 and I_3, giving $r_p = \Theta((\log d)^{-C_\rho e_p})$.

We now treat the case $\mathrm{ge}(r^*) > p \geq 1$. Write

$$\begin{aligned} r_p &= \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\bar{r}^*(z)\exp(\bar{r}^*(z))\mathrm{He}_p(z)] \\ &= \mathbb{E}[\rho^{-1}r^*(z)\exp(\rho^{-1}r^*(z))\mathrm{He}_p(z)] \\ &\quad - \mathbb{E}\big[\big(\rho^{-1}r^*(z)\exp(\rho^{-1}r^*(z)) - \mathrm{sign}(r^*(z))(\log d)^{-1}\exp((\log d)^{-1})\big)\,\mathrm{He}_p(z)\mathbf{1}(|r^*(z)| \geq \rho/\log d)\big]\,. \end{aligned}$$

Since $p < \mathrm{ge}(r^*)$, $\mathbb{E}[\rho^{-1}r^*(z)\exp(\rho^{-1}r^*(z))\mathrm{He}_p(z)] = 0$. By Lemma 17,

$$\begin{aligned} &\mathbb{E}\big[|\rho^{-1}r^*(z)\exp(\rho^{-1}r^*(z))\mathrm{He}_p(z)\mathbf{1}(|r^*(z)| \geq \rho/\log d)|\big] \\ &\leq \mathbb{E}[|\rho^{-1}r^*(z)\exp(\rho^{-1}r^*(z))\mathrm{He}_p(z)|^2]^{1/2}\mathbb{P}[|r^*(z)| \geq \rho/\log d]^{1/2} = O(d^{-C}), \\ &\mathbb{E}\big[|\,\mathrm{sign}(r^*(z))(\log d)^{-1}\exp((\log d)^{-1})\mathrm{He}_p(z)\mathbf{1}(|r^*(z)| \geq \rho/\log d)|\big] \\ &\leq (\log d)^{-1}e^{(\log d)^{-1}}\mathbb{E}[|\mathrm{He}_p(z)|^2]^{1/2}\mathbb{P}[|r^*(z)| \geq \rho/\log d]^{1/2} = O(d^{-C}). \end{aligned}$$

Therefore $r_p = O(d^{-C})$. □

E.2 Initialization

Lemma 19. Under Assumption 5 (ii), $\theta_\kappa^\top w_{\kappa,n}^0 \geq s^{-1/2}$ with probability $\Theta(1)$.

Proof. Write $w^0 := w_{\kappa,n}^0$ and let $\tilde{w} := \Sigma_\kappa w^0/\|\Sigma_\kappa w^0\|$. By assumption, $\tilde{w} \sim \mathrm{Unif}(S^{d-1} \cap V_\kappa)$, so identifying $V_\kappa \cong \mathbb{R}^s$ the argument of Lemma 1 gives $\theta_\kappa^\top \tilde{w} \geq 2s^{-1/2}$ with probability $\Theta(1)$. Since $\|\Sigma_\kappa^\perp w^0\|^2 \leq c_r s^{-1/2}$ by assumption, we have $\|\Sigma_\kappa w^0\| \geq \sqrt{1 - c_r s^{-1/2}} \geq 1/2$. Therefore $\theta_\kappa^\top w^0 = \|\Sigma_\kappa w^0\| \cdot \theta_\kappa^\top \tilde{w} \geq s^{-1/2}$. □

The following lemma shows that the Hermite sign conditions in Assumption 5 (ii) are satisfied with positive probability under a natural random initialization.

Lemma 20. Let $a_n \sim \mathrm{Unif}\{\pm\pi_\kappa\}$ and $\xi_{n,i} \sim \mathrm{Unif}\{\pm\delta_i\}$ for $0 \leq i \leq q$, where $(\delta_i)_{i=0}^q$ satisfies $\delta_0 = \delta_1 = \delta_2 = 1$ and $0 < \delta_{i+2} < \frac{i}{(i+1)(i+2)}\delta_i$ for every $i = 1, 2, \ldots, q-2$.

Then $a_n\sigma_n = \sum_{i=0}^q a_n\xi_{n,i}\mathrm{He}_i$ has its i-th Hermite coefficient ($1 \leq i \leq q$) taking the desired sign with probability $\Omega(1)$. Taking these a_n, σ_n, and $b_n = 0$ as initial values therefore gives a strong model initialization satisfying Assumption 5 (ii).

Proof. The coefficient $\beta_{\kappa,n,i} = \sqrt{i!}\,a_n\xi_{n,i}$ takes any given sign with probability 2^{-q}. By the construction of $\xi_{n,i}$, we have $i|\beta_{\kappa,n,i}| > \sqrt{(i+2)(i+1)}|\beta_{\kappa,n,i+2}|$ for every neuron and every $i = 1, 2, \ldots, q$. Combined with $\bar{\alpha}_2 = \tilde{\Theta}(1)$ from Lemma 18, the desired conclusion follows. □

E.3 Error of the Teacher Signal in the Subspace

Lemma 21. Let $x \sim \mathcal{N}(0, I_d)$. Let σ_n be a degree-q polynomial with uniformly bounded coefficients over $n = 1, 2, \ldots, N$, and let $f(x) = \frac{1}{N}\sum_{n=1}^N a_n\sigma_n(w_n^\top x + b_n)$. Suppose $\max_n |b_n| = \tilde{O}(1)$, and that there exists $\theta \in S^{d-1}$ such that $|w_n^\top\theta| \geq 1 - \tilde{\varepsilon}$ for all n. Let $V \subset \mathbb{R}^d$ be a subspace, and let P_V, $P_V^\perp$ denote the orthogonal projections onto V and its complement respectively. Then the following hold for the L^1 norm of f restricted to V.

(i) If $\theta \in V$, then $\mathbb{E}_x[|f(P_V x)|] \leq \mathbb{E}_x[|f(x)|] + \tilde{O}(N^{-1}\|a\|_1\tilde{\varepsilon})$.

(ii) If $\theta \perp V$, then $\mathbb{E}[|f(P_V x)|] = \tilde{O}(N^{-1}\|a\|_1 + \tilde{\varepsilon})$.

Proof. Set $u_n = w_n - (w_n^\top\theta)\theta$, so that

$$\|u_n\| = \sqrt{\|w_n - (w_n^\top\theta)\theta\|^2} = \sqrt{1 - (w_n^\top\theta)^2} \leq \sqrt{2\tilde{\varepsilon}}. \tag{10}$$

For $x \sim \mathcal{N}(0, I_d)$, write $x_1 = P_V x$ and $x_2 = P_V^\perp x$; these are independent.

Proof of (i). Set $\delta_n = w_n^\top P_V^\perp x$ and expand $\sigma_n(w_n^\top x + b_n)$ in a Taylor series around $w_n^\top P_V x + b_n$:

$$\begin{aligned} f(x) &= \frac{1}{N}\sum_{n=1}^{N} a_n \sigma(w_n^\top P_V x + b_n + \delta_n) \\ &= \frac{1}{N}\sum_{n=1}^{N} a_n \sigma(w_n^\top P_V x + b_n) + \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{q} \frac{1}{i!} a_n (\delta_n)^i \sigma_n^{(i)}(w_n^\top P_V x + b_n) \\ &= f(P_V x) + \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{q} \frac{1}{i!} a_n (\delta_n)^i \sigma_n^{(i)}(w_n^\top P_V x + b_n) \end{aligned} \tag{11}$$

Since $\mathbb{E}_{P_V^\perp x}[\delta_n] = 0$ and $\frac{(P_V^\perp u_n)^\top}{\|P_V^\perp u_n\|} x \sim \mathcal{N}(0,1)$, we have by (10)

$$|\delta_n| = |w_n^\top P_V^\perp x| = |u_n^\top P_V^\perp x| \le \|u_n\| \left| \frac{(P_V^\perp u_n)^\top}{\|P_V^\perp u_n\|} x \right| = \tilde{O}(\tilde{\varepsilon}^{1/2})$$

with high probability. Taking the expectation of (11) over $x_2 = P_V^\perp x$,

$$\begin{aligned} |f(x_1)| &= \left| \mathbb{E}_{x_2}[f(x_1 + x_2)] - \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{q} \frac{1}{i!} a_n \mathbb{E}_{x_2}[(\delta_n)^i] \sigma_n^{(i)}(w_n^\top x_1 + b_n) \right| \\ &\le \mathbb{E}_{x_2}[|f(x_1 + x_2)|] + \frac{1}{N}\sum_{n=1}^{N}\sum_{i=2}^{q} \frac{1}{i!} \left| a_n \mathbb{E}_{x_2}[(\delta_n)^i] \sigma_n^{(i)}(w_n^\top x_1 + b_n) \right| \end{aligned}$$

Taking the expectation over $x_1 = P_V x$ as well, and using the high-probability bound $|\delta_n|^i = \tilde{O}(\tilde{\varepsilon}^{1/2})$,

$$\mathbb{E}_{x_1,x_2}[|f(x_1)|] \le \mathbb{E}_{x_1,x_2}[|f(x_1 + x_2)|] + \tilde{O}(N^{-1}\|a\|_1 \tilde{\varepsilon}).$$

Proof of (ii). Since $\theta \perp V$, we have $w_n^\top P_V x + b_n = u_n^\top P_V x + b_n$, so

$$f(P_V x) = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=0}^{q} \frac{1}{i!} a_n (u_n^\top P_V x)^i \sigma_n^{(i)}(b_n).$$

By the same argument as in (i), $\mathbb{E}_x[u_n^\top P_V x] = 0$ and $\mathbb{E}_x[|u_n^\top P_V x|^i] = \tilde{O}(\tilde{\varepsilon}^{1/2})$ for $i \ge 2$, giving

$$\mathbb{E}_x[|f(P_V x)|] \le \frac{1}{N}\sum_{n=1}^{N} |a_n \sigma_n(b_n)| + \tilde{O}(\tilde{\varepsilon}) = \tilde{O}(N^{-1}\|a\|_1 + \tilde{\varepsilon}).$$

□

E.4 Gradient Decomposition

The following holds for the effective gradient g_k^t. Part (i) in Lemma 22 separates the signal and noise terms for the target task κ, analogously to Lemma 4. Part (ii) describes the effect of the teacher signal on neurons belonging to tasks $k \neq \kappa$; no clear signal term exists in this case. The task weighting introduces a scaling factor $\chi_k = \frac{N_\kappa \pi_k}{N_k \pi_\kappa}$ for the gradient.

Lemma 22 (Formal version of Section 3.2)**.** The gradient decomposes as follows.

(i) For $k = \kappa$: setting $\tilde{w}_\kappa^t = \Sigma_\kappa w_\kappa^t / \|\Sigma_\kappa w_\kappa^t\|$,

$$g_\kappa^t = \sum_{i=1}^{q} \lambda_\kappa \left[i\bar{\alpha}_i \beta_{\kappa,i} (\theta_\kappa^\top \tilde{w}_\kappa^t)^{i-1} \theta_\kappa + \sqrt{(i+2)(i+1)}\bar{\alpha}_i \beta_{\kappa,i+2} (\theta_\kappa^\top \tilde{w}_\kappa^t)^i \tilde{w}_\kappa^t \right] + Z^t + R^t.$$

Here Z^t is a mean-zero random variable with $\|Z^t\| = \tilde{O}(s^{1/2})$ with high probability, and $|v^\top Z^t| = \tilde{O}(1)$ with high probability for every $v \in S^{d-1}$. With probability λ_κ, $\|\Sigma_\kappa^\perp Z^t\| = \tilde{O}((1-\lambda_\kappa)s^{1/2})$, and $\mathbb{E}[\|\Sigma_\kappa^\perp Z^t\|^2] \lesssim (1-\lambda_\kappa)s$. The remainder R^t satisfies $\|R^t\| = \tilde{O}((\varepsilon_w \vee$

$\tilde{\varepsilon}_w \vee (1-\lambda_\kappa) \vee \|\Sigma_\kappa^\perp w_\kappa^t\|^2) s^{1/2})$ with high probability, $\|\Sigma_\kappa^\perp R^t\| = \tilde{O}((1-\lambda_\kappa)s^{1/2})$ with high probability, and for every $v \in S^{d-1}$,

$$|v^\top R^t| = \tilde{O}(\varepsilon_w \vee \tilde{\varepsilon}_w \vee (1-\lambda_\kappa) \vee \|\Sigma_\kappa v\| \|\Sigma_\kappa^\perp w_\kappa^t\|^2)$$

with high probability. Furthermore, $\|g_\kappa^t\| = \tilde{O}(s^{1/2})$ with high probability, and $g_\kappa^{t\top} v = \tilde{O}(1)$ with high probability for every $v \in \mathbb{R}^d$ with $\|v\| = O(1)$.

(ii) For $k \neq \kappa$:

$$g_k^t = \chi_k (Z^t + R^t), \qquad \chi_k = \frac{N_\kappa \pi_k}{N_k \pi_\kappa}.$$

Here Z^t is mean-zero with $\|Z^t\| = \tilde{O}(s^{1/2})$ with high probability and $|v^\top Z^t| = \tilde{O}(1)$ with high probability for every $v \in S^{d-1}$. With probability λ_κ, $\Sigma_\kappa^\perp Z^t = 0$, and $\mathbb{E}[\|\Sigma_\kappa^\perp Z^t\|^2] \lesssim (1-\lambda_\kappa)s$. The remainder satisfies $\|R^t\| = \tilde{O}((\varepsilon_w \vee \tilde{\varepsilon}_w \vee (1-\lambda_\kappa) \vee \|\Sigma_\kappa w_k^t\|) s^{1/2})$ with high probability, and for every $v \in S^{d-1}$,

$$|v^\top R^t| = \tilde{O}(\varepsilon_w \vee \tilde{\varepsilon}_w \vee (1-\lambda_\kappa) \vee \|\Sigma_\kappa v\| \|\Sigma_\kappa w_k^t\|)$$

with high probability. Furthermore, $\|g_k^t\| = \tilde{O}(\chi_k s^{1/2})$ with high probability, and $g_k^{t\top} v = \tilde{O}(\chi_k)$ with high probability for every $v \in \mathbb{R}^d$ with $\|v\| = O(1)$.

Proof. Proof of (i). By Lemma 18, $|\bar{\alpha}_2| = \tilde{\Theta}(1)$, so by Assumption 5 (ii) the absolute values of the Hermite coefficients of $\pi_\kappa^{-1} a_\kappa \sigma_\kappa(\cdot + b_\kappa)$ satisfy $|\beta_{\kappa,i}| \le C_\beta = \tilde{O}(1)$ for all $i \le q$. Define

$$\begin{aligned}
R_1^t &= \nabla_{w_\kappa} \mathbb{E}_{x\sim\mu}[\bar{y}^t a_\kappa \sigma_\kappa(w_\kappa^{t\top} x + b_\kappa)] - \nabla_{w_\kappa} \mathbb{E}_{x\sim\mu}[\bar{r}_\kappa^*(\theta_\kappa^\top x) \exp(\bar{r}_\kappa^*(\theta_\kappa^\top x)) a_\kappa \sigma_\kappa(w_\kappa^{t\top} x + b_\kappa)], \\
R_2^t &= \nabla_{w_\kappa} \sum_{k' \neq \kappa} \lambda_{k'} \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_{k'})}[\bar{r}_\kappa^*(\theta_\kappa^\top x) \exp(\bar{r}_\kappa^*(\theta_\kappa^\top x)) a_\kappa \sigma_\kappa(w_\kappa^{t\top} x + b_\kappa)], \\
R_3^t &= \lambda_\kappa \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_\kappa)} \left[\bar{r}_\kappa^*(\theta_\kappa^\top x) \exp(\bar{r}_\kappa^*(\theta_\kappa^\top x)) a_\kappa \left(\sigma_\kappa'(w_\kappa^{t\top} x + b_\kappa) - \sigma_\kappa'(\tilde{w}_\kappa^{t\top} x + b_\kappa)\right) x\right], \\
R^t &= \pi_\kappa^{-1}(R_1^t + R_2^t + R_3^t).
\end{aligned}$$

Then

$$\begin{aligned}
&\nabla_{w_\kappa} \mathbb{E}_{x\sim\mu}[\bar{y}^t a_\kappa \sigma_\kappa(w_\kappa^{t\top} x + b_\kappa)] \\
&= \lambda_\kappa \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_\kappa)} \left[\bar{r}_\kappa^*(\theta_\kappa^\top x) \exp(\bar{r}_\kappa^*(\theta_\kappa^\top x)) a_\kappa \sigma_\kappa'(\tilde{w}_\kappa^{t\top} x + b_\kappa) x\right] + R_1^t + R_2^t + R_3^t.
\end{aligned}$$

For R_1^t, we have

$$\begin{aligned}
R_1^t &= \nabla_{w_\kappa} \mathbb{E}_{x\sim\mu} \left[\bar{y}^t a_\kappa \sigma_\kappa(w_\kappa^{t\top} x + b_\kappa)\right] - \nabla_{w_\kappa} \mathbb{E}_{x\sim\mu} \left[\bar{r}_\kappa^*(\theta_\kappa^\top x) \exp(\bar{r}_\kappa^*(\theta_\kappa^\top x) a_\kappa \sigma_\kappa(w_\kappa^{t\top} x + b_\kappa)\right] \\
&= \mathbb{E}_{x\sim\mu} \left[\left(\bar{r}^w(x) \exp(\bar{r}^w(x)) - \bar{r}_\kappa^*(\theta_\kappa^\top x) \exp(\bar{r}_\kappa^*(\theta_\kappa^\top x)\right) a_\kappa \sigma_\kappa'(w_\kappa^{t\top} x + b_\kappa) x\right] \\
&= \mathbb{E}_{x\sim\mu} \left[\sum_{j=0}^\infty \frac{1}{j!} \left(\left(\bar{r}^w(x)^{j+1}\right) - \left(\bar{r}_\kappa^*(\theta_\kappa^\top x)\right)^{j+1}\right) a_\kappa \sigma_\kappa'(w_\kappa^{t\top} x + b_\kappa) x\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
&\|R_1^t\| \\
&\le \sum_{j=0}^\infty \frac{1}{j!} \mathbb{E}_{x\sim\mu} \left[\left|\left(\bar{r}^w(x) - \bar{r}_\kappa^*(\theta_\kappa^\top x)\right)\right| \left[(\bar{r}^w(x))^j + (\bar{r}^w(x))^{j-1} \bar{r}_\kappa^*(\theta_\kappa^\top x) + \cdots + (\bar{r}^w(x))^j\right]\right| \left\|a_\kappa \sigma_\kappa'(w_\kappa^{t\top} x + b_\kappa) x\right\|] \\
&\le \left(\sum_{j=0}^\infty \frac{j+1}{j!} (\log d)^{-j}\right) \mathbb{E}_{x\sim\mu} \left[\left|\left(\bar{r}^w(x) - \bar{r}_\kappa^*(\theta_\kappa^\top x)\right)\right| \left\|a_\kappa \sigma_\kappa'(w_\kappa^{t\top} x + b_\kappa) x\right\|\right]
\end{aligned}$$

Since $\pi_\kappa^{-1}\|a_\kappa\sigma'_\kappa(w_\kappa^{t\top}x+b_\kappa)x\| = \tilde{O}(s^{1/2})$ with high probability for $x\sim\mu$, and by Lemma 21, $\mathbb{E}_{x\sim\mu}[|\bar{r}^w(x)-\bar{r}^*_\kappa(\theta_\kappa^\top x)|] = \tilde{O}(\varepsilon_w\vee\tilde{\varepsilon}_w\vee(1-\lambda_\kappa))$, we obtain $\pi_\kappa^{-1}\|R_1^t\| = \tilde{O}((\varepsilon_w\vee\tilde{\varepsilon}_w\vee(1-\lambda_\kappa))s^{1/2})$.

