TWO-TIMESCALE FINE-TUNING PROVABLY LEARNS NEW FEATURES FOR TWO-LAYER RELU NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning pre-trained models on specialized tasks with scarce data is central to modern deep learning. Despite its empirical success, theoretical understanding of fine-tuning remains limited. We introduce a Gaussian multi-index setting to study fine-tuning from pre-trained weights, where the teacher network has $m + 1$ features, m of which are learned during pre-training and one of which must be learned during fine-tuning. For two-layer ReLU networks, we show that two-timescale training, i.e., updating the outer weights infinitely faster than the hidden ones, learns the new task-specific feature while preserving the pre-trained ones in the model representation. Moreover, only $\mathcal{O}(d)$ fine-tuning samples are required for this recovery, independently of the number of pre-trained features. In contrast, with random initialization, the same number of samples is insufficient to recover the target parameters. Our results therefore demonstrate that pre-training can induce an implicit bias with a clear statistical advantage over random initialization, enabling feature learning from scarce fine-tuning data.

1 INTRODUCTION

With the recent rise of foundation models, deep learning architectures are increasingly trained in two stages: they are first pre-trained on large and diverse datasets, and then fine-tuned on more specialized and typically much smaller datasets. This pre-training/fine-tuning paradigm has become central to modern machine learning and consistently outperforms *single-task* learning, where a model is trained from scratch directly on the target task (Kornblith et al., 2019).

Despite this empirical success, our theoretical understanding of fine-tuning remains limited. A large body of work has studied the training of neural networks in the *single-task* setting, often through simplified architectures such as two-layer ReLU neural networks (Chizat & Bach, 2018; Lyu & Li, 2020; Boursier et al., 2022). These analyses, however, typically consider training from an uninformative random initialization. Fine-tuning is fundamentally different: optimization starts from a highly structured initialization that already encodes information about related tasks. This raises a basic question: *Does pre-training merely make optimization easier, helping fine-tuning reach a solution that could also be learned from the fine-tuning data alone? Or does the pre-trained initialization provide side information that changes what can be learned from these limited data?* We show that the latter can indeed occur. Pre-training can induce a favorable implicit bias during fine-tuning, leading to better generalization even when the target task requires learning features that are absent from the pre-trained representation.

To understand this advantage, we need to explain how fine-tuning can learn the missing features from limited target data without relearning the entire representation. Existing theoretical analyses have largely focused on the preservation and reuse of pre-trained features (Shachaf et al., 2021; Kumar et al., 2022; Malladi et al., 2023). Yet empirical evidence shows that fine-tuning can also modify the representation and learn new task-specific features (Peters et al., 2019). Our work addresses this tension by showing how fine-tuning can acquire a new relevant feature while preserving the useful features already learned during pre-training. In particular, we show that such a feature can be learned from a pre-trained initialization with limited target data, even when the same feature fails to emerge from random initialization.

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

Contributions. We introduce a Gaussian multi-index setting designed to capture the key aspects of fine-tuning while remaining amenable to a precise analysis. We show that, under two-timescale training—where the outer weights are adapted much faster than the hidden weights—and in the limit where the *void* features inherited from pre-training have vanishing scale, the training dynamics converge to a limiting process in which, during an initial phase, only the void features are effectively updated. The pre-trained features therefore remain intact while new features are learned.

We then show that these limiting dynamics can recover a new neuron, corresponding to a feature unseen during pre-training, from only $\mathcal{O}(d)$ fine-tuning samples. Importantly, this sample complexity is independent of the number of relevant features already learned during pre-training. In contrast, learning the same model from scratch requires a sample complexity that grows with the number of features, highlighting a statistical advantage of fine-tuning over single-task learning.

Our analysis combines a detailed characterization of the optimization dynamics (Section 3.1) with sharp statistical concentration arguments (Section 3.2), yielding optimal statistical guarantees for the considered fine-tuning procedure.

1.1 Related work

Theory of fine-tuning. The theoretical literature on fine-tuning from pre-trained weights remains limited, especially compared with the vast literature on single-task learning. Several works (Shachaf et al., 2021; Kumar et al., 2022; Wu et al., 2022; Lee et al., 2023; Lauditi et al., 2026) study fine-tuning in linear networks and identify settings in which pre-trained features can be successfully or unsuccessfully adapted to a downstream task. In a related line of work, Shachaf et al. (2021); Malladi et al. (2023); Tomihari & Sato (2024) study fine-tuning from an NTK perspective. These analyses, however, do not capture the learning of new, task-specific features during fine-tuning, and instead focus primarily on adapting the pre-trained representation to the downstream task. With the exception of Shachaf et al. (2021), who consider the simultaneous fine-tuning of multiple layers under restrictive assumptions on both the pre-training and fine-tuning tasks, these approaches essentially reduce fine-tuning to a linear adaptation around the pre-trained weights. They therefore cannot capture the rich feature-learning dynamics that may arise during fine-tuning.We instead study fine-tuning of a two-layer ReLU network in the feature-learning regime. Our downstream task requires the pre-trained model not only to reuse features acquired during pre-training, but also to learn a genuinely new, task-specific feature. This setting allows us to show that fine-tuning from pre-trained weights can learn a feature that would not be learned from random initialization.