For R_2^t: since $\pi_\kappa^{-1}\|\bar{r}^*_\kappa(\theta_\kappa^\top x)\exp(\bar{r}^*_\kappa(\theta_\kappa^\top x))a_\kappa\sigma'_\kappa(w_\kappa^{t\top}x+b_\kappa)x\| = \tilde{O}(s^{1/2})$ for $x\sim\mathcal{N}(0,\Sigma_{k'})$ with $k'\neq\kappa$, we get $\pi_\kappa^{-1}\|R_2^t\| = \tilde{O}((1-\lambda_\kappa)s^{1/2})$.

For R_3^t: by Taylor expansion,

$$|\sigma'_\kappa(w_\kappa^{t\top}x+b_\kappa)-\sigma'_\kappa(\tilde{w}_\kappa^{t\top}x+b_\kappa)| \le \left|\sum_{i=1}^{q-1}\frac{1}{i!}((w_\kappa^t-\tilde{w}_\kappa^t)^\top x)^i\sigma_\kappa^{(i+1)}(\tilde{w}_\kappa^{t\top}x+b_\kappa)\right|.$$

For $x\sim\mathcal{N}(0,\Sigma_\kappa)$, $(w_\kappa^t-\tilde{w}_\kappa^t)^\top x = (\|\Sigma_\kappa w_\kappa^t\|-1)\tilde{w}_\kappa^{t\top}x$, and since $1-\|\Sigma_\kappa w_\kappa^t\| = \frac{\|\Sigma_\kappa^\perp w_\kappa^t\|^2}{1+\|\Sigma_\kappa w_\kappa^t\|} \le \|\Sigma_\kappa^\perp w_\kappa^t\|^2$ and $\tilde{w}_\kappa^{t\top}x\sim\mathcal{N}(0,1)$, we get $\pi_\kappa^{-1}\|R_3^t\| = \tilde{O}(\|\Sigma_\kappa^\perp w_\kappa^t\|^2 s^{1/2})$.

Combining, $\|R^t\| = \tilde{O}((\varepsilon_w\vee\tilde{\varepsilon}_w\vee(1-\lambda_\kappa)\vee\|\Sigma_\kappa^\perp w_\kappa^t\|^2)s^{1/2})$. Since $\Sigma_\kappa^\perp R^t = \pi_\kappa^{-1}\Sigma_\kappa^\perp(R_1^t+R_2^t)$, we have $\|\Sigma_\kappa^\perp R^t\| = \tilde{O}((1-\lambda_\kappa)s^{1/2})$. For every $v\in S^{d-1}$, since $v^\top x$ is sub-Gaussian with parameter $\|\Sigma_k v\|$ for $x\sim\mu_k$, $|v^\top R^t| = \tilde{O}(\varepsilon_w\vee\tilde{\varepsilon}_w\vee(1-\lambda_\kappa)\vee\|\Sigma_\kappa v\|\|\Sigma_\kappa^\perp w_\kappa^t\|^2)$ with high probability.

By the same computation as in Lemma 4,

$$\begin{aligned}&\lambda_\kappa\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_\kappa)}\left[\bar{r}^*_\kappa(\theta_\kappa^\top x)\exp(\bar{r}^*_\kappa(\theta_\kappa^\top x))a_\kappa\sigma'_\kappa(\tilde{w}_\kappa^{t\top}x+b_\kappa)x\right]\\ &=\sum_{i=1}^{q}\lambda_\kappa\left[i\bar{\alpha}_i\tilde{\beta}_{\kappa,i}(\theta_\kappa^\top\tilde{w}_\kappa^t)^{i-1}\theta_\kappa+\sqrt{(i+2)(i+1)}\bar{\alpha}_i\tilde{\beta}_{\kappa,i+2}(\theta_\kappa^\top\tilde{w}_\kappa^t)^i\tilde{w}_\kappa^t\right].\end{aligned}$$

Setting $Z^t = \mathbb{E}_{x\sim\mu}[g_\kappa^t]-g_\kappa^t$, we have $\mathbb{E}[Z^t]=0$, $\|Z^t\| = \tilde{O}(s^{1/2})$, and $|v^\top Z^t| = \tilde{O}(1)$ for every $v\in S^{d-1}$. By construction, with probability λ_κ, $\|\Sigma_\kappa^\perp Z^t\| = \tilde{O}((1-\lambda_\kappa)s^{1/2})$, giving $\mathbb{E}[\|\Sigma_\kappa^\perp Z^t\|^2]\lesssim(1-\lambda_\kappa)s$.

Proof of (ii). Define $\tilde{w}_k^t = \Sigma_k w_k^t/\|\Sigma_k w_k^t\|$. By an argument identical to (i), $|\beta_{k,i}|\le C_\beta = \tilde{O}(1)$. Define $R_1^t, R_2^t, R_3^t, R_4^t$ as

$$\begin{aligned}R_1^t &= \nabla_{w_k}\mathbb{E}_{x\sim\mu}[\bar{y}^t a_k\sigma_k(w_k^{t\top}x+b_k)]-\nabla_{w_k}\mathbb{E}_{x\sim\mu}[\bar{r}^*_\kappa(\theta_\kappa^\top x)\exp(\bar{r}^*_\kappa(\theta_\kappa^\top x))a_k\sigma_k(w_k^{t\top}x+b_k)],\\ R_2^t &= \nabla_{w_k}\sum_{k'\neq\kappa}\lambda_{k'}\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_{k'})}[\bar{r}^*_\kappa(\theta_\kappa^\top x)\exp(\bar{r}^*_\kappa(\theta_\kappa^\top x))a_k\sigma_k(w_k^{t\top}x+b_k)],\\ R_3^t &= \lambda_\kappa\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_\kappa)}\left[\bar{r}^*_\kappa(\theta_\kappa^\top x)\exp(\bar{r}^*_\kappa(\theta_\kappa^\top x))a_k\left(\sigma'_k(w_k^{t\top}x+b_k)-\sigma'_k(\tilde{w}_k^{t\top}x+b_k)\right)x\right],\\ R_4^t &= \lambda_\kappa\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_\kappa)}\left[\bar{r}^*_\kappa(\theta_\kappa^\top x)\exp(\bar{r}^*_\kappa(\theta_\kappa^\top x))a_k\sigma'_k(\tilde{w}_k^{t\top}x+b_k)x\right] = \lambda_\kappa\bar{\alpha}_1\sigma'_k(b_k)\theta_\kappa,\end{aligned}$$

and set $R^t = \pi_k^{-1}(R_1^t+R_2^t+R_3^t+R_4^t)$. By the same bounds as in (i), $\pi_k^{-1}\|R_1^t\| = \tilde{O}((\varepsilon_w\vee\tilde{\varepsilon}_w\vee(1-\lambda_\kappa))s^{1/2})$ and $\pi_k^{-1}\|R_2^t\| = \tilde{O}((1-\lambda_\kappa)s^{1/2})$. For R_3^t, since $w_k^{t\top}x/\|\Sigma_\kappa w_k^t\|\sim\mathcal{N}(0,1)$ for $x\sim\mathcal{N}(0,\Sigma_\kappa)$,

$$R_3^t = \lambda_\kappa\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_\kappa)}\left[\bar{r}^*_\kappa(\theta_\kappa^\top x)\exp(\bar{r}^*_\kappa(\theta_\kappa^\top x))a_k\left(\sum_{i=1}^{q-1}\frac{1}{i!}((w_k^t-\tilde{w}_k^t)^\top x)^i\sigma_k^{(i+1)}(\tilde{w}_k^{t\top}x+b_k)\right)x\right]$$

$$=\sum_{i=1}^{q-1}\frac{1}{i!}\lambda_\kappa a_k\sigma_k^{(i+1)}(b_k)\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_\kappa)}\left[\bar{r}^*_\kappa(\theta_\kappa^\top x)\exp(\bar{r}^*_\kappa(\theta_\kappa^\top x))(w_k^{t\top}x)^i x\right].$$

giving $\pi_k^{-1}\|R_3^t\| = \tilde{O}(\|\Sigma_\kappa w_k^t\|s^{1/2})$. By Assumption 5 and Lemma 18, $\pi_k^{-1}\|R_4^t\| = \tilde{O}(d^{-C})$ for a sufficiently large constant C. Setting $Z^t = \chi_k^{-1}(\mathbb{E}_{x\sim\mu}[g_k^t]-g_k^t)$, one verifies that R^t, Z^t, g_k^t satisfy the stated bounds by the same argument as in (i). □

For the rest of the analysis, define the effective error of the strong model training as

$$\mathcal{E}_s = \varepsilon_w\vee\tilde{\varepsilon}_w\vee(1-\lambda_\kappa).$$

Lemma 23. Let $\eta = \eta^t \le c_\eta s^{-1}$ and $\theta_\kappa^\top w_\kappa^t \ge \frac{1}{2} s^{-1/2}$. Then

$$\begin{aligned}
&\theta_\kappa^\top w_\kappa^t + \eta\lambda_\kappa \theta_\kappa^\top P_{w_\kappa^t}^\perp (Z^t + R^t) - \eta^2 C_1^2 (\theta_\kappa^\top w_\kappa^t) s \\
&\quad + \eta\lambda_\kappa \sum_{i=1}^{q} \Big[i\bar{\alpha}_i \beta_{\kappa,i} \|\Sigma_\kappa w_\kappa^t\|^{-(i-1)} (\theta_\kappa^\top w_\kappa^t)^{i-1} (1 - (\theta_\kappa^\top w_\kappa^t)^2) \\
&\qquad + \sqrt{(i+2)(i+1)} \bar{\alpha}_i \beta_{\kappa,i+2} \|\Sigma_\kappa w_\kappa^t\|^{-(i+1)} \|\Sigma_\kappa^\perp w_\kappa^t\|^2 (\theta_\kappa^\top w_\kappa^t)^{i+1} \Big] \\
&\le \theta_\kappa^\top w_\kappa^{t+1} \\
&\le \theta_\kappa^\top w_\kappa^t + \eta\lambda_\kappa \theta_\kappa^\top P_{w_\kappa^t}^\perp (Z^t + R^t) \\
&\quad + \eta\lambda_\kappa \sum_{i=1}^{q} \Big[i\bar{\alpha}_i \beta_{\kappa,i} \|\Sigma_\kappa w_\kappa^t\|^{-(i-1)} (\theta_\kappa^\top w_\kappa^t)^{i-1} (1 - (\theta_\kappa^\top w_\kappa^t)^2) \\
&\qquad + \sqrt{(i+2)(i+1)} \bar{\alpha}_i \beta_{\kappa,i+2} \|\Sigma_\kappa w_\kappa^t\|^{-(i+1)} \|\Sigma_\kappa^\perp w_\kappa^t\|^2 (\theta_\kappa^\top w_\kappa^t)^{i+1} \Big],
\end{aligned}$$

where Z^t, R^t satisfy the conditions in Lemma 22. Furthermore, $|\theta_\kappa^\top w_\kappa^{t+1} - \theta_\kappa^\top w_\kappa^t| = \tilde{O}(\eta)$.

Proof. From $\tilde{w}_\kappa^t = \Sigma_\kappa w_\kappa^t / \|\Sigma_\kappa w_\kappa^t\|$,

$$\theta_\kappa^\top \tilde{w}_\kappa^t = \frac{\theta_\kappa^\top w_\kappa^t}{\|\Sigma_\kappa w_\kappa^t\|}, \quad \theta_\kappa^\top P_{w_\kappa^t}^\perp \tilde{w}_\kappa^t = \theta_\kappa^\top \tilde{w}_\kappa^t - (\theta_\kappa^\top w_\kappa^t)(w_\kappa^{t\top} \tilde{w}_\kappa^t) = \left(\frac{1}{\|\Sigma_\kappa w_\kappa^t\|} - \|\Sigma_\kappa w_\kappa^t\| \right) \theta_\kappa^\top w_\kappa^t. \tag{12}$$

Setting $g = g_\kappa^t$ and arguing as in Lemma 5,

$$\begin{aligned}
\theta_\kappa^\top w_\kappa^{t+1} &= \theta_\kappa^\top \frac{w_\kappa^t - \eta P_{w_\kappa^t}^\perp(-g)}{\|w_\kappa^t - \eta P_{w_\kappa^t}^\perp(-g)\|} \\
&\ge \theta_\kappa^\top w_\kappa^t + \eta \theta_\kappa^\top P_{w_\kappa^t}^\perp g - \frac{1}{2}(\theta_\kappa^\top w_\kappa^t)\eta^2 \|g\|^2 - \frac{1}{2}\eta^3 |\theta_\kappa^\top P_{w_\kappa^t}^\perp g| \|g\|^2 \\
&\ge \theta_\kappa^\top w_\kappa^t + \eta\lambda_\kappa \theta_\kappa^\top P_{w_\kappa^t}^\perp (Z^t + R^t) - \frac{1}{2}(\theta_\kappa^\top w_\kappa^t)\eta^2 C_1^2 s - \frac{1}{2}\eta^3 C_1^3 s^{\frac{3}{2}} \\
&\quad + \eta\lambda_\kappa \sum_{i=1}^{q} \Big[i\alpha_i \beta_{\kappa,i} (\theta_\kappa^\top \tilde{w}_\kappa^t)^{i-1} (\theta_\kappa^\top P_{w_\kappa^t}^\perp \theta_\kappa) + \sqrt{(i+2)(i+1)} \alpha_i \beta_{\kappa,i+2} (\theta_\kappa^\top \tilde{w}_\kappa^t)^i (\theta_\kappa^\top P_{w_\kappa^t}^\perp \tilde{w}_\kappa^t) \Big] \\
&\ge \theta_\kappa^\top w_\kappa^t + \eta\lambda_\kappa \theta_\kappa^\top P_{w_\kappa^t}^\perp (Z^t + R^t) - \eta^2 C_1^2 (\theta_\kappa^\top w_\kappa^t) s \\
&\quad + \eta\lambda_\kappa \sum_{i=1}^{q} \Big[i\alpha_i \beta_{\kappa,i} \|\Sigma_\kappa w_\kappa^t\|^{-(i-1)} (\theta_\kappa^\top w_\kappa^t)^{i-1} (1 - (\theta_\kappa^\top w_\kappa^t)^2) \\
&\qquad + \sqrt{(i+2)(i+1)} \alpha_i \beta_{\kappa,i+2} \|\Sigma_\kappa w_\kappa^t\|^{-(i+1)} (1 - \|\Sigma_\kappa w_\kappa^t\|^2) (\theta_\kappa^\top w_\kappa^t)^{i+1} \Big]
\end{aligned}$$

where the second inequality uses Lemma 22, and the third uses $\eta \le c_\eta s^{-1}$, $\theta_\kappa^\top w_\kappa^t \ge \frac{1}{2} s^{-1/2}$, and (12). The upper bound follows from

$$\begin{aligned}
\theta_\kappa^\top w_\kappa^{t+1} &\le \theta_\kappa^\top (w_\kappa^t - \eta P_{w_\kappa^t}^\perp(-g)) \\
&= \theta_\kappa^\top w_\kappa^t + \eta\lambda_\kappa \theta_\kappa^\top P_{w_\kappa^t}^\perp (Z^t + R^t) \\
&\quad + \eta\lambda_\kappa \sum_{i=1}^{q} \Big[i\alpha_i \beta_{\kappa,i} \|\Sigma_\kappa w_\kappa^t\|^{-(i-1)} (\theta_\kappa^\top w_\kappa^t)^{i-1} (1 - (\theta_\kappa^\top w_\kappa^t)^2) \\
&\qquad + \sqrt{(i+2)(i+1)} \alpha_i \beta_{\kappa,i+2} \|\Sigma_\kappa w_\kappa^t\|^{-(i+1)} (1 - \|\Sigma_\kappa w_\kappa^t\|^2) (\theta_\kappa^\top w_\kappa^t)^{i+1} \Big].
\end{aligned}$$

Finally, $|\theta_\kappa^\top w_\kappa^{t+1} - \theta_\kappa^\top w_\kappa^t| \le \eta |\theta_\kappa^\top P_{w_\kappa^t}^\perp g| + \frac{1}{2} |\theta_\kappa^\top w_\kappa^t| \eta^2 \|g\|^2 + \frac{1}{2}\eta^3 |\theta_\kappa P_{w_\kappa^t}^\perp g| \|g\|^2 = \tilde{O}(\eta)$, since $\eta \le c_\eta s^{-1}$. □

E.5 Upper Bound on Complement-Subspace Deviation

Lemma 24 (Formal version of Section 3.2)**.** Let $\eta = \eta^t \le c_\eta s^{-1}$. For every $t = 0, 1, 2, \ldots$, if $\theta_\kappa^\top w_\kappa^t \ge 0$ then $\|\Sigma_\kappa^\perp w_\kappa^{t+1}\|^2 - \|\Sigma_\kappa^\perp w_\kappa^t\|^2 \le 7C_1\eta$. Moreover, let $\tau, \tau' > 0$ and suppose $\theta_\kappa^\top w_\kappa^t \ge 0$ for $t = \tau, \tau+1, \ldots, \tau+\tau'-1$. Then there exists a constant $C > 0$ independent of τ, τ' such that, with probability $1-\delta'$, for every $t = 0, 1, \ldots, \tau'$,

$$\|\Sigma_\kappa^\perp w_\kappa^{\tau+t}\|^2 \le \|\Sigma_\kappa^\perp w_\kappa^\tau\|^2 + \sum_{t'=0}^{t-1}\Big[2C_1\eta\|\Sigma_\kappa^\perp w_\kappa^{\tau+t'}\|^4 + 3C_1^2\eta^2\|\Sigma_\kappa^\perp w_\kappa^{\tau+t'}\|^2 + 8C_1\eta\,\mathcal{E}_s\Big] + C\Big[\Big((1-\lambda_\kappa)^{\frac{1}{2}} + \|\Sigma_\kappa^\perp w_\kappa^\tau\|^2 + (C_1\eta\tau')^2\Big)\delta'^{-\frac{1}{2}}\eta\tau'^{\frac{1}{2}} + (1-\lambda_\kappa)\delta'^{-1}\eta^2 s\tau'\Big].$$

Furthermore, defining

$$\begin{aligned} Q^0 &= \|\Sigma_\kappa^\perp w_\kappa^\tau\|^2 \vee C_1\eta + 8C_1\eta\,\mathcal{E}_s\,\tau' \\ &\quad + C\Big[(1-\lambda_\kappa)^{\frac{1}{2}}\delta'^{-\frac{1}{2}}\eta\tau'^{\frac{1}{2}} + (1-\lambda_\kappa)\delta'^{-1}\eta^2 s\tau' + C_1(\|\Sigma_\kappa^\perp w_\kappa^\tau\|^2 + 7C_1\eta\tau')\eta\tau'^{\frac{1}{2}}\Big], \\ Q^t &= Q^{t-1} + 5C_1\eta(Q^{t-1})^2, \end{aligned}$$

we have $\|\Sigma_\kappa^\perp w_\kappa^{\tau+t}\|^2 \le Q^t$ for all $t = 0, 1, \ldots, \tau'$.

Proof. First, note that

$$\Sigma_\kappa^\perp P_{w_\kappa^t}^\perp \theta_\kappa = -(\theta_\kappa^\top w_\kappa^t)\Sigma_\kappa^\perp w_\kappa^t, \quad \Sigma_\kappa^\perp P_{w_\kappa^t}^\perp \tilde{w}_\kappa^t = -(\tilde{w}_\kappa^{t\top} w_\kappa^t)\Sigma_\kappa^\perp w_\kappa^t = -\|\Sigma_\kappa w_\kappa^t\|\Sigma_\kappa^\perp w_\kappa^t. \quad (13)$$

Set $g = g_\kappa^t$ and define

$$\xi^t = 1 - \eta\lambda_\kappa\sum_{i=1}^q\Big(i\bar{\alpha}_i\beta_{\kappa,i} + \sqrt{(i+2)(i+1)}\bar{\alpha}_i\beta_{\kappa,i+2}\Big)(\theta_\kappa^\top\tilde{w}_\kappa^t)^{i-1}(\theta_\kappa^\top w_\kappa^t).$$

When $\theta_\kappa^\top w_\kappa^t \ge 0$, we have $0 \le \xi^t \le 1$. By Lemma 22 and (13),

$$\Sigma_\kappa^\perp P_{w_\kappa^t}^\perp g = -\sum_{i=1}^q\lambda_\kappa\Big(i\bar{\alpha}_i\beta_{\kappa,i} + \sqrt{(i+2)(i+1)}\bar{\alpha}_i\beta_{\kappa,i+2}\Big)(\theta_\kappa^\top\tilde{w}_\kappa^t)^{i-1}(\theta_\kappa^\top w_\kappa^t)\Sigma_\kappa^\perp w_\kappa^t + \Sigma_\kappa^\perp P_{w_\kappa^t}^\perp(Z^t + R^t).$$

Hence

$$\begin{aligned} \|\Sigma_\kappa^\perp w_\kappa^{t+1}\|^2 &= \frac{(w_\kappa^t + \eta P_{w_\kappa^t}^\perp g)^\top\Sigma_\kappa^\perp(w_\kappa^t + \eta P_{w_\kappa^t}^\perp g)}{\|w_\kappa^t + \eta P_{w_\kappa^t}^\perp g\|^2} \\ &\le \|\xi^t\Sigma_\kappa^\perp w_\kappa^t + \eta\Sigma_\kappa^\perp P_{w_\kappa^t}^\perp(Z^t + R^t)\|^2 \\ &= (\xi^t)^2\|\Sigma_\kappa^\perp w_\kappa^t\|^2 + 2\eta\xi^t w_\kappa^{t\top}\Sigma_\kappa^\perp P_{w_\kappa^t}^\perp(Z^t + R^t) + \eta^2\|\Sigma_\kappa^\perp P_{w_\kappa^t}^\perp(Z^t + R^t)\|^2 \\ &\le (\xi^t)^2\|\Sigma_\kappa^\perp w_\kappa^t\|^2 + 2\eta\xi^t(w_\kappa^{t\top}\Sigma_\kappa^\perp Z^t - \|\Sigma_\kappa^\perp w_\kappa^t\|^2 w_\kappa^{t\top}Z^t + w_\kappa^t\Sigma_\kappa^\perp R^t - \|\Sigma_\kappa^\perp w_\kappa^t\|^2 w_\kappa^t R^t) \\ &\quad + 2\eta^2\left[\|\Sigma_\kappa^\perp Z^t\|^2 + \|\Sigma_\kappa^\perp R^t\|^2 + \|\Sigma_\kappa^\perp w_\kappa^t\|^2\left((w_\kappa^{t\top}Z^t)^2 + (w_\kappa^{t\top}R^t)^2\right)\right] \\ &\le (\xi^t)^2\|\Sigma_\kappa^\perp w_\kappa^t\|^2 + 2C_1\eta\xi^t\|\Sigma_\kappa^\perp w_\kappa^t\|^4 + 3C_1^2\eta^2\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \\ &\quad + 4C_1\eta\xi^t\mathcal{E}_s\|\Sigma_\kappa^\perp w_\kappa^t\| + 4C_1^2\eta^2\mathcal{E}_s^2 s \\ &\quad + 2\left(\eta\xi^t w_\kappa^{t\top}\Sigma_\kappa^\perp Z^t - \eta\xi^t\|\Sigma_\kappa^\perp w_\kappa^t\|^2 w_\kappa^{t\top}Z^t + \eta^2\|\Sigma_\kappa^\perp Z^t\|^2\right) \\ &\le \|\Sigma_\kappa^\perp w_\kappa^t\|^2 + 2C_1\eta\|\Sigma_\kappa^\perp w_\kappa^t\|^4 + 3C_1^2\eta^2\|\Sigma_\kappa^\perp w_\kappa^t\|^2 + 8C_1\eta\mathcal{E}_s \\ &\quad + 2\left(\eta\xi^t w_\kappa^{t\top}\Sigma_\kappa^\perp Z^t - \eta\xi^t\|\Sigma_\kappa^\perp w_\kappa^t\|^2 w_\kappa^{t\top}Z^t + \eta^2\|\Sigma_\kappa^\perp Z^t\|^2\right). \end{aligned} \quad (14)$$

Here the second inequality uses $w_\kappa^{t\top}\Sigma_\kappa^\perp P_{w_\kappa^t}^\perp Z^t = w_\kappa^t\Sigma_\kappa^\perp Z^t - \|\Sigma_\kappa^\perp w_\kappa^t\|^2 w_\kappa^t Z^t$ and similarly for R^t, together with

$$\|\Sigma_\kappa^\perp P_{w_\kappa^t}^\perp(Z^t + R^t)\|^2 \le 2(\|\Sigma_\kappa^\perp Z^t\|^2 + \|\Sigma_\kappa^\perp R^t\|^2) + 2\|\Sigma_\kappa^\perp w_\kappa^t\|^2((w_\kappa^{t\top}Z^t)^2 + (w_\kappa^{t\top}R^t)^2).$$

The third inequality uses $w_\kappa^{t\top}\Sigma_\kappa^\perp R^t \le C_1\,\mathcal{E}_s\|\Sigma_\kappa^\perp w_\kappa^t\|$, $w_\kappa^{t\top}R^t \le C_1(\mathcal{E}_s \vee \|\Sigma_\kappa^\perp w_\kappa^t\|^2)$, $\|\Sigma_\kappa^\perp R^t\|^2 \le C_1^2\mathcal{E}_s^2 s$, $(w_\kappa^{t\top}R^t)^2 \le C_1^2(\mathcal{E}_s^2 \vee \|\Sigma_\kappa^\perp w^t\|^4)$, and $(w_\kappa^{t\top}Z^t)^2 \le C_1^2$. The fourth uses $\theta_\kappa^\top w_\kappa^t \ge 0$. The dominant fluctuation terms in (14) are $2C_1\eta\|\Sigma_\kappa^\perp w_\kappa^t\|^4$, $2\eta\xi^t w_\kappa^{t\top}\Sigma_\kappa^\perp Z^t$, and $-2\eta\xi^t\|\Sigma_\kappa^\perp w_\kappa^t\|^2 w_\kappa^{t\top}Z^t$, from which $\|\Sigma_\kappa^\perp w_\kappa^{t+1}\|^2 - \|\Sigma_\kappa^\perp w_\kappa^t\|^2 \le 7C_1\eta$.

Summing (14) over $t' = 0,\dots,t-1$ gives the stated bound, where the stochastic terms are controlled via Doob's maximal inequality for the components in $V_\kappa^\perp$, exploiting the low occurrence probability $1-\lambda_\kappa$, and standard concentration inequalities for the components in V_κ, as detailed below. The sequences $\{\sum_{t'=0}^{t-1}\xi^{\tau+t'}w_\kappa^{(\tau+t')\top}\Sigma_\kappa^\perp Z^{\tau+t'}\}$ and $\{\sum_{t'=0}^{t-1}\xi^{\tau+t'}\|\Sigma_\kappa^\perp w_\kappa^{\tau+t'}\|^2 w_\kappa^{(\tau+t')\top}Z^{\tau+t'}\}$ are martingales, and $\{\sum_{t'=0}^{t-1}\|\Sigma_\kappa^\perp Z^{\tau+t'}\|^2\}$ is a submartingale satisfying

$$\mathbb{E}[(\xi^t w_\kappa^{t\top}\Sigma_\kappa^\perp Z^t)^2] \lesssim (1-\lambda_\kappa)^2,\quad \mathbb{E}[(\xi^t\|\Sigma_\kappa^\perp w_\kappa^t\|^2 w_\kappa^{t\top}Z^t)^2] \lesssim \|\Sigma_\kappa^\perp w_\kappa^t\|^4,\quad \mathbb{E}[\|\Sigma_\kappa^\perp Z^t\|^2] \lesssim (1-\lambda_\kappa)s,$$

Markov's and Doob's inequalities give, with probability $1-\delta'$,

$$\max_{0\le t\le\tau'}\left|\sum_{t'=0}^{t-1}\xi^{\tau+t'}w_\kappa^{(\tau+t')\top}\Sigma_\kappa^\perp Z^{\tau+t'}\right|^2 \lesssim \delta'^{-1}(1-\lambda_\kappa)\tau',$$

$$\max_{0\le t\le\tau'}\sum_{t'=0}^{t-1}\|\Sigma_\kappa^\perp Z^{\tau+t'}\|^2 \lesssim \delta'^{-1}(1-\lambda_\kappa)s\tau'.$$

Furthermore,

$$\sum_{t'=0}^{t-1}\xi^{\tau+t'}\|\Sigma_\kappa^\perp w_\kappa^{\tau+t'}\|^2 w_\kappa^{(\tau+t')\top}Z^{\tau+t'} \le C_1(\|\Sigma_\kappa^\perp w_\kappa^\tau\|^2 + 7C_1\eta\tau')\sqrt{\tau'}.$$

Thus, for $t = 1,2,\dots\tau'$,

$$\begin{aligned}&2\eta\sum_{t'=0}^{t-1}\left(\xi^{\tau+t'}w_\kappa^{(\tau+t')\top}\Sigma_\kappa^\perp Z^{\tau+t'} - \xi^{\tau+t'}\|\Sigma_\kappa^\perp w_\kappa^{\tau+t'}\|^2 w_\kappa^{(\tau+t')\top}Z^{\tau+t'} + \eta\|\Sigma_\kappa^\perp Z^{\tau+t'}\|^2\right)\\ &\lesssim (1-\lambda_\kappa)^{\frac12}\delta'^{-\frac12}\eta\tau'^{\frac12} + (1-\lambda_\kappa)\delta'^{-1}\eta^2 s\tau' + C_1(\|\Sigma_\kappa^\perp w_\kappa^\tau\|^2 + 7C_1\eta\tau')\eta\tau'^{\frac12}.\end{aligned}$$

Combining these estimates yields the bound for $\|\Sigma_\kappa^\perp w_\kappa^{\tau+t}\|^2$. The sequence Q^t is then constructed so that $Q^t \ge C_1\eta$ and $Q^t \ge \|\Sigma_\kappa^\perp w_\kappa^{\tau+t}\|^2$ follows by induction. □

E.6 Weak Alignment

Lemma 25. Let $\eta^t = \eta \le c_\eta s^{-3/2}$, $\mathcal{E}_s \le c_\varepsilon s^{-1/2}$, and $\tau = \lfloor\frac12 s^{-1/2}C_1^{-1}\eta^{-1}\rfloor$. Suppose $\theta_\kappa^\top w_\kappa^0 \ge s^{-1/2}$, $\|\Sigma_\kappa^\perp w_\kappa^0\|^2 \le c_r\theta_\kappa^\top w_\kappa^0$, and $\theta_\kappa^\top w_\kappa^t \le c_1$ for all $t = 0,1,\dots,\tau$. Then, with probability $1-\eta\tau\delta$,

$$\theta_\kappa^\top w_\kappa^\tau \ge \left(1+\tfrac12\bar\alpha_2\beta_{\kappa,2}\eta\tau\right)\theta_\kappa^\top w_\kappa^0,$$

and furthermore $\|\Sigma_\kappa^\perp w_\kappa^\tau\|^2 \le c_r\theta_\kappa^\top w_\kappa^\tau$.