More recently, Lippl & Lindsey (2024); Anguita et al. (2026) have studied how the scale of the pre-trained initialization affects the optimization dynamics of fine-tuning and their implicit bias. These works focus primarily on optimization and do not provide a statistical analysis translating the resulting dynamics into improved generalization guarantees on the fine-tuning task. Jones-McCormick et al. (2025) establish a statistical advantage of fine-tuning over single-task learning in a single-index model. In their setting, however, the feature relevant to the fine-tuning task can already be recovered from the unsupervised pre-training task via PCA. By contrast, our fine-tuning procedure must learn a feature that is absent from the pre-trained representation, while preserving the useful features acquired during pre-training.

Parameter efficient fine-tuning. Given the enormous number of parameters in foundation models, parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2022) adapt only a small subset of the model parameters. We view these methods as a distinct class of algorithms, with different expressivity and applications from methods that update all model weights. They have also attracted increasing theoretical interest (see e.g., Jang et al., 2024; Dayi & Chen, 2024; Zhang et al., 2025; Kim et al., 2025). This line of work is complementary to ours, although our setting is inspired by Dayi & Chen (2024) and our fine-tuning task could be solved by LoRA, as explained in Section 3.

Learning ReLU neurons. Our theoretical setting can be viewed as an extension of the widely studied problem of learning a single ReLU neuron (Soltanolkotabi, 2017a; Yehudai & Ohad, 2020; Xu & Du, 2023). While this literature considers learning a single neuron from scratch, we study the learning of a new ReLU neuron in the presence of m neurons already learned during pre-training. Extending our setting to multiple new features would be a natural and important direction, but is outside our scope. Learning multiple neurons remains largely unsolved in fully general settings

(Zhong et al., 2017; Zhang et al., 2019). Interestingly, this literature typically trains only the hidden layer. Our analysis instead crucially relies on training both layers at different learning rates. This two layer structure yields the limiting dynamics in Equation (4), which are key to learning the new feature while preserving the pre-trained ones.

Two-timescale optimization. Two-timescale optimization, in which different layers are trained with different learning rates, has been studied in several single-task settings (Marion & Berthier, 2023; Berthier et al., 2025; Bietti et al., 2025; Barboni et al., 2025), through the limiting dynamics obtained when the outer weights are trained infinitely faster than the inner ones. In these works, the two-timescale regime is primarily motivated by analytical tractability. Differential learning rates are also widely used in practice for fine-tuning, as they help preserve pre-trained features and mitigate feature distortion (Howard & Ruder, 2018). Thus, the two-timescale regime in our work is motivated both by its relevance to practical fine-tuning and by the tractable training dynamics it provides.

1.2 Outline.

We introduce our fine-tuning setting and derive the limiting dynamics of interest in Section 2. Section 3 presents our main results, showing these dynamics recover the optimal weights both at the population level and from $n \gtrsim d$ samples. Section 4 discusses the results and their relation to the existing literature. Finally, Section 5 illustrates our theoretical findings through a toy numerical experiment.

2 Setting and limit dynamics

We consider a teacher-student setting in which data $(x_i, y_i)_{i\in[n]}$ are generated i.i.d. as follows:

$$x_i \sim \mathcal{N}(0, \mathrm{I}_d) \qquad \text{and} \qquad y_i = \textstyle\sum_{i=1}^{m+1} \sigma(x^\top w_i^\star),$$

where $\sigma(\cdot) = \max(0, \cdot)$ is the ReLU activation. The labels are generated by a two-layer teacher network of width $m+1$, whose output weights are all equal to one. On these data, we train a two-layer ReLU network parametrized by $\theta = (a, W) \in \mathbb{R}^{m+1} \times \mathbb{R}^{(m+1)\times d}$ by minimizing the empirical square loss

$$\mathcal{L}_n(\theta) \coloneqq \tfrac{1}{n} \textstyle\sum_{i=1}^n \left(f_\theta(x_i) - y_i\right)^2 \qquad \text{where} \qquad f_\theta(x) = \textstyle\sum_{i=1}^{m+1} a_i \sigma(x^\top w_i).$$

Our goal is to understand fine-tuning from a pre-trained initialization that has already learned m of the $m+1$ teacher features. Specifically, we assume that pre-training has recovered the features $w_1^\star, \dots, w_m^\star$, whereas the remaining feature $w_{m+1}^\star$ was absent from the pre-training task and must therefore be learned during fine-tuning. Motivated by this picture, we initialize the hidden weights as follows, for a small $\alpha > 0$,

$$w_i(0) = w_i^\star \text{ for any } i \in [m], \qquad w_{m+1}(0) = \alpha \breve{w} \text{ with } \breve{w} \sim \mathrm{Unif}(\mathbb{S}_{d-1}). \tag{1}$$

This initialization thus idealizes finite-time pre-training on data generated only by the first m teacher features; it is discussed further in Section 4.