Proof. By Lemma 23, if $\theta_\kappa^\top w_\kappa^t \ge \frac12 s^{-1/2}$ then $|\theta_\kappa^\top w_\kappa^{t+1} - \theta_\kappa^\top w_\kappa^t| \le C_1\eta$ with high probability, so by induction $\theta_\kappa^\top w_\kappa^t \ge \frac12 s^{-1/2}$ for all $t = 0,1,\dots,\tau$. Applying Lemma 24 with $\tau \leftarrow 0$, $\tau' \leftarrow \tau$, $\delta' \leftarrow \eta\tau\delta$, define

$$Q^0 = \|\Sigma_\kappa^\perp w_\kappa^0\|^2 \vee C_1\eta + 8C_1\eta\,\mathcal{E}_s\,\tau + C\Big[(1-\lambda_\kappa)^{1/2}\delta^{-1/2}\eta^{1/2} + (1-\lambda_\kappa)\delta^{-1}\eta s + C_1(\|\Sigma_\kappa^\perp w_\kappa^0\|^2 + 7C_1\eta\tau)\eta\tau^{1/2}\Big],$$

$$Q^t = Q^{t-1} + 5C_1\eta(Q^{t-1})^2.$$

With probability $1-\eta\tau\delta$, $\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \le Q^t$ for all $t = 0,1,\dots,\tau$. By the Bihari–LaSalle inequality, $Q^t \le \frac{Q^0}{1-5C_1\eta Q^0 t}$, and in particular

$$Q^\tau \le (1+6C_1\eta\tau Q^0)Q^0$$

$$
\begin{aligned}
&\le (1 + 6C_1\eta\tau(\|\Sigma_\kappa^\perp w_\kappa^0\|^2 \vee C_1\eta + c_3 s^{-1}))(\|\Sigma_\kappa^\perp w_\kappa^0\|^2 \vee C_1\eta + c_3 s^{-1}) \\
&\le (1 + 7C_1 c_1 c_r \eta\tau)(\|\Sigma_\kappa^\perp w_\kappa^0\|^2 \vee C_1\eta + c_3 s^{-1}) \\
&\le (1 + \frac{1}{2}\alpha_2\beta_{\kappa,2}\eta\tau)(\|\Sigma_\kappa^\perp w_\kappa^0\|^2 \vee c_r s^{-\frac{1}{2}}) = (1 + \frac{1}{2}\alpha_2\beta_{\kappa,2}\eta\tau) c_r(\theta_\kappa^\top w_\kappa^0). \qquad (15)
\end{aligned}
$$

where the second inequality uses $\eta \le c_\eta s^{-3/2}$ and $\mathcal{E}_s \le c_\varepsilon s^{-1/2}$ to bound

$$
\begin{aligned}
&8C_1\eta\mathcal{E}_s\tau + C\left[(1-\lambda_\kappa)^{\frac{1}{2}}\delta^{-\frac{1}{2}}\eta^{\frac{1}{2}} + (1-\lambda_\kappa)\delta^{-1}\eta s + C_1(\|\Sigma_\kappa^\perp w_\kappa^\tau\|^2 + 7C_1\eta\tau)\eta\tau^{\frac{1}{2}}\right] \\
&\le \left(4c_\varepsilon + C\left[c_\varepsilon^{\frac{1}{2}}\delta^{-\frac{1}{2}}c_\eta^{\frac{1}{2}} + c_\varepsilon\delta^{-1}c_\eta + 4C_1^{\frac{1}{2}}c_\eta^{\frac{1}{2}}\right]\right) s^{-1} \\
&\le c_3 s^{-1}
\end{aligned}
$$

the third uses $\|\Sigma_\kappa^\perp w_\kappa^0\|^2 \vee C_1\eta + 5C_1 s^{-1} \le \frac{7}{6} c_r(\theta_\kappa^\top w_\kappa^0) \le \frac{7}{6} c_r c_1$, and the fourth uses $\eta\tau c_r s^{-1/2} = \frac{1}{2} C_1^{-1} c_r s^{-1} \gtrsim c_3 s^{-1}$.

By Lemma 23,

$$
\begin{aligned}
\theta_\kappa^\top w_\kappa^{t+1} &\ge \theta_\kappa^\top w_\kappa^t + \eta\lambda_\kappa \theta_\kappa^\top P_{w_\kappa^t}^\perp (Z^t + R^t) - \eta^2 C_1^2 (\theta_\kappa^\top w_\kappa^t) s \\
&\quad + \eta\lambda_\kappa \sum_{i=1}^{q} \Big[i\alpha_i\beta_{\kappa,i}\|\Sigma_\kappa w_\kappa^t\|^{-(i-1)}(\theta_\kappa^\top w_\kappa^t)^{i-1}(1 - (\theta_\kappa^\top w_\kappa^t)^2) \\
&\qquad + \sqrt{(i+2)(i+1)}\alpha_i\beta_{\kappa,i+2}\|\Sigma_\kappa w_\kappa^t\|^{-(i+1)}\|\Sigma_\kappa^\perp w_\kappa^t\|^2(\theta_\kappa^\top w_\kappa^t)^{i+1}\Big] \\
&\ge \theta_\kappa^\top w_\kappa^t + 2\eta\lambda_\kappa\alpha_2\beta_{\kappa,2}(\theta_\kappa^\top w_\kappa^t)(1-c_1) + \eta\lambda_\kappa\theta_\kappa^\top P_{w_\kappa^t}^\perp (Z^t + R^t).
\end{aligned}
$$

where we used $\eta \le c_\eta s^{-3/2}$, $\|\Sigma_\kappa w_\kappa^t\|^2 = \Theta(1)$, and $\theta_\kappa^\top w_\kappa^t$, $\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \le c_1$ for the second inequality. Thus,

$$
\begin{aligned}
\theta_\kappa^\top w_\kappa^\tau &\ge \theta_\kappa^\top w_\kappa^0 + \eta\lambda_\kappa \sum_{t=0}^{\tau-1} \Big[2(1-c_1)\alpha_2\beta_{\kappa,2}(\theta_\kappa^\top w_\kappa^t) + \theta_\kappa^\top P_{w_\kappa^t}^\perp (Z^t + R^t)\Big] \\
&\ge \theta_\kappa^\top w_\kappa^0 + 2(1-c_1)\eta\lambda_\kappa\alpha_2\beta_{\kappa,2}\tau(\theta_\kappa^\top w_\kappa^0 - \frac{1}{2}C_1\eta\tau) + \eta\lambda_\kappa \sum_{t=0}^{\tau-1}\Big[\theta_\kappa^\top P_{w_\kappa^t}^\perp (Z^t + R^t)\Big] \\
&\ge \theta_\kappa^\top w_\kappa^0 + \alpha_2\beta_{\kappa,2}\eta\tau(\theta_\kappa^\top w_\kappa^0) - C_1\eta\left(\tau^{\frac{1}{2}} + (\mathcal{E}_s \vee Q^\tau)\tau\right) \\
&\ge \theta_\kappa^\top w_\kappa^0 + \alpha_2\beta_{\kappa,2}\eta\tau\theta_\kappa^\top w_\kappa^0 - \eta\tau(\sqrt{2}c_\eta^{\frac{1}{2}}C_1^{\frac{3}{2}}s^{-\frac{1}{2}} + C_1 c_\varepsilon s^{-\frac{1}{2}} + 2C_1 c_r(\theta_\kappa^\top w_\kappa^0)) \\
&\ge (1 + \frac{1}{2}\alpha_2\beta_{\kappa,2}\eta\tau)\theta_\kappa^\top w_\kappa^0. \qquad (16)
\end{aligned}
$$

holds with high probability. The second inequality uses $|\theta_\kappa^\top w_\kappa^{t+1} - \theta_\kappa^\top w_\kappa^t| \le C_1\eta$ with high probability. The third uses $C_1\eta\tau \le \frac{1}{2}\theta_\kappa^\top w_\kappa^0$ and $(1-c_1)\lambda_\kappa \ge \frac{2}{3}$. The fourth uses $Q^\tau \le 2c_r(\theta_\kappa^\top w_\kappa^0)$, which follows from (15), together with $\mathcal{E}_s \le c_\varepsilon s^{-1/2}$. From (15) and (16), $\|\Sigma_\kappa^\perp w_\kappa^\tau\|^2 \le Q^\tau \le c_r\theta_\kappa^\top w_\kappa^\tau$. □

Lemma 26. Let $\eta^t = \eta \le c_\eta s^{-3/2}$, $\mathcal{E}_s \le c_\varepsilon s^{-1/2}$, $\theta_\kappa^\top w_\kappa^0 \ge s^{-1/2}$, and $\|\Sigma_\kappa^\perp w_\kappa^0\|^2 \le c_r s^{-1/2}$. With probability $1 - \tilde{\delta}$, there exists $t_1 \le T_{1,1} = \Theta(\eta^{-1}\log s)$ such that $\theta_\kappa^\top w_\kappa^{t_1} > c_1$. Furthermore, with probability $1 - \tilde{\delta} - \delta$, $\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \le 3c_r s^{-1/2}$ for all $t = 0, 1, \ldots, t_1 - 1$.

Proof. Suppose $\theta_\kappa^\top w_\kappa^t \le c_1$ for all $t = 0, 1, \ldots, T_{1,1}$. Set $\tau = \lfloor \frac{1}{2} s^{-1/2} C_1^{-1}\eta^{-1} \rfloor$ and $T_{1,1} = \lceil \frac{3\log s}{2\bar{\alpha}_2\beta_{\kappa,2}\eta\tau} \rceil \tau$. Applying Lemma 25 repeatedly with probability $1 - \frac{3\log s}{2\bar{\alpha}_2\beta_{\kappa,2}}\delta$,

$$
\log \frac{\theta_\kappa^\top w_\kappa^{T_{1,1}}}{\theta_\kappa^\top w_\kappa^0} \ge \left\lceil \frac{3\log s}{2\bar{\alpha}_2\beta_{\kappa,2}\eta\tau} \right\rceil \log\big(1 + \tfrac{1}{2}\bar{\alpha}_2\beta_{\kappa,2}\eta\tau\big) \ge \frac{3\log s}{2\bar{\alpha}_2\beta_{\kappa,2}\eta\tau} \cdot \frac{\bar{\alpha}_2\beta_{\kappa,2}\eta\tau}{3} \ge \log s^{1/2}.
$$

Hence $\theta_\kappa^\top w_\kappa^{T_{1,1}} \ge 1$, contradicting $\theta_\kappa^\top w_\kappa^{T_{1,1}} \le c_1$. Let t_1 be the first time $\theta_\kappa^\top w_\kappa^t > c_1$. For $t = 0, 1, \ldots, t_1 - 1$, applying Lemma 24 gives $\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \le Q^t$ with probability $1 - \delta$, where $Q^0 \le \|\Sigma_\kappa^\perp w_\kappa^0\|^2 + \frac{1}{2}c_3 s^{-1/2}$. By the Bihari–LaSalle inequality, $\|\Sigma_\kappa^\perp w_\kappa^{t_1 - 1}\|^2 \le Q^{t_1 - 1} \le 2Q^0 \le 3c_r s^{-1/2}$. □

E.7 Amplification of Alignment

We reset the time index at t_1 (when weak alignment is achieved) and write $t \leftarrow t - t_1$.

Lemma 27. Let $\eta^t = \eta \le c_\eta s^{-3/2}$, $\mathcal{E}_s \le c_\varepsilon s^{-1/2}$, and $\tau = \lfloor \frac{1}{2} c_1 C_1^{-1} \eta^{-1} \rfloor$. Suppose $\theta_\kappa^\top w_\kappa^0 \ge c_1$, $\|\Sigma_\kappa^\perp w_\kappa^0\|^2 \le c_r \theta_\kappa^\top w_\kappa^0$, and $\theta_\kappa^\top w_\kappa^t \le 1 - c_1$ for all $t = 0, 1, \dots, \tau$. Then, with probability $1 - \eta\tau\delta$,

$$\theta_\kappa^\top w_\kappa^\tau \ge \big(1 + \tfrac{1}{2} c_1 \bar{\alpha}_2 \beta_{\kappa,2} \eta\tau\big)\, \theta_\kappa^\top w_\kappa^0,$$

and $\|\Sigma_\kappa^\perp w_\kappa^\tau\|^2 \le c_r \theta_\kappa^\top w_\kappa^\tau$.

Proof. The argument parallels Lemma 25. For $t = 0, 1, \dots, \tau$, we have $\theta_\kappa^\top w_\kappa^t \ge \frac{1}{2} c_1 \ge \frac{1}{2} s^{-1/2}$. Applying Lemma 24 with $\tau \leftarrow 0$, $\tau' \leftarrow \tau$, $\delta' \leftarrow \eta\tau\delta$ defines Q^0 and Q^t; the dominant term of Q^0 is $\|\Sigma_\kappa^\perp w_\kappa^0\|^2 \vee C_1\eta + 8C_1\eta\,\mathcal{E}_s\,\tau$, giving $Q^0 \le \|\Sigma_\kappa^\perp w_\kappa^0\|^2 \vee C_1\eta + s^{-1/2}$. By the Bihari–LaSalle inequality Lemma 38,

$$\begin{aligned}
Q^\tau &\le (1 + 6C_1\eta\tau Q^0) Q^0 \\
&\le (1 + 7c_r C_1 \eta\tau)(\|\Sigma_\kappa^\perp w_\kappa^0\|^2 \vee C_1\eta + s^{-1/2}) \\
&\le (1 + \tfrac{1}{2} c_1 \bar{\alpha}_2 \beta_{\kappa,2} \eta\tau)(\|\Sigma_\kappa^\perp w_\kappa^0\|^2 \vee c_r c_1) \le (1 + \tfrac{1}{2} c_1 \bar{\alpha}_2 \beta_{\kappa,2} \eta\tau) c_r (\theta_\kappa^\top w_\kappa^0).
\end{aligned}$$

By Lemma 23,

$$\begin{aligned}
\theta_\kappa^\top w_\kappa^{t+1} &\ge \theta_\kappa^\top w_\kappa^t + \eta\lambda_\kappa \theta_\kappa^\top P_{w_\kappa^t}^\perp (Z^t + R^t) - \eta^2 C_1^2 (\theta_\kappa^\top w_\kappa^t) s \\
&\quad + \eta\lambda_\kappa \sum_{i=1}^{q} \Big[i\alpha_i \beta_{\kappa,i} \|\Sigma_\kappa w_\kappa^t\|^{-(i-1)} (\theta_\kappa^\top w_\kappa^t)^{i-1} (1 - (\theta_\kappa^\top w_\kappa^t)^2) \\
&\qquad + \sqrt{(i+2)(i+1)}\, \alpha_i \beta_{\kappa,i+2} \|\Sigma_\kappa w_\kappa^t\|^{-(i+1)} \|\Sigma_\kappa^\perp w_\kappa^t\|^2 (\theta_\kappa^\top w_\kappa^t)^{i+1} \Big] \\
&\ge \theta_\kappa^\top w_\kappa^t + c_1 \eta \alpha_2 \beta_{\kappa,2} (\theta_\kappa^\top w_\kappa^t) + \eta\lambda_\kappa \theta_\kappa^\top P_{w_\kappa^t}^\perp (Z^t + R^t).
\end{aligned}$$

where we used $1 - (\theta_\kappa^\top w_\kappa^t)^2 \ge 2c_1 - c_1^2$, $\eta \le c_\eta s^{-3/2}$, and $\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \le Q^\tau \le 3c_r \theta_\kappa^\top w_\kappa^t$ for the second inequality. Then we have

$$\begin{aligned}
\theta_\kappa^\top w_\kappa^\tau &\ge \theta_\kappa^\top w_\kappa^0 + c_1 \alpha_2 \beta_{\kappa,2} \eta\tau (\theta_\kappa^\top w_\kappa^0 - \frac{1}{2} C_1 \eta\tau) + \eta\lambda_\kappa \sum_{t=0}^{\tau-1} \Big[\theta_\kappa^\top P_{w_\kappa^t}^\perp (Z^t + R^t) \Big] \\
&\ge \theta_\kappa^\top w_\kappa^0 + c_1 \alpha_2 \beta_{\kappa,2} \eta\tau (\theta_\kappa^\top w_\kappa^0 - \frac{1}{2} C_1 \eta\tau) - C_1 \eta \left(\tau^{\frac{1}{2}} + (\mathcal{E}_s \vee Q^\tau)\tau \right) \\
&\ge (1 + \frac{1}{2} c_1 \alpha_2 \beta_{\kappa,2} \eta\tau) \theta_\kappa^\top w_\kappa^0.
\end{aligned}$$

with high probability. Hence $\|\Sigma_\kappa^\perp w_\kappa^\tau\|^2 \le Q^\tau \le c_r \theta_\kappa^\top w_\kappa^\tau$. □

Lemma 28. Let $\eta^t = \eta \le c_\eta s^{-3/2}$, $\mathcal{E}_s \le c_\varepsilon s^{-1/2}$, $\theta_\kappa^\top w_\kappa^0 \ge c_1$, and $\|\Sigma_\kappa^\perp w_\kappa^0\|^2 \le c_r c_1$. With probability $1 - \tilde{\delta}$, there exists $t_2 \le T_{1,2} = \tilde{\Theta}(\eta^{-1})$ such that $\theta_\kappa^\top w_\kappa^{t_2} > 1 - c_1$ and $\|\Sigma_\kappa^\perp w_\kappa^{t_2}\|^2 \le 2c_r$.

Proof. Suppose $\theta_\kappa^\top w_\kappa^t \le 1 - c_1$ for all $t = 0, 1, \dots, T_{1,2}$. Set $\tau = \lfloor \frac{1}{2} c_1 C_1^{-1} \eta^{-1} \rfloor$ and $T_{1,2} = \lceil \frac{3 \log s}{2 c_1 \bar{\alpha}_2 \beta_{\kappa,2} \eta\tau} \rceil \tau$. Applying Lemma 27 repeatedly with probability $1 - \frac{3 \log s}{2\bar{\alpha}_2 \beta_{\kappa,2}} c_1^{-1} \delta$,

$$\log \frac{\theta_\kappa^\top w_\kappa^{T_{1,2}}}{\theta_\kappa^\top w_\kappa^0} \ge \left\lceil \frac{3 \log s}{2 c_1 \bar{\alpha}_2 \beta_{\kappa,2} \eta\tau} \right\rceil \log\big(1 + \tfrac{1}{2} c_1 \bar{\alpha}_2 \beta_{\kappa,2} \eta\tau\big) \ge \log s^{1/2},$$

so $\theta_\kappa^\top w_\kappa^{T_{1,2}} \ge 1$, contradicting $\theta_\kappa^\top w_\kappa^{T_{1,2}} \le 1 - c_1$. Let t_2 be the first time $\theta_\kappa^\top w_\kappa^t > 1 - c_1$, and set $t_2' = \lfloor t_2/\tau \rfloor \tau$. At time t_2', $\|\Sigma_\kappa^\perp w_\kappa^{t_2'}\|^2 \le c_r \theta_\kappa^\top w_\kappa^{t_2'} \le c_r$. By the same argument as in Lemma 27, $\|\Sigma_\kappa^\perp w_\kappa^{t_2}\|^2 \le (1 + \frac{1}{2} c_1 \bar{\alpha}_2 \beta_{\kappa,2} \eta\tau) c_r (\theta_\kappa^\top w_\kappa^{t_2'}) \le 2c_r$. □

E.8 Strong Alignment

Lemma 29. Let $\bar{\varepsilon} \geq \mathcal{E}_s$, $\eta = \eta^t \leq c_\eta \bar{\varepsilon} s^{-1} \wedge c_\eta \bar{\varepsilon}^2$, and $\|\Sigma_\kappa^\perp w_\kappa^0\|^2 \leq 3c_r$. Let $\tau > 0$ and suppose $\theta_\kappa^\top w_\kappa^t \geq \frac{1}{2}$ and $\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \geq \bar{\varepsilon}$ for all $t = 0, 1, \ldots, \tau - 1$. Then for every $t = 0, 1, \ldots, \tau$,

$$\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \leq \left(1 - \tfrac{1}{5} c_w \eta\right)^t \|\Sigma_\kappa^\perp w_\kappa^0\|^2 + c_r \bar{\varepsilon}$$

with high probability.

Proof. Setting ξ^t as in the proof of Lemma 24, for $\theta_\kappa^\top w_\kappa^t \geq 0$ we have by the same computation as in (14),

$$\begin{aligned}\|\Sigma_\kappa^\perp w_\kappa^{t+1}\|^2 \leq (\xi^t)^2 \|\Sigma_\kappa^\perp w_\kappa^t\|^2 + 2C_1 \eta \xi^t \|\Sigma_\kappa^\perp w_\kappa^t\|^4 + 3C_1^2 \eta^2 \|\Sigma_\kappa^\perp w_\kappa^t\|^2 \\ + 4C_1 \eta \xi^t \, \mathcal{E}_s \|\Sigma_\kappa^\perp w_\kappa^t\| + 4C_1^2 \eta^2 \mathcal{E}_s^2 s \\ + 2\left(\eta \xi^t w_\kappa^{t\top} \Sigma_\kappa^\perp Z^t - \eta \xi^t \|\Sigma_\kappa^\perp w_\kappa^t\|^2 w_\kappa^{t\top} Z^t + \eta^2 \|\Sigma_\kappa^\perp Z^t\|^2\right).\end{aligned}$$

Fix $t \leq \tau$ and assume $\|\Sigma_\kappa^\perp w_\kappa^{t'}\|^2 \leq 4c_r$ for all $t' \leq t - 1$. Using $(\xi^{t'})^2 \leq \xi^{t'}$, $\theta_\kappa^\top w_\kappa^{t'} \geq \frac{1}{2}$, and $\|\Sigma_\kappa^\perp w_\kappa^{t'}\| \geq \bar{\varepsilon}$, and bounding $4C_1 \eta \xi^{t'} \mathcal{E}_s \|\Sigma_\kappa^\perp w_\kappa^{t'}\| \leq 8c_r C_1 \bar{\varepsilon} \eta$, $4C_1^2 \eta^2 \mathcal{E}_s^2 s \leq 4C_1^2 c_\eta \bar{\varepsilon}^3 \eta$, and $\eta \|\Sigma_\kappa^\perp Z^{t'}\|^2 \leq c_\eta C_1 \bar{\varepsilon}$,

$$\begin{aligned}\|\Sigma_\kappa^\perp w_\kappa^{t'+1}\|^2 &\leq \|\Sigma_\kappa^\perp w_\kappa^{t'}\|^2 - \frac{1}{4}\lambda_\kappa c_w \eta \|\Sigma_\kappa^\perp w_\kappa^{t'}\|^2 + 2C_1 \eta \xi^{t'} \|\Sigma_\kappa^\perp w_\kappa^{t'}\|^4 + 3C_1^2 \eta^2 \|\Sigma_\kappa^\perp w_\kappa^{t'}\|^2 \\ &\quad + 4C_1 \eta \xi^{t'} \mathcal{E}_s \|\Sigma_\kappa^\perp w_\kappa^{t'}\| + 4C_1^2 \eta^2 \mathcal{E}_s^2 s \\ &\quad + 2\left(\eta \xi^{t'} w_\kappa^{t'\top} \Sigma_\kappa^\perp Z^{t'} - \eta \xi^{t'} \|\Sigma_\kappa^\perp w_\kappa^{t'}\|^2 w_\kappa^{t'\top} Z^{t'} + \eta^2 \|\Sigma_\kappa^\perp Z^{t'}\|^2\right) \\ &\leq (1 - \frac{1}{5} c_w \eta) \|\Sigma_\kappa^\perp w_\kappa^{t'}\|^2 + 2\left(\eta \xi^{t'} w_\kappa^{t'\top} \Sigma_\kappa^\perp Z^{t'} - \eta \xi^{t'} \|\Sigma_\kappa^\perp w_\kappa^{t'}\|^2 w_\kappa^{t'\top} Z^{t'}\right).\end{aligned}$$

Unrolling this recursion,

$$\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \leq (1 - \tfrac{1}{5} c_w \eta)^t \|\Sigma_\kappa^\perp w_\kappa^0\|^2 + 2\eta \sum_{t'=0}^{t-1} (1 - \tfrac{1}{5} c_w \eta)^{t-t'-1} \left(\xi^{t'} w_\kappa^{t'\top} \Sigma_\kappa^\perp Z^{t'} - \xi^{t'} \|\Sigma_\kappa^\perp w_\kappa^{t'}\|^2 w_\kappa^{t'\top} Z^{t'}\right).$$

The conditional variance of the martingale increment at step t' is at most $(1 - \frac{1}{5} c_w \eta)^{2(t-t'-1)} C_1^2$. Summing over $t' = 0, \ldots, t - 1$ and bounding the resulting geometric series gives

$$\sum_{t'=0}^{t-1} \left(1 - \tfrac{1}{5} c_w \eta\right)^{2(t-t'-1)} C_1^2 \leq \frac{C_1^2}{1 - \left(1 - \frac{1}{5} c_w \eta\right)^2} \leq \frac{5C_1^2}{c_w \eta}.$$

Hence each weighted sum is at most $\sqrt{5C_1^2/(c_w \eta)}$ with high probability, and since $\eta \leq c_\eta \bar{\varepsilon}^2$,

$$\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \leq (1 - \tfrac{1}{5} c_w \eta)^t \|\Sigma_\kappa^\perp w_\kappa^0\|^2 + 4\sqrt{5} C_1 c_w^{-1/2} \sqrt{\eta} \leq (1 - \tfrac{1}{5} c_w \eta)^t \|\Sigma_\kappa^\perp w_\kappa^0\|^2 + c_r \bar{\varepsilon}.$$

This gives $\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \leq 4c_r$, completing the induction. □

Lemma 30. Let $\eta \leq c_\eta s^{-1}$, $\mathcal{E}_s \leq c_\varepsilon s^{-1/2}$, $\theta_\kappa^\top w_\kappa^0 \geq 1 - 3c_1$, and suppose $\theta_\kappa^\top w_\kappa^t \leq 1 - c_2$ and $\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \leq 4c_r$ for all $t = 0, 1, \ldots, \tau$. Then, with high probability,

$$\theta_\kappa^\top w_\kappa^t \geq \theta_\kappa^\top w_\kappa^0 - C_2 \eta + \bar{\alpha}_2 \beta_{\kappa,2} c_2 \eta t$$

Proof. Fix $t \leq \tau$ and assume $\theta_\kappa^\top w_\kappa^{t'} \geq 1 - 4c_1$ for all $t' \leq t - 1$. By Lemma 23, using $\eta \leq c_\eta s^{-1}$, $1 - (\theta_\kappa^\top w_\kappa^{t'})^2 \geq c_2$, $\|\Sigma_\kappa^\perp w_\kappa^{t'}\|^2 \leq 4c_r$, and $c_r \lesssim c_2$,

$$\begin{aligned}\theta_\kappa^\top w_\kappa^{t'+1} &\geq \theta_\kappa^\top w_\kappa^{t'} + \eta \lambda_\kappa \theta_\kappa^\top P_{w_\kappa^{t'}}^\perp (Z^{t'} + R^{t'}) - \eta^2 C_1^2 (\theta_\kappa^\top w_\kappa^{t'}) s \\ &\quad + \eta \lambda_\kappa \sum_{i=1}^{q} \Big[i \alpha_i \beta_{\kappa,i} \|\Sigma_\kappa w_\kappa^{t'}\|^{-(i-1)} (\theta_\kappa^\top w_\kappa^{t'})^{i-1} (1 - (\theta_\kappa^\top w_\kappa^{t'})^2)\end{aligned}$$

$$+\sqrt{(i+2)(i+1)}\alpha_i\beta_{\kappa,i+2}\|\Sigma_\kappa w_\kappa^{t'}\|^{-(i+1)}\|\Sigma_\kappa^\perp w_\kappa^{t'}\|^2(\theta_\kappa^\top w_\kappa^{t'})^{i+1}\Big]$$
$$\geq \theta_\kappa^\top w_\kappa^{t'} + \frac{3}{2}\eta\alpha_2\beta_{\kappa,2}c_2 + \eta\lambda_\kappa\theta_\kappa^\top P_{w_\kappa^{t'}}^\perp(Z^{t'}+R^{t'}).$$

Using $\sum_{t'=0}^{t-1}\eta\lambda_\kappa\theta_\kappa^\top P_{w_\kappa^{t'}}^\perp R^{t'} \leq \eta C_1(\mathcal{E}_s\vee 2c_r) \leq \frac{1}{4}\eta\alpha_2\beta_{\kappa,2}c_2$ and

$$\sum_{t'=0}^{t-1}\eta\lambda_\kappa\theta_\kappa^\top P_{w_\kappa^{t'}}^\perp Z^{t'} \leq \eta C_1\sqrt{t} \leq \begin{cases} C_2\eta & (t\leq C_2) \\ \frac{1}{4}\bar{\alpha}_2\beta_{\kappa,2}c_2\eta t & (t>C_2)\end{cases} \quad \text{with high probability,}$$

we obtain $\theta_\kappa^\top w_\kappa^t \geq \theta_\kappa^\top w_\kappa^0 - C_2\eta + \bar{\alpha}_2\beta_{\kappa,2}c_2\eta t$. In particular $\theta_\kappa^\top w_\kappa^t \geq 1-4c_1$, completing the induction. □

Lemma 31. Let $\mathcal{E}_s \leq \bar{\varepsilon} \leq c_r$, $\eta \leq c_\eta\bar{\varepsilon}s^{-1}\wedge c_\eta\bar{\varepsilon}^2$, $\theta_\kappa^\top w_\kappa^0 \geq 1-3c_1$. Let $\tau>0$ and suppose $\|\Sigma_\kappa^\perp w_\kappa^t\| \leq 2\bar{\varepsilon}$ and $\theta_\kappa^\top w_\kappa^t \leq 1-C_2\bar{\varepsilon}$ for all $t=0,1,\ldots,\tau-1$. Then

$$\theta_\kappa^\top w_\kappa^t \geq 1-(1-\eta\bar{\alpha}_2\beta_{\kappa,2})^t(1-\theta_\kappa^\top w_\kappa^0) - c_3\bar{\varepsilon}.$$

Proof. Fix $t\leq\tau$ and assume $\theta_\kappa^\top w_\kappa^{t'} \geq 1-4c_1$ for all $t'\leq t-1$. By the same argument as in Lemma 30,

$$\begin{aligned}\theta_\kappa^\top w_\kappa^{t'+1} &\geq \theta_\kappa^\top w_\kappa^{t'} + \eta\lambda_\kappa\theta_\kappa^\top P_{w_\kappa^{t'}}^\perp(Z^{t'}+R^{t'}) - \eta^2C_1^2(\theta_\kappa^\top w_\kappa^{t'})s \\ &\quad + \eta\lambda_\kappa\sum_{i=1}^q\Big[i\alpha_i\beta_{\kappa,i}\|\Sigma_\kappa w_\kappa^{t'}\|^{-(i-1)}(\theta_\kappa^\top w_\kappa^{t'})^{i-1}(1-(\theta_\kappa^\top w_\kappa^{t'})^2) \\ &\qquad +\sqrt{(i+2)(i+1)}\alpha_i\beta_{\kappa,i+2}\|\Sigma_\kappa w_\kappa^{t'}\|^{-(i+1)}\|\Sigma_\kappa^\perp w_\kappa^{t'}\|^2(\theta_\kappa^\top w_\kappa^{t'})^{i+1}\Big] \\ &\geq \theta_\kappa^\top w_\kappa^{t'} + \frac{3}{2}\eta\alpha_2\beta_{\kappa,2}(1-\theta_\kappa^\top w_\kappa^{t'}) + \eta\lambda_\kappa\theta_\kappa^\top P_{w_\kappa^{t'}}^\perp(Z^{t'}+R^{t'}).\end{aligned}$$

Since $\eta\lambda_\kappa\theta_\kappa^\top P_{w_\kappa^{t'}}^\perp R^{t'} \leq \eta C_1(\mathcal{E}_s\vee 2\bar{\varepsilon}) \leq \frac{1}{2}\eta\bar{\alpha}_2\beta_{\kappa,2}(1-\theta_\kappa^\top w_\kappa^{t'})$,

$$\begin{aligned}1-\theta_\kappa^\top w_\kappa^t &\leq (1-\eta\alpha_2\beta_{\kappa,2})^t(1-\theta_\kappa^\top w_\kappa^0) + \eta\lambda_\kappa\sum_{t'=0}^{t-1}(1-\eta\alpha_2\beta_{\kappa,2})^{t-t'-1}\theta_\kappa^\top P_{w_\kappa^{t'}}^\perp Z^{t'} \\ &\leq (1-\eta\alpha_2\beta_{\kappa,2})^t(1-\theta_\kappa^\top w_\kappa^0) + \frac{\sqrt{\eta}C_1\lambda_\kappa}{\sqrt{\alpha_2\beta_{\kappa,2}}}.\end{aligned}$$

where the weighted martingale sum is bounded using the conditional variance of the martingale increments. Since $\eta\leq c_\eta\bar{\varepsilon}^2$, the second term is at most $c_3\bar{\varepsilon}$, giving the stated bound. In particular $\theta_\kappa^\top w_\kappa^t \geq 1-4c_1$, completing the induction. □

Lemma 32. Let $\mathcal{E}_s \leq c_\varepsilon s^{-1/2}$ and $C_2\,\mathcal{E}_s \leq \tilde{\varepsilon}$. Set

$$\eta^t = \begin{cases}\eta_1 \leq c_\eta s^{-3/2} & (0\leq t\leq \Delta T_1+\Delta T_2-1) \\ \eta_2 \leq c_\eta C_2^{-1}\tilde{\varepsilon}s^{-1}\wedge c_\eta C_2^{-2}\tilde{\varepsilon}^2 & (\Delta T_1+\Delta T_2\leq t\leq \Delta T_1+\Delta T_2+T_{1,3}+T_{1,4}-1),\end{cases}$$

where $\Delta T_1 = T_{1,1}-t_1$, $\Delta T_2 = T_{1,2}-t_2$, $T_{1,3}=\tilde{\Theta}(\eta_2^{-1}\log\tilde{\varepsilon}^{-1})$, and $T_{1,4}=\tilde{\Theta}(\eta_2^{-1}\log\tilde{\varepsilon}^{-1})$. Then for every neuron with $\theta_\kappa^\top w_\kappa^0\geq 1-c_1$ and $\|\Sigma_\kappa^\perp w_\kappa^0\|^2\leq 2c_r$,

$$\theta_\kappa^\top w_\kappa^{\Delta T_1+\Delta T_2+T_{1,3}+T_{1,4}} > 1-2\tilde{\varepsilon}$$

with high probability.

Proof. Phase 1: $0\leq t\leq\Delta T_1+\Delta T_2$. Set $\tau=\lceil 5c_w^{-1}\eta_1^{-1}\rceil$ and $\bar{\varepsilon}=s^{-1/2}$, so $\eta_1\leq c_\eta\bar{\varepsilon}s^{-1}$.

We first control $\|\Sigma_\kappa^\perp w_\kappa^t\|^2$. By Lemma 23, $\theta_\kappa^\top w_\kappa^t \geq 1-C_1\eta_1\tau\geq\frac{1}{2}$ for all $t\leq\tau$, so Lemma 29 is applicable throughout $[0,\tau]$. If $\|\Sigma_\kappa^\perp w_\kappa^t\|^2\geq\bar{\varepsilon}$ throughout $[0,\tau]$, then Lemma 29 gives $\|\Sigma_\kappa^\perp w_\kappa^\tau\|^2 \leq e^{-c_w\eta_1\tau/5}\|\Sigma_\kappa^\perp w_\kappa^0\|^2 + c_r\bar{\varepsilon}\leq 2c_r$, and $\|\Sigma_\kappa^\perp w_\kappa^t\|^2\leq 3c_r$ for all $t\leq\tau$. If instead $\|\Sigma_\kappa^\perp w_\kappa^{t'}\|^2<\bar{\varepsilon}$ for some $t'\leq\tau$, then by Lemma 24 the quantity increases by at most $7C_1\eta_1$ per step, so $\|\Sigma_\kappa^\perp w_\kappa^{t'+1}\|^2\leq$

$\bar{\varepsilon} + 7C_1\eta_1 \le c_r$, and Lemma 29 applies again from $t'+1$. Repeating this argument at each such t' yields $\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \le 3c_r$ for all $t \le \Delta T_1 + \Delta T_2$.