2.1 Two-timescale dynamics

Fine-tuning is typically performed by gradient-based optimization of the empirical loss $\mathcal{L}_n(\theta)$. For the sake of the analysis, we model this optimization by gradient flow, which describes its continuous-time behavior in the small-step-size limit:

$$\dot{\theta}(t) \in -\partial \mathcal{L}_n(\theta(t)),$$

where ∂ stands for the Clarke subdifferential, which accounts for the non-differentiability of ReLU. In the regime studied in this work, updating all layers at the same rate can substantially modify the pre-trained features and thereby degrade their generalization properties (see Section 4). A natural alternative, considered in several previous works, is to use a larger learning rate for the output layer than for the hidden layer. This allows the output coefficients to adapt rapidly to the downstream task while limiting the displacement of the pre-trained features. In the limit where the ratio between the output- and hidden-layer learning rates tends to infinity, the dynamics converge to a two-timescale regime in which the output weights are instantaneously fitted to their optimal values,

while the hidden weights evolve on a much slower timescale. We focus on this theoretical limit, which corresponds to the following training dynamics:

$$\dot{W}(t) \in -\partial_W \mathcal{L}_n(a_n(W(t)), W(t)) \qquad \text{where} \qquad a_n(W) \coloneqq \operatorname*{argmin}_{a \in \mathbb{R}^{m+1}} \mathcal{L}_n(a, W). \tag{2}$$

The argmin above may not be unique, making the dynamics of Equation (2) ill-defined. However, with our choice of initialization and sufficiently many samples, the argmin is unique along the whole trajectory, making the dynamics well defined. In particular, the trajectory remains within a set $\mathcal{T} \subseteq \mathbb{R}^{(m+1)\times d}$ that satisfies Assumption 1 below (see Theorem 1 for details).

Assumption 1 (Conditioning of the empirical feature matrix in $\mathcal{T}$)**.** *The set $\mathcal{T}$ is compact and for any $W \in \mathcal{T}$, $\lambda_{\min}\left(\Sigma(W, \mathbf{X})\Sigma(W, \mathbf{X})^\top\right) > 0$, where $\Sigma(W, \mathbf{X}) \in \mathbb{R}^{(m+1)\times n}$ is the empirical feature matrix, given by $\Sigma(W, \mathbf{X})_{ik} = \sigma(w_i^\top x_k)$ and $\lambda_{\min}$ denotes the smallest eigenvalue function.*

2.2 Separated neuron dynamics

Studying the dynamics of Equation (2) remains extremely challenging in full generality. We therefore exploit the specific structure of the pre-trained initialization in Equation (1). Importantly, the ReLU activation is homogeneous, meaning that $\sigma(\lambda z) = \lambda\sigma(z)$ for any $\lambda \geq 0$. As a first consequence, the norm of each neuron w_i remains constant throughout training.

Lemma 1. *Let $W \in \mathbb{R}^{(m+1)\times d}$ follow Equation* (2)*, then for any $i \in [m+1]$, $\|w_i(t)\| = \|w_i(0)\|$.*

In consequence, one only has to track the normalized neurons $\mathsf{w}_i \coloneqq \frac{w_i}{\|w_i\|}$ during the dynamics. Using the homogeneity of the ReLU activation, one can easily derive the dynamics followed by the normalized neurons. If we denote by W the matrix whose i-th row is given by w_i and W follows the differential inclusion of Equation (2), then its row-normalized version satisfies the following differential inclusion for almost any $t \geq 0$:

$$\dot{\mathsf{w}}_i(t) \in -\tfrac{1}{\|w_i(0)\|^2}\partial_{\mathsf{w}_i}\mathcal{L}_n(a_n(\mathsf{W}(t)), \mathsf{W}(t)) \qquad \text{for all } i \in [m+1]. \tag{3}$$

Notably, the evolution rate of neuron i in Equation (3) scales as $\|w_i(0)\|^{-2}$. When initialized by Equation (1), as α goes to 0, the last neuron—that does not correspond to any pre-trained feature—therefore evolves on an infinitely faster timescale than the other neurons. In the limit, only this neuron evolves, while the other neurons, inherited from the pre-training procedure, remain fixed. This limiting dynamics is formalized in Proposition 1 below.