We then control $\theta_\kappa^\top w_\kappa^t$ via Lemma 30. If $\theta_\kappa^\top w_\kappa^t \le 1 - c_2$ throughout $[0, \tau]$, then $\theta_\kappa^\top w_\kappa^\tau \ge \theta_\kappa^\top w_\kappa^0 - C_2\eta_1 + \bar{\alpha}_2\beta_{\kappa,2}c_2\eta_1\tau \ge 1 - c_1$, and $\theta_\kappa^\top w_\kappa^t \ge 1 - 2c_1$ for all $t \le \tau$. If instead $\theta_\kappa^\top w_\kappa^{t'} > 1 - c_2$ for some $t' \le \tau$, then by Lemma 23 the alignment decreases by at most $C_2\eta_1$ per step, so $\theta_\kappa^\top w_\kappa^{t'+1} \ge 1 - c_2 - C_2\eta_1 \ge 1 - 2c_1$, and the same argument applies from $t'+1$. Repeating at each such t' gives $\theta_\kappa^\top w_\kappa^t \ge 1 - 2c_1$ for all $t \le \Delta T_1 + \Delta T_2$.

Phase 2: $\Delta T_1 + \Delta T_2 \le t \le \Delta T_1 + \Delta T_2 + T_{1,3} + T_{1,4}$. Set $\bar{\varepsilon} = C_2^{-1}\tilde{\varepsilon}$ and define

$$T_{1,3} = \left\lceil 5c_w^{-1}\eta_2^{-1}\log\bar{\varepsilon}^{-1}\right\rceil, \qquad T_{1,4} = \left\lceil \eta_2^{-1}\bar{\alpha}_2^{-1}\beta_{\kappa,2}^{-1}\log\bar{\varepsilon}^{-1}\right\rceil.$$

By the same argument as Phase 1, $\theta_\kappa^\top w_\kappa^t \ge 1 - 3c_1$ and $\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \le 4c_r$ throughout this phase. If $\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \ge \bar{\varepsilon}$ for all $t \le T$, then Lemma 29 gives

$$\|\Sigma_\kappa^\perp w_\kappa^{\Delta T_1+\Delta T_2+T_{1,3}}\|^2 \le (1 - \tfrac{1}{5}c_w\eta_2)^{T_{1,3}} \cdot 4c_r + c_r\bar{\varepsilon} \le 4c_r\bar{\varepsilon} < \bar{\varepsilon},$$

a contradiction. Hence there exists $t_3 \le T_{1,3}$ with $\|\Sigma_\kappa^\perp w_\kappa^{\Delta T_1+\Delta T_2+t_3}\|^2 < \bar{\varepsilon}$. Applying Lemma 29 repeatedly thereafter keeps $\|\Sigma_\kappa^\perp w_\kappa^t\|^2 \le 2\bar{\varepsilon}$. If $\theta_\kappa^\top w_\kappa^{\Delta T_1+\Delta T_2+t} \le 1 - C_2\bar{\varepsilon}$ for all t in $[t_3, T_{1,3}+T_{1,4}]$, then Lemma 31 gives

$$\theta_\kappa^\top w_\kappa^{\Delta T_1+\Delta T_2+T_{1,3}+T_{1,4}} \ge 1 - (1 - \eta_2\bar{\alpha}_2\beta_{\kappa,2})^{T_{1,4}-t_3}(1 - \theta_\kappa^\top w_\kappa^{\Delta T_1+\Delta T_2+t_3}) - c_3\bar{\varepsilon} \ge 1 - 3c_1\bar{\varepsilon} - c_3\bar{\varepsilon} > 1 - C_2\bar{\varepsilon},$$

a contradiction. Hence there exists $t_4 \le T_{1,4}$ with $\theta_\kappa^\top w_\kappa^{\Delta T_1+\Delta T_2+T_{1,3}-t_3+t_4} > 1 - C_2\bar{\varepsilon}$. Applying Lemma 31 for the remaining steps gives $\theta_\kappa^\top w_\kappa^{\Delta T_1+\Delta T_2+T_{1,3}+T_{1,4}} \ge 1 - C_2\bar{\varepsilon} - C_1\eta_2 \ge 1 - 2\tilde{\varepsilon}$. □

E.9 Non-Forgetting of Other Tasks

Lemma 33. Let $\eta = \eta^t$ with $\chi_k C_1\eta s^{1/2} \le 1$. Then for every $t' = 1, 2, \ldots,$

$$\|\Sigma_\kappa w_k^{t'}\|^2 \le \Big(\|\Sigma_\kappa w_k^0\|^2 + 2\chi_k C_1\eta\Big(\sqrt{t'} + \mathcal{E}_s\, t'\Big) + 2(\chi_k C_1\eta)^2 s\, t'\Big)\exp(2\chi_k C_1\eta\, t').$$

Proof. Set $g = g_k^t$ using the notation of Lemma 22. Since $\eta\|P_{w_k^t}^\perp g\| \le \chi_k C_1\eta s^{1/2} \le 1$ with high probability,

$$\begin{aligned}
\|\Sigma_\kappa w_k^{t+1}\|^2 &= \frac{(w_k^t + \eta P_{w_k^t}^\perp g)^\top \Sigma_\kappa (w_k^t + \eta P_{w_k^t}^\perp g)}{\|w_k^t + \eta P_{w_k^t}^\perp g\|^2} \\
&\le (w_k^t + \eta P_{w_k^t}^\perp g)^\top \Sigma_\kappa (w_k^t + \eta P_{w_k^t}^\perp g)(1 - \eta^2\|P_{w_k^t}^\perp g\|^2) \\
&= (\|\Sigma_\kappa w_k^t\|^2 + 2\eta w_k^t \Sigma_\kappa P_{w_k^t}^\perp g + \eta^2\|\Sigma_\kappa P_{w_k^t}^\perp g\|^2)(1 - \eta^2\|P_{w_k^t}^\perp g\|^2) \\
&\le \|\Sigma_\kappa w_k^t\|^2 + 2\eta w_k^t \Sigma_\kappa P_{w_k^t}^\perp g + \eta^2\|\Sigma_\kappa P_{w_k^t}^\perp g\|^2 + \eta^2\|P_{w_k^t}^\perp g\|^2 |2\eta w_k^t \Sigma_\kappa P_{w_k^t}^\perp g| \\
&\le \|\Sigma_\kappa w_k^t\|^2 + 2\eta w_k^t \Sigma_\kappa P_{w_k^t}^\perp g + 2(\chi_k C_1\eta)^2 s.
\end{aligned}$$

By Lemma 22 (ii),

$$\begin{aligned}
2\eta w_k^{t\top}\Sigma_\kappa P_{w_k^t}^\perp g &= 2\chi_k\eta w_k^{t\top}\Sigma_\kappa P_{w_k^t}^\perp (Z^t + R^t) \\
&\le 2\chi_k\eta\Big[w_k^{t\top}\Sigma_\kappa P_{w_k^t}^\perp Z^t + C_1\,\mathcal{E}_s + C_1\|\Sigma_\kappa w_k^t\|^2\Big].
\end{aligned}$$

Summing over $t = 0, \ldots, t'-1$ and bounding the martingale term $\sum_{t=0}^{t'-1} w_k^{t\top}\Sigma_\kappa P_{w_k^t}^\perp Z^t \le C_1\sqrt{t'}$ with high probability,

$$\|\Sigma_\kappa w_k^{t'}\|^2 \le \|\Sigma_\kappa w_k^0\|^2 + 2\chi_k C_1\eta\left(\sum_{t=0}^{t'-1}\|\Sigma_\kappa w_k^t\|^2 + \sqrt{t'} + \mathcal{E}_s\, t'\right) + 2(\chi_k C_1\eta)^2 s\, t'.$$

The stated bound follows by Grönwall's inequality. □

Lemma 34. Assume the hypotheses of Lemma 31: $\mathcal{E}_s \le c_\varepsilon s^{-1/2}$, $\mathcal{E}_s \le \bar\varepsilon$, and $\theta_k^\top w_k^0 \ge \frac{3}{4}$. Set

$$\eta^t = \begin{cases} \eta_1 \le c_\eta s^{-3/2} & (0 \le t \le T_{1,1}+T_{1,2}-1) \\ \eta_2 \le c_\eta \bar\varepsilon s^{-1} \wedge c_\eta \bar\varepsilon^2 & (T_{1,1}+T_{1,2} \le t \le T_{1,1}+T_{1,2}+T_{1,3}+T_{1,4}-1). \end{cases}$$

(i) If $\chi_k \le c_\chi$, then

$$\theta_k^\top w_k^{T_{1,1}+T_{1,2}} \ge \theta_k^\top w_k^0 - \chi_k C_3\big(\|\Sigma_\kappa w_k^0\|^2 + \mathcal{E}_s\big) - \chi_k C_3\sqrt{\eta_1} - \chi_k^2 C_3 \eta_1 s.$$

(ii) If additionally $\chi_k \le c_2(\log \bar\varepsilon^{-1})^{-1}$, then

$$\begin{aligned} &\theta_k^\top w_k^{T_{1,1}+T_{1,2}+T_{1,3}+T_{1,4}} \qquad (17)\\ &\ge \theta_k^\top w_k^{T_{1,1}+T_{1,2}} - \chi_k C_2 \log\bar\varepsilon^{-1}\Big(\|\Sigma_\kappa w_k^{T_{1,1}+T_{1,2}}\|^2 + \mathcal{E}_s\Big) - \chi_k C_2\sqrt{\eta_2 \log\bar\varepsilon^{-1}} - \chi_k^2 C_2 \eta_2 s \log\bar\varepsilon^{-1} \\ &\ge \theta_k^\top w_k^0 - \chi_k(C_3 \vee 2C_2\log\bar\varepsilon^{-1})\Big(\|\Sigma_\kappa w_k^0\|^2 + \mathcal{E}_s + c_3(\chi_k s^{-\frac{3}{4}} + \chi_k^2 s^{-\frac{1}{2}})\Big) - \chi_k c_3 \bar\varepsilon \log\bar\varepsilon^{-1}. \end{aligned}$$

Proof. Set $g = g_k^t$ using the notation of Lemma 22. Since $\chi_k \le c_\chi$, we have $\chi_k C_1 \eta^t s^{1/2} \le \frac{1}{2}$ at every step. When $\theta_k^\top w_k^t \ge \frac{1}{2}$, since $\theta_k^\top(w_k^t + \eta^t P_{w_k^t}^\perp g) \ge \theta_k^\top w_k^t - \chi_k C_1 \eta^t \ge 0$, we have $\theta_k^\top w_k^{t+1} \ge 0$ and

$$\theta_k^\top w_k^{t+1} \ge \theta_k^\top w_k^t + \eta^t \theta_k^\top P_{w_k^t}^\perp g - \tfrac{1}{2}(\eta^t)^2\|g\|^2(\theta_k^\top w_k^t) - \tfrac{1}{2}(\eta^t)^3|\theta_k^\top P_{w_k^t}^\perp g|\|g\|^2. \qquad (18)$$

For $\eta = \eta_1$ or $\eta = \eta_2$, set $\bar t = \chi_k C_1 \eta t$ and define

$$P^t = \theta_k^\top w_k^0 - \bar t e^{2\bar t}\Big[\|\Sigma_\kappa w_k^0\|^2 + \tfrac{4}{3}\sqrt{\chi_k C_1 \eta\, \bar t} + \bar t(\mathcal{E}_s + \chi_k C_1 \eta s)\Big] - \sqrt{\chi_k C_1 \eta\, \bar t}\Big(1 + \mathcal{E}_s\sqrt{t}\Big) - \chi_k C_1 \eta s\, \bar t. \qquad (19)$$

We prove by induction that $\theta_k^\top w_k^t \ge P^t$ for all $t = 0, 1, \dots, T$ whenever $P^T \ge \frac{1}{2}$. The base case $t = 0$ is immediate. Given the hypothesis up to $t' - 1$, by (18) and the proof of Lemma 33,

$$\theta_k^\top w_k^{t'} \ge \theta_k^\top w_k^0 - \chi_k C_1 \eta \left(\sum_{t=0}^{t'-1} \|\Sigma_\kappa w_k^t\|^2 + \sqrt{t'} + \mathcal{E}_s\, t'\right) - (\chi_k C_1 \eta)^2 s\, t',$$

and bounding $\sum_{t=0}^{t'-1}\|\Sigma_\kappa w_k^t\|^2$ via Lemma 33 yields

$$\begin{aligned} \sum_{t=0}^{t'-1}\|\Sigma_\kappa w_k^t\|^2 &\le \sum_{t=0}^{t'-1}\Big(\|\Sigma_\kappa w_k^0\|^2 + 2\chi_k C_1\eta(\sqrt{t} + \mathcal{E}_s t) + 2(\chi_k C_1\eta)^2 s t\Big)\exp(2\chi_k C_1 \eta t) \\ &\le \exp(2\chi_k C_1 \eta t')\left[\|\Sigma_\kappa w_k^0\|^2 t' + 2\chi_k C_1\eta\left(\frac{2}{3}t'^{\frac{3}{2}} + \frac{1}{2}\mathcal{E}_s t'^2\right) + (\chi_k C_1\eta)^2 s t'^2\right] \end{aligned}$$

and

$$\begin{aligned} &\theta_k^\top w_k^0 - \chi_k C_1 \eta t' \exp(2\chi_k C_1\eta t')\left[\|\Sigma_\kappa w_k^0\|^2 + \frac{4}{3}\chi_k C_1\eta\sqrt{t'} + \chi_k C_1\eta\mathcal{E}_s t' + (\chi_k C_1\eta)^2 s t'\right] \\ &\quad - \chi_k C_1\eta\left(\sqrt{t'} + \mathcal{E}_s t'\right) - (\chi_k C_1\eta)^2 s t' \\ &= \theta_k^\top w_k^0 - \bar t\exp(2\bar t)\left[\|\Sigma_\kappa w_k^0\|^2 + \frac{4}{3}\sqrt{\chi_k C_1\eta\bar t} + \bar t\,(\mathcal{E}_s + \chi_k C_1\eta s)\right] \\ &\quad - \sqrt{\chi_k C_1\eta\bar t}\left(1 + \mathcal{E}_s\sqrt{t'}\right) - \chi_k C_1\eta s\bar t \\ &= P^{t'}. \end{aligned}$$

Hence $\theta_k^\top w_k^{t'} \ge P^{t'}$. Since P^t is decreasing in t, the induction goes through whenever $P^T \ge \frac{1}{2}$.

Proof of (i). Let $\eta = \eta_1 \le c_\eta s^{-3/2}$ and set $\tau_1 = \lfloor \frac{1}{2} s^{-1/2} C_1^{-1}\eta_1^{-1}\rfloor$ and $\tau_2 = \lfloor \frac{1}{2} c_1 C_1^{-1}\eta_1^{-1}\rfloor$. Then

$$T_{1,1} + T_{1,2} = \left\lceil \frac{3\log s}{2\alpha_2\beta_{\kappa,2}\eta_1\tau_1}\right\rceil \tau_1 + \left\lceil \frac{3\log s}{2c_1\alpha_2\beta_{\kappa,2}\eta_1\tau_2}\right\rceil \tau_2$$

$$\le \frac{3\log s}{2\alpha_2\beta_{\kappa,2}\eta_1} + \tau_1 + \frac{3\log s}{2c_1\alpha_2\beta_{\kappa,2}\eta_1} + \tau_2 \le C_2\eta_1^{-1}.$$

Hence $\bar{t} \coloneqq \chi_k C_1\eta_1 T \le \chi_k C_1 C_2 \le c_3$. Substituting $T = T_{1,1} + T_{1,2}$ into (19) and using $\bar{t} \le \chi_k C_1 C_2$ and $T \le C_2\eta_1^{-1}$,

$$\begin{aligned}
P^T &\ge \theta_k^\top w_k^0 - \chi_k C_1 C_2 e^{2c_3}\Big[\|\Sigma_\kappa w_k^0\|^2 + \tfrac{4}{3}\chi_k C_1\sqrt{C_2\eta_1} + \chi_k C_1 C_2(\mathcal{E}_s + \chi_k C_1\eta_1 s)\Big] \\
&\quad - \chi_k C_1\sqrt{C_2\eta_1} - \chi_k C_1 C_2\,\mathcal{E}_s - \chi_k^2 C_1^2 C_2\eta_1 s \\
&\ge \theta_k^\top w_k^0 - \chi_k C_1 C_2 e^{2c_3}\|\Sigma_\kappa w_k^0\|^2 - \big(1 + \tfrac{4}{3}c_3 e^{2c_3}\big)\chi_k C_1\sqrt{C_2\eta_1} \\
&\quad - \chi_k C_1 C_2\big(1 + c_3 e^{2c_3}\big)\mathcal{E}_s - \big(1 + e^{2c_3}\big)\chi_k^2 C_1^2 C_2\eta_1 s \\
&\ge \theta_k^\top w_k^0 - 2\chi_k C_1 C_2\|\Sigma_\kappa w_k^0\|^2 - 2\chi_k C_1\sqrt{C_2\eta_1} - 2\chi_k C_1 C_2\,\mathcal{E}_s - 3\chi_k^2 C_1^2 C_2\eta_1 s \\
&\ge \theta_k^\top w_k^0 - \chi_k C_3\big(\|\Sigma_\kappa w_k^0\|^2 + \mathcal{E}_s\big) - \chi_k C_3\sqrt{\eta_1} - \chi_k^2 C_3\eta_1 s \ge \tfrac{2}{3},
\end{aligned}$$

where the second inequality expands the bracket, the third uses $c_3 \le \frac{3}{8}$ and $e^{2c_3} \le 2$, the fourth uses $C_1 \lesssim C_2 \lesssim C_3$, and the last uses $\theta_k^\top w_k^0 \ge \frac{3}{4}$, $\mathcal{E}_s \le c_\varepsilon s^{-1/2}$, and $\eta_1 \le c_\eta s^{-3/2}$.