Proposition 1. *Consider $T \in \mathbb{R}_+$, any positive sequence $(\alpha_k)_{k\in\mathbb{N}}$ such that $\alpha_k \overset{k\to\infty}{\longrightarrow} 0$, and corresponding solutions W^{α_k} of Equation* (3)*, where $W^{\alpha_k}(0)$ is initialized by Equation* (1) *with scale α_k. Suppose that for k large enough, the trajectory $\left(\mathsf{W}^{\alpha_k}(\alpha_k^2 t)\right)_{t\in[0,T]}$ is included within some set $\mathcal{T}$, independent of k, satisfying Assumption 1. Then, one can extract a subsequence $\alpha_{\varphi(k)}$ such that, as $k \to \infty$, $(\mathsf{W}(\alpha_{\varphi(k)}^2 t))_{t\geq 0}$ converges uniformly towards $\mathsf{W}^\circ(t)$ on $[0, T]$, where W° is a solution of the following differential inclusion for almost any $t \geq 0$, with $\mathsf{w}^\circ_{m+1}(0) = \breve{w}$,*

$$\begin{cases} \mathsf{w}_i^\circ(t) = \frac{w_i^\star}{\|w_i^\star\|} & \text{for any } i \in [m], \\ \dot{\mathsf{w}}^\circ_{m+1}(t) \in -\partial_{\mathsf{w}_{m+1}}\mathcal{L}_n(a_n(\mathsf{W}^\circ(t)), \mathsf{W}^\circ(t)). \end{cases} \tag{4}$$

Proposition 1 implies that, as $\alpha \to 0$, the trajectories $\mathsf{W}^\alpha(\alpha^2\cdot)$ converge uniformly, on any compact of $\mathbb{R}_+$, towards the set of solutions of the limit differential inclusion (4). Importantly, the convergence is only to a set of trajectories, since Equation (4) may admit multiple solutions due to the non-differentiability of the loss. Characterizing these different solutions is the focus of Section 3.

Proposition 1 also requires the trajectories to remain in a set $\mathcal{T}$ satisfying Assumption 1. In Section 3, we show that, with high probability, all trajectories of the limiting process remain in such a set (see Theorem 2). Lemma 14, given in Appendix D, then implies that, for α sufficiently small, the trajectories $\mathsf{W}^\alpha(\alpha^2\cdot)$ also remain in a set $\mathcal{T}$ satisfying Assumption 1.

3 Two-timescale small initialization dynamics

The two-timescale dynamics in Equation (2) are difficult to analyze directly. In the limit $\alpha \to 0$, however, they reduce to the simpler dynamics in Equation (4), which we study throughout the remainder of the paper. In this limiting regime, the pre-trained features w_i°, $i \in [m]$, remain fixed, and

only the new feature evolves. The dynamics are nevertheless nontrivial because the corresponding output weights a_i° of the pre-trained features continue to evolve and must be carefully controlled. In particular, as shown in the proof of Lemma 2, these output weights are not initially equal to their optimal value, namely 1, and reach this value only at convergence. Early in training, they instead compensate for the fact that the $(m+1)$-th feature has not yet been learned. This transient compensation also explains why the *Linear Probe then Fine-Tune* method (see Section 4) does not achieve zero test loss in our setting, as confirmed experimentally in Section 5.

For the same reason, the dynamics are more complex than simply freezing the first m neurons—both their inner and output weights—and training only the $(m+1)$-th neuron. Such a procedure would be closer to LoRA fine-tuning and, in our setting, would reduce to learning a single ReLU neuron. To keep the analysis tractable, we make the following assumption on the teacher features.

Assumption 2. *The teacher features $(w_i^\star)_{i\in[m+1]}$ form an orthonormal system of $\mathbb{R}^d$.*

Assumption 2 simplifies the analysis by yielding a closed-form expression for $a_n(W)$, at least in the population limit studied in Section 3.1. Orthogonality of features is a standard assumption in the literature on learning from multi-neuron teacher networks (Safran & Shamir, 2018; Simsek et al., 2023; Dayi & Chen, 2024). We discuss the dependence of our results on this assumption in Section 4.

3.1 Warm up: gradient flow on population loss

Before studying the training dynamics of Equation (4), we first focus on the population loss case, where fine-tuning is done with access to an infinite number of data points. More precisely, we here consider the population loss dynamics given by the following ODE[1], for any $t \geq 0$:

$$w_i(t) = w_i^\star \text{ for any } i \in [m], \qquad \text{and} \qquad \dot{w}_{m+1}(t) \in -\nabla_{w_{m+1}} \mathcal{L}(a(W(t)), W(t)); \tag{5}$$

where $w_{m+1}(0) = \breve{w}$, $a(W) \coloneqq \operatorname{argmin}_{a\in\mathbb{R}^{m+1}} \mathcal{L}(a, W)$ and

$$\mathcal{L}(a, W) \coloneqq \mathbf{E}_{x\sim\mathcal{N}(0,\mathrm{I}_d)} \left[\left(\sum_{i=1}^{m+1} a_i(t)\sigma(x^\top w_i(t)) - \sum_{i=1}^{m+1} \sigma(x^\top w_i^\star) \right)^2 \right].$$

Although this simpler setting avoids any statistical consideration, it already provides a good grasp of what is happening from an optimization point of view. We show in this section that w_{m+1} converges to $w_{m+1}^\star$ in the population loss dynamics. For that, we merely need to track the angular deviation between these two vectors. In the population case, explicit computations can be done, relying on the arc-cosine kernel formula (Cho & Saul, 2009), for any $u, v \in \mathbb{S}_{d-1}$:

$$\mathbf{E}_{x\sim\mathcal{N}(0,\mathrm{I}_d)} \left[\sigma(x^\top u)\sigma(x^\top v) \right] = \frac{1}{2\pi} \left(\sqrt{1-(u^\top v)^2} + (\pi - \arccos(u^\top v))u^\top v \right).$$

In particular, for some functions H and D depending on this formula and defined precisely by Equation (9) in Appendix C.2, one can express precisely the loss and its gradient.