Proof of (ii). Re-indexing $t \mapsto t - T_{1,1} - T_{1,2}$, consider $0 \le t \le T_{1,3} + T_{1,4}$ with $\eta = \eta_2$. We have

$$T_{1,3} + T_{1,4} = \big\lceil 5c_w^{-1}\eta_2^{-1}\log\bar{\varepsilon}^{-1}\big\rceil + \big\lceil \eta_2^{-1}\alpha_2^{-1}\beta_{\kappa,2}^{-1}\log\bar{\varepsilon}^{-1}\big\rceil \le C_1\eta_2^{-1}\log\bar{\varepsilon}^{-1},$$

so $\bar{t} \coloneqq \chi_k C_1\eta_2 T \le \chi_k C_1^2\log\bar{\varepsilon}^{-1} \le c_1$. Substituting $T = T_{1,3} + T_{1,4}$ into (19) and using $\bar{t} \le \chi_k C_1^2\log\bar{\varepsilon}^{-1}$ and $T \le C_1\eta_2^{-1}\log\bar{\varepsilon}^{-1}$,

$$\begin{aligned}
P^T &\ge \theta_k^\top w_k^0 - \chi_k C_1^2\log\bar{\varepsilon}^{-1}\, e^{2c_1}\left[\|\Sigma_\kappa w_k^0\|^2 + \tfrac{4}{3}\sqrt{\chi_k^2 C_1^3\eta_2\log\bar{\varepsilon}^{-1}} + \chi_k C_1^2\log\bar{\varepsilon}^{-1}(\mathcal{E}_s + \chi_k C_1\eta_2 s)\right] \\
&\quad - \sqrt{\chi_k^2 C_1^3\eta_2\log\bar{\varepsilon}^{-1}} - \sqrt{\chi_k^2 C_1^4(\log\bar{\varepsilon}^{-1})^2}\,\mathcal{E}_s - \chi_k^2 C_1^3\eta_2 s\log\bar{\varepsilon}^{-1} \\
&\ge \theta_k^\top w_k^0 - 2\chi_k C_1^2\log\bar{\varepsilon}^{-1}\big(\|\Sigma_\kappa w_k^0\|^2 + \mathcal{E}_s\big) - 2\sqrt{\chi_k^2 C_1^3\eta_2\log\bar{\varepsilon}^{-1}} - 3\chi_k^2 C_1^3\eta_2 s\log\bar{\varepsilon}^{-1} \\
&\ge \theta_k^\top w_k^0 - \chi_k C_2\log\bar{\varepsilon}^{-1}\big(\|\Sigma_\kappa w_k^0\|^2 + \mathcal{E}_s\big) - \chi_k C_2\sqrt{\eta_2\log\bar{\varepsilon}^{-1}} - \chi_k^2 C_2\eta_2 s\log\bar{\varepsilon}^{-1},
\end{aligned}$$

establishing the first line of (17). For the second line, substitute (i) and Lemma 33 into (17) to bound $\|\Sigma_\kappa w_k^{T_{1,1}+T_{1,2}}\|^2$,

$$\begin{aligned}
&\theta_k^\top w_k^{T_{1,1}+T_{1,2}+T_{1,3}+T_{1,4}} \\
&\ge \theta_k^\top w_k^0 - \chi_k C_3\big(\|\Sigma_\kappa w_k^0\|^2 + \mathcal{E}_s\big) - \chi_k C_3\sqrt{\eta_1} - \chi_k^2 C_3\eta_1 s \\
&\quad - \chi_k C_2\log\bar{\varepsilon}^{-1}\Big(\|\Sigma_\kappa w_k^0\|^2 + 2\chi_k C_1\eta_1\Big(\sqrt{T_{1,1}+T_{1,2}} + \mathcal{E}_s(T_{1,1}+T_{1,2})\Big) + 4(\chi_k C_1\eta_1)^2 s(T_{1,1}+T_{1,2})\Big) \\
&\qquad \times \exp(2\chi_k C_1\eta_1(T_{1,1}+T_{1,2})) \\
&\quad - \chi_k C_2\log\bar{\varepsilon}^{-1}\,\mathcal{E}_s - \chi_k C_2\sqrt{\eta_2\log\bar{\varepsilon}^{-1}} - \chi_k^2 C_2\eta_2 s\log\bar{\varepsilon}^{-1} \\
&\ge \theta_k^\top w_k^0 - \chi_k C_3\big(\|\Sigma_\kappa w_k^0\|^2 + \mathcal{E}_s\big) - \chi_k C_3\sqrt{c_\eta}\,s^{-\frac{3}{4}} - \chi_k^2 C_3 c_\eta\, s^{-\frac{1}{2}} \\
&\quad - \chi_k C_2\log\bar{\varepsilon}^{-1}\Big(\|\Sigma_\kappa w_k^0\|^2 + 2\chi_k C_1\Big(\sqrt{C_2 c_\eta}\,s^{-\frac{3}{4}} + \mathcal{E}_s C_2\Big) + 4\chi_k^2 C_1^2 C_2 c_\eta\, s^{-\frac{1}{2}}\Big)\, e^{2\chi_k C_1 C_2 c_\eta} \\
&\quad - \chi_k C_2\log\bar{\varepsilon}^{-1}\,\mathcal{E}_s - \chi_k C_2\sqrt{c_\eta\bar{\varepsilon}^2\log\bar{\varepsilon}^{-1}} - \chi_k^2 C_2 c_\eta\bar{\varepsilon}\log\bar{\varepsilon}^{-1} \\
&\ge \theta_k^\top w_k^0 - \chi_k(C_3 \vee 2C_2\log\bar{\varepsilon}^{-1})\Big(\|\Sigma_\kappa w_k^0\|^2 + \mathcal{E}_s + 2\chi_k C_1\sqrt{C_2 c_\eta}\,s^{-\frac{3}{4}} + 4\chi_k^2 C_1^2 C_2 c_\eta\, s^{-\frac{1}{2}}\Big) \\
&\quad - \big(\chi_k C_2\sqrt{c_\eta} + \chi_k^2 C_2 c_\eta\big)\,\bar{\varepsilon}\log\bar{\varepsilon}^{-1} \\
&\ge \theta_k^\top w_k^0 - \chi_k(C_3 \vee 2C_2\log\bar{\varepsilon}^{-1})\Big(\|\Sigma_\kappa w_k^0\|^2 + \mathcal{E}_s + c_3(\chi_k s^{-\frac{3}{4}} + \chi_k^2 s^{-\frac{1}{2}})\Big) - \chi_k c_3\bar{\varepsilon}\log\bar{\varepsilon}^{-1},
\end{aligned}$$

where the third inequality uses $T_{1,1} + T_{1,2} \le C_2\eta_1^{-1}$ and $\chi_k C_1 C_2 \le c_3$. □

F Full proof of Theorem 2: SFT forgetting

In this section we show that, when the strong model is trained by SFT on a specific task (as in Algorithm 3), catastrophic forgetting of previously learned features occurs during first-layer training.

As in Section E, we rescale the learning rate by $\eta \leftarrow N_\kappa \pi_\kappa^{-1} \eta$ to normalize for the signal of task κ. Under this rescaling, the update rule for the weight w_k associated with task k in Algorithm 3 reads

$$w_k^{t+1} \leftarrow w_k^t + \eta g_k^t, \quad \text{where} \quad g_k^t = \chi_k \pi_k^{-1} \nabla_{w_k} \bar{y}^t a_k \sigma_k(w_k^\top x + b_k).$$

By the same arguments as in Section E.8, $\Theta(N_\kappa)$ of the N_κ neurons associated with task κ achieve alignment $1 - \tilde{\varepsilon}$ within

$$\tilde{\Theta}\Big(d^{\frac{p}{2}} s^{\frac{p-2}{2}} \vee d\tilde{\varepsilon}^{-1} \log \tilde{\varepsilon}^{-1} \vee \tilde{\varepsilon}^{-2} \log \tilde{\varepsilon}^{-1}\Big)$$

steps of online SGD. For neurons associated with task $k \neq \kappa$, however, alignment with the true feature θ_k may be destroyed. This section derives the conditions under which such catastrophic forgetting occurs.

Set $c_{l,k} = p\alpha_p \beta_{k,p} = \tilde{\Theta}(1)$ and $c_{u,k} = q^2 \max_{p \le i \le q} |\alpha_i \beta_{k,i}| = \tilde{\Theta}(1)$. We further fix constants of order $\mathrm{polylog}\, d$ satisfying

$$\begin{cases} C_1 \lesssim c_1^{-1} \lesssim c_2^{-1} \\ c_{u,k}^{-1} \le c_{l,k}^{-1} \end{cases} \lesssim C_2 \lesssim c_\eta^{-1} = \tilde{O}(1).$$

These constants are distinct from those defined analogously in Sections D and E.

Assumption 6 (SFT initialization conditions)**.** Fix $k \neq \kappa$ with $\chi_k \lesssim c_1^{-1}$ and $n \in [N_k]$. The strong model is initialized as in Assumption 5, and the cross-task alignment satisfies

$$\theta_\kappa^\top w_{k,n}^0 \;\ge\; C_1\, s^{-1/2} \left(\frac{2^{p+1}(1+c_1)^{p-1}\, c_{u,\kappa}}{(1-\chi_k c_1)^{p-2}\, \chi_k\, c_{l,k}} \right)^{1/(p-2)}.$$

Remark. The intended setting is that the pre-trained strong model has task-k neurons satisfying $\theta_k^\top w_{k,n}^0 \approx 1$, then $\theta_\kappa^\top w_{k,n}^0 \approx \theta_k^\top \theta_\kappa$, and Assumption 6 reduces to requiring $\theta_k^\top \theta_\kappa \gtrsim s^{-1/2}$. The condition $\theta_k^\top \theta_\kappa \gtrsim s^{-1/2}$ holds when tasks k and κ are more similar than two random directions in the s-dimensional subspace, i.e., when the two tasks represent similar capabilities.

Under Assumption 6, the cross-task alignment $\theta_\kappa^\top w_{k,n}^0$ is large enough that task-k neurons converge to θ_κ faster than the task-κ neurons do, causing forgetting of the pre-trained feature θ_k.

Theorem 2 (SFT forgetting; formal version of Theorem 2)**.** Assume Assumptions 1 and 5 (i) with $\mathrm{IE}(\sigma_\kappa^*) = p \ge 3$, and that the strong model is trained by Algorithm 3 with learning rate $\eta^t \le c_\eta d^{-p/2}$. Let $t_{1,\kappa} = \min\big\{t : \theta_\kappa^\top w_{\kappa,n}^t \ge 1 - o(1) \text{ for some } n \in [N_\kappa]\big\}$. Then, for every $k \neq \kappa$ and $n \in [N_k]$ satisfying Assumption 6 whose Hermite coefficients satisfy $\alpha_p \tilde{\beta}_{k,p} > 0$ and $\alpha_i \tilde{\beta}_{k,i} \ge 0$ for all $i \ge p$, with high probability, $\theta_\kappa^\top w_{k,n}^t \ge 1 - o(1)$ for all $t \ge t_{1,\kappa}$.

The neurons in group k thus converge to θ_κ no later than the task-κ neurons do, so their alignment $\theta_k^\top w_{k,n}^t$ with the pre-trained feature θ_k is destroyed once task κ is learned.

Remark. Under the random initialization in Section E.2, $\Theta(N_k)$ neurons in group k satisfy the sign condition $\alpha_p \tilde{\beta}_{k,p} > 0$ and $\alpha_i \tilde{\beta}_{k,i} \ge 0$ for all $i \ge p$ with high probability.

The proof is given by Lemmas 35 to 37 below.

Lemma 35. Let $p \ge 2$, $\eta = \eta^t \le c_\eta d^{-\frac{p}{2}}$, and $\chi_k \lesssim c_1^{-1}$. Suppose $\theta_\kappa^\top w_k^0 \ge s^{-\frac{1}{2}}$. If $\theta_\kappa^\top w_k^t \le c_1$ for all $t = 1, 2, \ldots, \tau$, then for every $t \le \tau$,

$$\begin{aligned} &(1 - \chi_k c_1)\theta_\kappa^\top w_k^0 + \chi_k \eta \sum_{t'=0}^{t} (1 - c_1) c_{l,k} (\theta_\kappa^\top w_k^{t'})^{p-1} \\ &\le \theta_k^\top w_k^{t+1} \\ &\le (1 + \chi_k c_1)\theta_\kappa^\top w_k^0 + \chi_k \eta \sum_{t'=0}^{t} (1 + c_1) c_{u,k} (\theta_\kappa^\top w_k^{t'})^{p-1}. \end{aligned}$$

Moreover, defining sequences $(P_k^t)_{t=0}^{\tau+1}$ and $(Q_k^t)_{t=0}^{\tau+1}$ by $P_k^0 = (1 - \chi_k c_1)\theta_\kappa^\top w_k^0$, $Q_k^0 = (1 + \chi_k c_1)\theta_\kappa^\top w_k^0$, and for $t = 1, 2, \ldots, \tau$,

$$P_k^{t+1} = P_k^t + \chi_k \eta (1 - c_1) c_{l,k} (P_k^t)^{p-1},$$

Algorithm 3: Online SGD training of the strong model (SFT)

Input: Initialized strong model $r_{(\Theta_k)_{k=1}^K}$, $\Theta_k = (a_{n,k}, b_{n,k}, w_{n,k}^0)_{n=1}^{N_k}$, learning rate η^t, number of steps T, initialization scale C_b.

Phase I: first-layer training

for $t = 0, 1, \ldots, T-1$ **do**

$x^t \sim \mathcal{N}(0, I_d)$, $y^t = r_\kappa^*(\theta_\kappa^\top x) + \zeta$.

$w_{k,n}^{t+1} \leftarrow w_{k,n}^t + \eta^t y^t \tilde{\nabla}_w r^s_{((a_{k,n}, b_{k,n}, w_{k,n}^t)_{n=1}^{N_k})_{k=1}^K}(x^t)$

$w_{k,n}^{t+1} \leftarrow w_{k,n}^{t+1} / \|w_{k,n}^{t+1}\|, \quad (n = 1, \ldots, N_k), \quad (k = 1, \ldots, K)$

$\hat{w}_{k,n} \leftarrow w_{k,n}^T$

Output: $r_{\hat{\Theta}}(x)$ with $\hat{\Theta}_k = (\hat{a}_{k,n}, b_{k,n}, \hat{w}_{k,n})_{n=1}^{N_k}$.

$$Q_k^{t+1} = Q_k^t + \chi_k \eta (1 + c_1) c_{u,k} (Q_k^t)^{p-1},$$

we have $P_k^t \le \theta_\kappa^\top w_k^t \le Q_k^t$ for every $t \le \tau + 1$ with high probability.