Lemma 2. *Let Assumption 2 hold and $w_{m+1} \in \mathbb{S}_{d-1}$ such that $D(w_{m+1}) > 0$. Then, noting $W = [w_1^{\star\top}, \dots, w_m^{\star\top}, w_{m+1}^\top]$, $a(W)$ is uniquely defined and satisfies $a(W)_{m+1} = \frac{H(w_{m+1})}{D(w_{m+1})}$. Moreover, the loss and its subdifferential satisfy*

$$\mathcal{L}(a(W), W) = D(w_{m+1}^\star) - \frac{H(w_{m+1})^2}{D(w_{m+1})},$$

$$\nabla_{w_{m+1}} \mathcal{L}(a(W), W) = -\frac{H(w_{m+1})}{D(w_{m+1})} \left(\mathrm{I}_d - w_{m+1}w_{m+1}^\top \right) \Psi(w_{m+1}),$$

where $\Psi(w) \coloneqq 2\nabla H(w) - \frac{H(w)}{D(w)} \nabla D(w)$.

Lemma 2 then provides an intricate closed-form expression of the training dynamics, in terms of the function H and D. From there, a careful analysis can be done to show convergence of the learned parameter towards the optimal choice, as given by Theorem 1 below.

[1] In the population case, the loss becomes differentiable.

Theorem 1. *Let Assumption 2 hold and initialize as* $w_{m+1}(0) \sim \mathrm{Unif}(\mathbb{S}_{d-1})$. *There exist positive universal constants* m_0, c *and* C *such that for any* $\delta \in (0, 1/2)$, *if* $m \geq m_0$ *and* $d \geq Cm^2 \ln(1/\delta)$, *then with probability at least* $1 - \delta$ *over the initialization, the solution of Equation* (5) *satisfies:*

$$w_{m+1}(t)^\top w^\star_{m+1} \geq 1 - e^{-c(t-Cm)} \qquad \textit{for any } t \geq Cm.$$

Sketch of proof. The key ingredient is Lemma 4 in Appendix C.3. Let $P_m w$ denote the orthogonal projection of w onto $\mathrm{Span}(w^\star_1, \dots, w^\star_m)$. The lemma relies on the orthogonal features assumption and shows that, as long as $\|P_m w_{m+1}(t)\|$ remains below some constant threshold,

1. $D(w_{m+1}(t)) \gtrsim 1$;
2. $\frac{\mathrm{d}}{\mathrm{d}t}(w_{m+1}(t)^\top w^\star_{m+1}) \gtrsim \frac{H(w_{m+1}(t))}{D(w_{m+1}(t))}\left(1 - w_{m+1}(t)^\top w^\star_{m+1}\right)$;
3. $\frac{\mathrm{d}}{\mathrm{d}t}\|P_m w_{m+1}(t)\| \lesssim \frac{H(w_{m+1}(t))}{D(w_{m+1}(t))}\left(\frac{1}{\sqrt{m}} + \|P_m w_{m+1}(t)\|\right)$.

These inequalities reveal the key mechanism behind the result. As long as $\frac{H(w_{m+1}(t))}{D(w_{m+1}(t))} > 0$, the component of $w_{m+1}(t)$ along $w^\star_{m+1}$ grows quickly towards 1, while its component in the span of the previously learned directions $w^\star_1, \dots, w^\star_m$ evolves at a much slower rate, of order $\frac{1}{\sqrt{m}} + \|P_m w_{m+1}(t)\|$. In particular, when m is sufficiently large and $\|P_m w_{m+1}(0)\|$ is sufficiently small, the undesirable component $P_m w_{m+1}(t)$ grows slowly enough to remain small throughout the trajectory. This ensures that the conditions of Lemma 4 continue to hold.

It remains to verify that the ratio H/D is indeed positive along the trajectory. Lemma 4 also implies that, provided

$$\|P_m w_{m+1}(0)\| \lesssim 1/\sqrt{m} \qquad \text{and} \qquad w_{m+1}(0)^\top w^\star_{m+1} \gtrsim -1/m, \tag{6}$$

we have $\frac{H(w_{m+1}(0))^2}{D(w_{m+1}(0))} \gtrsim \frac{1}{m^2}$. When $d \gtrsim m^2$, Equation (6) holds with high probability. Moreover, by Lemma 2, the quantity $\frac{H(w_{m+1})^2}{D(w_{m+1})}$ is, up to an additive constant, the opposite of the population loss. Since the population loss is non-increasing along the gradient flow, it follows that $\frac{H(w_{m+1}(t))^2}{D(w_{m+1}(t))}$ remains bounded away from zero for all $t \geq 0$. Together with item 1 above, this implies that $\frac{H(w_{m+1}(t))}{D(w_{m+1}(t))}$ remains positive (by continuity) and bounded away from zero throughout the trajectory.

The preceding argument therefore applies for all $t \geq 0$: the component along the new direction $w^\star_{m+1}$ grows substantially faster than the component along the previously learned directions. A more precise analysis of the ratio H/D then yields the exact convergence rate stated in Theorem 1. □

Theorem 1 shows that, when trained on the full population distribution, the learned feature $w_{m+1}(t)$ eventually converges towards the optimal parameter $w^\star_{m+1}$ at a linear rate. This convergence can nevertheless be substantially delayed by the initialization phase. Indeed, the convergence rate contains a time shift of order Cm. This reflects the slow dynamics at initialization: the update of $w_{m+1}(t)$ is initially only of order $1/m$ (due to the ratio $\frac{H(w_{m+1})}{D(w_{m+1})}$ at initialization), and the feature therefore requires a time of order m to escape this slow-evolution regime. Once this initial phase is overcome, the dynamics enter the linear convergence regime described by Theorem 1.

A notable difference with classical analyses of learning a single ReLU neuron is that, in our setting, convergence requires a suitable initialization. This is precisely what leads to the dimensional requirement $d \gtrsim m^2$—see Section 4 for further discussion. The reason is that the two-timescale dynamics of Equation (5) are not globally attracted to $w^\star_{m+1}$. Instead, the dynamics possess spurious local minima, distinct from the desired solution $w^\star_{m+1}$. Their existence is implied by Lemma 2, but their precise location is not characterized by our analysis. The initialization condition thus ensures that $w_{m+1}(0)$ lies outside the attraction basins of these spurious minima. Once this condition is satisfied, the dynamics are driven towards the desired solution $w^\star_{m+1}$.

3.2 Gradient flow on empirical loss

Let us now go back to the empirical loss case, through the study of the (possibly multiple) solutions W° of the differential inclusion given by Equation (4). In the empirical case, we can still derive

closed form expressions of the loss and its subdifferential, replacing the key quantities H and D, by their empirical counterpart defined in Equation (33). Given a large enough number of training samples, the different quantities of interest concentrate sufficiently close to their population counterpart, so that we can derive a similar analysis of the training dynamics. It then yields our main theorem.

Theorem 2. *Let Assumption 2 hold and initialize as* $\mathsf{w}^\circ_{m+1}(0) \sim \mathrm{Unif}(\mathbb{S}_{d-1})$. *There exist positive universal constants* m_0, c *and* C *such that for any* $\delta \in (0, 1/2)$, *if* $m \geq m_0$, $d \geq Cm^2 \ln(1/\delta)$ *and* $n \geq C\left(d + \ln(1/\delta)\right)$ *then with probability at least* $1 - \delta$ *over both the data and initialization, any solution of Equation* (4) *satisfies:*

$$\mathsf{w}^\circ_{m+1}(t)^\top w^\star_{m+1} \geq 1 - e^{-c(t-Cm)} \qquad \textit{for any } t \geq Cm.$$

Moreover under the same random event, there exists a set $\mathcal{T}$ *satisfying Assumption 1 such that* $\mathsf{W}^\circ(t) \in \mathcal{T}$ *for any* $t \geq 0$.

Sketch of proof. From an optimization perspective, the proof follows the same lines as the one of Theorem 1. The main additional challenge is to establish sufficiently tight concentration bounds for the relevant empirical quantities around their population counterparts. The main concentration result is provided by Proposition 2 in Appendix D.3. Similarly to the population case, it shows that, under the stated sample complexity, with high probability and as long as $\|P_m \mathsf{w}^\circ_{m+1}(t)\|$ is small enough,

1. $D_n(\mathsf{w}^\circ_{m+1}(t)) \gtrsim 1$;
2. $\frac{\mathrm{d}}{\mathrm{d}t}(\mathsf{w}^\circ_{m+1}(t)^\top w^\star_{m+1}) \gtrsim \frac{H_n(\mathsf{w}^\circ_{m+1}(t))}{D_n(\mathsf{w}^\circ_{m+1}(t))}\left(1 - \mathsf{w}^\circ_{m+1}(t)^\top w^\star_{m+1}\right)$;
3. $\frac{\mathrm{d}}{\mathrm{d}t}\|P_m \mathsf{w}^\circ_{m+1}(t)\| \lesssim \frac{H_n(\mathsf{w}^\circ_{m+1}(t))}{D_n(\mathsf{w}^\circ_{m+1}(t))}\left(\frac{1}{\sqrt{m}} + \|P_m \mathsf{w}^\circ_{m+1}(t)\| + \varepsilon\right)$;

for a small $\varepsilon > 0$. The proof of these three statements relies on two complementary concentration arguments. First, we establish uniform concentration on the sphere $\mathbb{S}_{d-1}$ for the different empirical quantities involved. These bounds directly yield points 1 and 3, and also point 2 whenever w°_{m+1} lies outside a neighborhood of $w^\star_{m+1}$ (see Appendix D.3.1). The required uniform concentration bounds are obtained from empirical process concentration arguments, as detailed in Appendix E.1, i.e., with tight chaining and VC dimension techniques (we refer to van der Vaart & Wellner, 1996, for a useful introduction to those techniques).