Proof. For any $k \in [K]$, the gradient term decomposes as

$$g_k^t = \chi_k \sum_{i=p}^{q} \left[i\alpha_i \beta_{k,i} (\theta_\kappa^\top w_k^t)^{i-1} \theta_\kappa + \sqrt{(i+2)(i+1)} \alpha_i \beta_{k,i+2} (\theta_\kappa^\top w_k^t)^i w_k^t \right] + \chi_k Z^t,$$

where Z^t has mean 0 and satisfies $\|Z^t\| \le c_1 d^{\frac{1}{2}}$ with high probability and $|v^\top Z^t| \le c_1$ with high probability for every $v \in S^{d-1}$. If $\theta_\kappa^\top w_k^{t+1} \ge \frac{1}{2} d^{-\frac{1}{2}}$, then by the same argument as in Lemma 5,

$$\begin{aligned}
\theta_\kappa^\top w_k^{t+1} &\ge (\theta_\kappa^\top w_k^t + \eta \theta_\kappa^\top P_{w_k^t}^\perp g)(1 - \tfrac{1}{2}\eta^2 \|P_{w_k^t}^\perp g\|^2) \\
&\ge \theta_\kappa^\top w_k^t + \chi_k \eta c_{l,k} (\theta_\kappa^\top w_k^t)^{p-1} (1 - (\theta_\kappa^\top w_k^t)^2) - \chi_k^2 \eta^2 c_1^2 (\theta_\kappa^\top w_k^t) d + \chi_k \eta \theta_\kappa^\top P_{w_\kappa^t}^\perp Z^t,
\end{aligned}$$

where we used $\chi_k \eta c_1 \le 1$ and hence $\eta |\theta_\kappa^\top P_{w_k^t}^\perp g| \le 1$ with high probability. By an argument analogous to Lemma 6,

$$\theta_\kappa^\top w_k^{t+1} \ge (1 - \chi_k c_1) \theta_\kappa^\top w_k^0 + \chi_k \eta (1 - c_1) \sum_{t'=0}^{t} c_{l,k} (\theta_\kappa^\top w_k^{t'})^{p-1}. \tag{20}$$

On the other hand,

$$\begin{aligned}
\theta_\kappa^\top w_k^{t+1} &\le \theta_\kappa^\top w_k^t + \eta \theta_\kappa^\top P_{w_k^t}^\perp g \\
&= \theta_\kappa^\top w_k^t + \chi_k \eta \sum_{i=p}^{q} i\alpha_i \beta_{k,i} (\theta_\kappa^\top w_k^t)^{i-1} \theta_\kappa^\top P_{w_k^t}^\perp \theta_\kappa + \chi_k \eta \theta_\kappa^\top P_{w_k^t}^\perp Z^t \\
&\le \theta_\kappa^\top w_k^t + \chi_k \eta c_{u,k} (\theta_\kappa^\top w_k^t)^{p-1} (1 - (\theta_\kappa^\top w_k^t)^2) + \chi_k \eta \theta_\kappa^\top P_{w_k^t}^\perp Z^t,
\end{aligned}$$

and by an argument analogous to Lemma 6,

$$\theta_\kappa^\top w_k^{t+1} \le (1 + \chi_k c_1) \theta_\kappa^\top w_k^0 + \chi_k \eta (1 + c_1) \sum_{t'=0}^{t} c_{u,k} (\theta_\kappa^\top w_k^{t'})^{p-1}. \tag{21}$$

Combining (20) and (21) gives the desired bounds. □

In the following, we shift the time origin to $t_{1,k}$, the first time at which weak alignment of w_k^t is achieved.

Lemma 36. Let $\chi_k \lesssim c_1^{-1}$. Suppose $\theta_\kappa^\top w_k^0 \ge c_1$, $\eta = \eta_t \le c_\eta d^{-\frac{p}{2}}$, and $\theta_\kappa^\top w_k^t \le 1 - c_1$ for all $t \le \tau$. Then,

$$(1 - \chi_k c_1) \theta_\kappa^\top w_k^0 + \chi_k \eta \sum_{t'=0}^{t} c_1 c_{l,k} (\theta_\kappa^\top w_k^{t'})^{p-1} \le \theta_\kappa^\top w_k^{t+1}$$

$$\leq (1+\chi_k c_1)\theta_\kappa^\top w_k^0 + \chi_k \eta \sum_{t'=0}^{t} c_1 c_{u,k} (\theta_\kappa^\top w_k^{t'})^{p-1}.$$

Moreover, defining $(P_k^t)_{t=0}^{\tau+1}$ and $(Q_k^t)_{t=0}^{\tau+1}$ by $P_k^0 = (1-\chi_k c_1)\theta_\kappa^\top w_k^0$, $Q_k^0 = (1+\chi_k c_1)\theta_\kappa^\top w_k^0$, $P_k^{t+1} = P_k^t + \chi_k \eta c_1 c_{l,k} (P_k^t)^{p-1}$, and $Q_k^{t+1} = Q_k^t + \chi_k \eta c_1 c_{u,k} (Q_k^t)^{p-1}$, we have $P_k^t \leq \theta_\kappa^\top w_k^t \leq Q_k^t$ for every $t \leq \tau + 1$.

Proof. The proof is analogous to that of Lemma 8. □

Lemma 37. Let $p \geq 3$, $\eta = \eta_t \leq c_\eta d^{-\frac{p}{2}}$, $\chi_k \lesssim c_1^{-1}$, $\theta_\kappa^\top w_\kappa^t \geq s^{-\frac{1}{2}}$, and $c_1^{-(p-2)}(1+c_1)^{p-2}(\theta_\kappa^\top w_\kappa^0)^{p-2} \leq \frac{1}{2}$. If $\theta_\kappa^\top w_\kappa^t > c_1$, then

$$t \geq \frac{(1+c_1)^{-(p-1)}(\theta_\kappa^\top w_\kappa^0)^{-(p-2)}}{4(p-2)\eta c_{u,\kappa}}.$$

On the other hand, if

$$t \geq 2^{p-2} \frac{(1-c_1)^{-1}(\theta_\kappa^\top w_k^0)^{-(p-2)} + c_1^{-(p-1)}}{(1-\chi_k c_1)^{p-2}(p-2)\chi_k \eta c_{l,k}},$$

then $\theta_\kappa^\top w_k^t > 1 - c_1$. Furthermore, if $c_1 \gtrsim \theta_\kappa^\top w_k^0 \geq C_1 s^{-\frac{1}{2}} \left(\frac{2^{p+1}(1+c_1)^{p-1} c_{u,\kappa}}{(1-\chi_k c_1)^{p-2}\chi_k c_{l,k}}\right)^{\frac{1}{p-2}}$, then $\theta_\kappa^\top w_k^t > 1 - c_1$ for every $t \geq t_{1,\kappa}$.

Proof. By Lemma 35 and the Bihari–LaSalle inequality, for every $t \leq t_{1,\kappa}$,

$$\theta_\kappa^\top w_\kappa^t \leq \frac{(1+c_1)\theta_\kappa^\top w_\kappa^0}{(1-\eta(1+c_1)c_{u,k}(p-2)((1+c_1)\theta_\kappa^\top w_\kappa^0)^{p-2} t_{1,\kappa})^{\frac{1}{p-2}}},$$

so a necessary condition for $\theta_\kappa^\top w_\kappa^{t_{1,\kappa}} > c_1$ is

$$t_{1,\kappa} > \frac{1 - c_1^{-(p-2)}(1+c_1)^{p-2}(\theta_\kappa^\top w_\kappa^0)^{p-2}}{\eta(1+c_1)^{p-1} c_{u,\kappa}(p-2)(\theta_\kappa^\top w_\kappa^0)^{p-2}}.$$

Since $\theta_\kappa^\top w_\kappa^{t_{1,\kappa}} \leq c_1 + C_1\eta$, for every $t \leq t_{2,\kappa}$,

$$\theta_\kappa^\top w_\kappa^{t_{1,\kappa}+t} \leq \frac{(1+c_1)(c_1+C_1\eta)}{(1-\eta c_1 c_{u,\kappa}(p-2)((1+c_1)(c_1+C_1\eta))^{p-2} t_{2,\kappa})^{\frac{1}{p-2}}},$$

so a necessary condition for $\theta_\kappa^\top w_\kappa^{t_{1,\kappa}+t_{2,\kappa}} > 1 - c_1$ is

$$t_{2,\kappa} > \frac{1 - (1-c_1)^{-(p-2)}(1+c_1)^{p-2}(c_1+C_1\eta)^{p-2}}{\eta c_1 c_{u,\kappa}(p-2)(1+c_1)^{p-2}(c_1+C_1\eta)^{p-2}}.$$

Therefore,

$$\begin{aligned} t_{1,\kappa} + t_{2,\kappa} &\geq \frac{1}{(1+c_1)^{p-2}(p-2)\eta c_{u,\kappa}} \\ &\quad \times \left(\frac{1 - c_1^{-(p-2)}(1+c_1)^{p-2}(\theta_\kappa^\top w_\kappa^0)^{p-2}}{(1+c_1)(\theta_\kappa^\top w_\kappa^0)^{p-2}} + \frac{1 - (1-c_1)^{-(p-2)}(1+c_1)^{p-2}(c_1+C_1\eta)^{p-2}}{c_1(c_1+C_1\eta)^{p-2}} \right) \\ &\geq \frac{(1+c_1)^{-(p-1)}(\theta_\kappa^\top w_\kappa^0)^{-(p-2)} + c_1^{-1}(c_1+C_1\eta)^{-(p-2)}}{4(p-2)\eta c_{u,\kappa}}, \end{aligned}$$

where the second inequality uses $c_1^{-(p-2)}(1+c_1)^{p-2}(\theta_\kappa^\top w_\kappa^0)^{p-2} \leq \frac{1}{2}$, $(1-c_1)^{-(p-2)}(c_1 + C_1\eta)^{p-2} \leq \frac{1}{2}$, and $(1+c_1)^{p-2} \leq 2$. On the other hand, by Lemma 35 and the Bihari–LaSalle inequality, for every $t \geq t_{1,k}$,

$$\theta_\kappa^\top w_k^t \geq \frac{(1-\chi_k c_1)\theta_\kappa^\top w_k^0}{\left(1 - \chi_k \eta(1-c_1) c_{l,k} 2^{-(p-1)}(p-2)((1-\chi_k c_1)\theta_\kappa^\top w_k^0)^{p-2} t_{1,k}\right)^{\frac{1}{p-2}}},$$

using $\chi_k\eta(1-c_1)c_{l,k} \le 1$. A sufficient condition for $\theta_\kappa^\top w_k^{t_{1,k}} > c_1$ is then

$$t_{1,k} > \frac{1 - c_1^{-(p-2)}((1-\chi_k c_1)\theta_\kappa^\top w_k^0)^{p-2}}{\chi_k\eta(1-c_1)c_{l,k}2^{-(p-1)}(p-2)((1-\chi_k c_1)\theta_\kappa^\top w_k^0)^{p-2}}.$$

Since $\theta_\kappa^\top w_k^{t_{1,k}} \ge c_1$, for every $t \ge t_{2,k}$,

$$\theta_\kappa^\top w_k^{t_{1,k}+t} \ge \frac{(1-\chi_k c_1)c_1}{\left(1 - \chi_k\eta c_1 c_{l,k}2^{-(p-1)}(p-2)((1-\chi_k c_1)c_1)^{p-2}t_{2,k}\right)^{\frac{1}{p-2}}},$$

using $\chi_k\eta c_1 c_{l,k} \le 1$. A sufficient condition for $\theta_\kappa^\top w_k^{t_{2,k}} > 1 - c_1$ is

$$t_{2,k} > \frac{1 - (1-c_1)^{-(p-2)}((1-\chi_k c_1)c_1)^{p-2}}{\chi_k\eta c_1 c_{l,k}2^{-(p-1)}(p-2)((1-\chi_k c_1)c_1)^{p-2}}.$$

Therefore,

$$\begin{aligned} t_{1,k} + t_{2,k} &> \frac{2^{p-1}}{(1-\chi_k c_1)^{p-2}(p-2)\chi_k\eta c_{l,k}} \\ &\quad \times \left(\frac{1 - c_1^{-(p-2)}((1-\chi_k c_1)\theta_\kappa^\top w_k^0)^{p-2}}{(1-c_1)(\theta_\kappa^\top w_\kappa^0)^{p-2}} + \frac{1-(1-c_1)^{-(p-2)}((1-\chi_k c_1)c_1)^{p-2}}{c_1^{p-1}}\right) \\ &\ge 2^{p-2}\frac{(1-c_1)^{-1}(\theta_\kappa^\top w_k^0)^{-(p-2)} + c_1^{-(p-1)}}{(1-\chi_k c_1)^{p-2}(p-2)\chi_k\eta c_{l,k}}, \end{aligned}$$

where the second inequality uses $c_1^{-(p-2)}((1-\chi_k c_1)\theta_\kappa^\top w_k^0)^{p-2} \le \frac{1}{2}$ and $(1-c_1)^{-(p-2)}((1-\chi_k c_1)c_1)^{p-2} \le \frac{1}{2}$. Finally, if $\theta_\kappa^\top w_k^0 \ge C_1 s^{-\frac{1}{2}}\left(\frac{2^{p+1}(1+c_1)^{p-1}c_{u,\kappa}}{(1-\chi_k c_1)^{p-2}\chi_k c_{l,k}}\right)^{\frac{1}{p-2}}$, then with high probability,

$$\left(\frac{\theta_\kappa^\top w_k^0}{\theta_\kappa^\top w_\kappa^0}\right)^{p-2} \ge \frac{2^{p+1}(1+c_1)^{p-1}c_{u,\kappa}}{(1-\chi_k c_1)^{p-2}\chi_k c_{l,k}},$$

where we used $\theta_\kappa^\top w_\kappa^0 \le C_1 s^{-\frac{1}{2}}$ with high probability, which follows from the sub-Gaussianity of the uniform distribution on the sphere. Using $(1-c_1)^{-1}(\theta_\kappa^\top w_k^0)^{-(p-2)} + c_1^{-(p-1)} \le 2(\theta_\kappa^\top w_k^0)^{-(p-2)}$,

$$\frac{(1+c_1)^{-(p-1)}(\theta_\kappa^\top w_\kappa^0)^{-(p-2)}}{4(p-2)\eta c_{u,\kappa}} \ge 2^{p-2}\frac{(1-c_1)^{-1}(\theta_\kappa^\top w_k^0)^{-(p-2)} + c_1^{-(p-1)}}{(1-\chi_k c_1)^{p-2}(p-2)\chi_k\eta c_{l,k}}.$$

Since $t_{1,\kappa} \ge t_{1,k} + t_{2,k}$, it follows that $\theta_\kappa^\top w_k^t > 1 - c_1$ for every $t \ge t_{1,\kappa}$. □

G Technical lemmas

Lemma 38 (Bihari–LaSalle and Gronwall inequalities; cf. Oko et al. (2024a); Ben Arous et al. (2021))**.** Let $p \ge 3$, $c > 0$, and let $(a^t)_{t=0}^\infty$ be a sequence of positive reals satisfying $a^{t+1} = a^t + c(a^t)^{p-1}$. Then

$$a^t \le \frac{a^0}{\left(1 - c(p-2)(a^0)^{p-2}t\right)^{\frac{1}{p-2}}}.$$

Moreover, if $a^t \le 1$ for all $t \le T-1$, then

$$a^t \ge \frac{a^0}{\left(1 - c(1+c)^{-(p-1)}(p-2)(a^0)^{p-2}t\right)^{\frac{1}{p-2}}}.$$

Proof. Since the integrand $x^{-(p-1)}$ is decreasing,

$$c = \frac{a^{t+1}-a^t}{(a^t)^{p-1}} \ge \int_{a^t}^{a^{t+1}} \frac{dx}{x^{p-1}} = \frac{1}{p-2}\left[\frac{1}{(a^t)^{p-2}} - \frac{1}{(a^{t+1})^{p-2}}\right].$$

Summing from 0 to $t-1$ gives

$$c(p-2)t \geq (a^0)^{-(p-2)} - (a^t)^{-(p-2)},$$

and rearranging yields the upper bound

$$a^t \leq \frac{1}{\left((a^0)^{-(p-2)} - c(p-2)t\right)^{\frac{1}{p-2}}} = \frac{a^0}{\left(1 - c(p-2)(a^0)^{p-2}t\right)^{\frac{1}{p-2}}}.$$

For the lower bound, note that $a^t \leq 1$ implies $a^{t+1} \leq (1+c)a^t$, so

$$c = \frac{a^{t+1} - a^t}{(a^t)^{p-1}} = \int_{a^t}^{a^{t+1}} \frac{dx}{(a^t)^{p-1}} \leq \int_{a^t}^{a^{t+1}} \frac{(1+c)^{p-1}}{(a^{t+1})^{p-1}} dx \leq \frac{(1+c)^{p-1}}{p-2}\left[\frac{1}{(a^t)^{p-2}} - \frac{1}{(a^{t+1})^{p-2}}\right].$$

Summing and rearranging gives

$$a^t \geq \frac{a^0}{\left(1 - c(1+c)^{-(p-1)}(p-2)(a^0)^{p-2}t\right)^{\frac{1}{p-2}}}.$$

□

Lemma 39 (Matrix Bernstein inequality (Vershynin, 2018))**.** Let $X_1, \ldots, X_N$ be independent, mean-zero, symmetric $d \times d$ random matrices with $\|X_i\| \leq K$ almost surely for all i. Then for all $t \geq 0$,

$$\mathbb{P}\left[\left\|\sum_{i=1}^N X_i\right\| \geq t\right] \leq 2d \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right),$$

where $\sigma^2 = \left\|\sum_{i=1}^N \mathbb{E}[X_i^2]\right\|$ is the operator norm of the variance of the sum.

Lemma 40 (Yu et al. (2015), Corollary 3)**.** Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{d\times d}$ be symmetric matrices with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$, respectively. Fix $j \in [d]$ and assume $\min(\lambda_{j-1} - \lambda_j,\ \lambda_j - \lambda_{j+1}) > 0$, where $\lambda_0 = \infty$ and $\lambda_{d+1} = -\infty$. If $v, \hat{v} \in \mathbb{R}^d$ satisfy $\Sigma v = \lambda_j v$ and $\hat{\Sigma}\hat{v} = \hat{\lambda}_j \hat{v}$, then

$$\sin\Theta(\hat{v}, v) \leq \frac{2\|\hat{\Sigma} - \Sigma\|}{\min(\lambda_{j-1} - \lambda_j,\ \lambda_j - \lambda_{j+1})},$$

where $\Theta(\hat{v}, v) \in [0, \pi]$ denotes the angle between $\hat{v}$ and v. Furthermore, if $\hat{v}^\top v \geq 0$, then

$$\|\hat{v} - v\| \leq \frac{2^{3/2}\|\hat{\Sigma} - \Sigma\|}{\min(\lambda_{j-1} - \lambda_j,\ \lambda_j - \lambda_{j+1})}.$$