The second argument is needed to control the dynamics when w°_{m+1} enters a neighborhood of $w^\star_{m+1}$. Borrowing techniques from Soltanolkotabi (2017a), we obtain the required lower bound in point 2 in this regime as well (see Appendix D.3.2). Once these properties are established with high probability, the remainder of the proof follows the same lines as the one of Theorem 1, with the additional ingredient of a sharper concentration bound at initialization, detailed in Appendix D.3.4. □

Interestingly, Theorem 2 shows that perfectly recovering the optimal parameter $w^\star_{m+1}$ and achieving zero test loss only requires a sample complexity of $n \gtrsim d$, independent of m. This is particularly striking when compared with learning all $m+1$ features from scratch: from an information-theoretic perspective, this would require at least $\Omega(md)$ samples. In contrast, the optimization considered here effectively leverages the information already encoded in the pre-trained weights. As a result, it only needs to learn the single new feature $w^\star_{m+1}$, leading to a sample complexity of the same order as that required for learning a single neuron.

4 Discussion and limitations

In this section, we discuss the limitations of our results, and also compare the two-timescale optimization scheme considered here to other standard methods.

Comparison with other fine-tuning methods. The most natural fine-tuning strategy is *full fine-tuning*, which adapts all model parameters using gradient-based methods with the same learning rate across layers. Despite its simplicity and strong empirical performance, full fine-tuning can substantially *distort* pre-trained features (Kumar et al., 2022), thereby losing the statistical benefits of pre-training and harming generalization. We observe the same phenomenon in our toy setting (see Section 5). At the opposite extreme, *linear probing* freezes the inner weights and adapts only the last

layer. This prevents feature distortion and can be particularly effective with limited fine-tuning data, but at the cost of expressivity: it cannot learn tasks requiring new or substantially modified features.

Several intermediate strategies aim to preserve pre-trained features while retaining this expressivity. *Two-timescale fine-tuning*, which we study here, assigns smaller learning rates to inner layers than to outer layers (Howard & Ruder, 2018). We show that this induces a favorable implicit regularization: new features can be learned while the pre-trained representation remains essentially unchanged. The mechanism is particularly striking in the limiting regime where the *void* features vanish at initialization: their small magnitude is compensated by large outer weights, amplifying their gradients and causing them to evolve much faster than the pre-trained features.

A related strategy is *Linear Probe then Finetune* (LP-FT), which first performs linear probing and then switches to full fine-tuning (Kumar et al., 2022). While effective in practice, LP-FT only delays feature distortion rather than fully preventing it. Indeed, even with infinitely many data points, linear probing does not initially assign the optimal output weights to the pre-trained features: in our setting, the weights satisfy $a_i(0) \neq 1$ for $i \in [m]$ (see proof of Lemma 2). To learn $w^\star_{m+1}$, the output weights must therefore be readjusted. However, because LP-FT does not preserve the two-timescale structure during this second phase, this readjustment is accompanied by a distortion of the pre-trained features. In our experiments, this distortion is comparable to that of full fine-tuning and is sufficient to deteriorate test performance (see Section 5). Finally, ℓ_2 *distance to starting point* (L2-SP) explicitly penalizes deviations from the pre-trained weights (Xuhong et al., 2018). This regularization directly preserves the pre-trained representation and is particularly well suited to our setting, as illustrated empirically in Section 5 and discussed in detail in Appendix A.1. However, in more general settings, this regularization can also hinder the adaptation of the pre-trained features, for instance when the new task requires small but non-negligible modifications of the pre-trained weights.

Requirement $d \gtrsim m^2$. As explained after Theorem 1, the condition $d \gtrsim m^2$ is primarily needed to ensure that, with high probability, the $(m+1)$-th neuron is initialized within the attraction basin of the target weight $w^\star_{m+1}$. In lower dimensions, such a guarantee cannot hold with high probability for a single randomly initialized neuron. Overparameterization could potentially circumvent this limitation: by training multiple new neurons, one could hope that at least one is initialized in the correct attraction basin, thereby relaxing the dimensional requirement. Although similar dynamics appear to persist in the overparameterized setting (see Appendix A.3), making this argument rigorous would require a substantially more intricate analysis, as the dynamics become considerably more challenging to characterize. This difficulty is already apparent in the single-task setting for learning a single ReLU (Xu & Du, 2023).

Pre-trained initialization of the weights. A key assumption of our work concerns the structure of the initialization: we assume that the first m neurons are initialized at the optimal weights $w^\star_1, \dots, w^\star_m$, while the $(m+1)$-th neuron is initialized with a small norm and a uniformly random direction. This setting is inspired by pre-training an overparameterized model on a teacher network with features $w^\star_1, \dots, w^\star_m$ (see, e.g., Bietti et al., 2025). If pre-training succeeds, one would expect the learned network to contain features closely aligned with these optimal weights, while overparameterization may leave additional *void* features that are not aligned with any of them. Assuming that such void features have small norm is also natural, as loss minimization tends to suppress unused features, and this effect is further reinforced by weight decay (Han et al., 2015).

We thus view Equation (1) as an idealized model of weights obtained after pre-training, chosen to make the resulting dynamics amenable to analysis. In Appendix A.4, we complement this idealized setting with experiments initialized from an actually pre-trained model. The resulting dynamics are consistent with our theoretical findings, supporting the relevance of our idealized initialization.

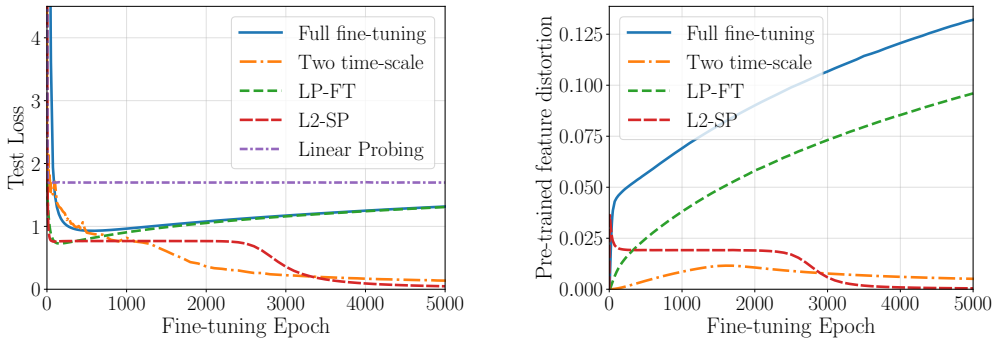
Orthogonality of features. Assumption 2 is primarily made for analytical convenience. In particular, orthogonality provides a simple and explicit expression for the population Gram matrix G, whose (i,j)-th entry is $G_{i,j} = \mathbf{E}_x\big[\sigma(x^\top w^\star_i)\sigma(x^\top w^\star_j)\big]$. This assumption can readily be relaxed to allow features with different norms and nonzero correlations. However, the analysis still requires an analogue of Item 2 in Lemma 4, which imposes an angular condition on the different features, albeit a substantially weaker one than orthogonality. At a high level, Item 2 ensures that, along the relevant portion of the dynamics, the signal associated with the target feature $w^\star_{m+1}$ dominates the signal associated with the pre-trained features.

5 EXPERIMENTS

We complement our theoretical results with numerical experiments in a slightly more general setting, where the teacher weights are drawn independently from a standard Gaussian distribution and are therefore neither normalized nor pairwise orthogonal. Before fine-tuning, we initialize the network by setting the first m weights to the corresponding optimal weights $w_1^\star, \dots, w_m^\star$, while initializing the $(m+1)$-th weight with a uniformly random direction and a small norm. With problem parameters $m = 20$, $d = 500$ and $n = 4000$, we then fine-tune the two-layer ReLU network using the different algorithms described in Section 4. Experimental details and additional experiments are provided in Appendix A.

Figure 1a shows the evolution of the test loss during fine-tuning for the different methods. LP-FT performs well during the initial epochs, but its performance subsequently deteriorates as the model starts overfitting the fine-tuning data and, at the same time, degrading its pre-trained features (see below). Full fine-tuning exhibits a similar, but more pronounced, behavior: without an initial linear probing phase, it starts distorting the pre-trained features from the beginning of training, which leads to even worse generalization. The other methods are less affected by feature distortion and overfitting. Linear probing cannot reduce the test loss below a certain threshold because of its limited expressivity. In contrast, both two-timescale and L2-SP perform much better, although they require substantially more epochs to reach their best performance. L2-SP performs well as we regularize only the hidden layer, a choice that is particularly well suited to our setting; regularizing all parameters substantially degrades performance. We provide additional discussion and details on L2-SP in Appendix A.1.

Figure 1a also highlights a trade-off between statistical performance and optimization speed: two-timescale and L2-SP achieve the best generalization by preserving the pre-trained features while learning the new feature, but might require more training epochs to do so. Importantly, all methods (except linear probing) eventually achieve nearly zero training loss. This indicates that their differences in test performance do not stem from their ability to minimize the training objective, but rather from the different forms of implicit regularization induced by their respective optimization schemes.



(a) Test loss.

(b) Distortion of pre-trained features.

Figure 1: Fine-tuning from pre-trained weights with different methods.

Figure 1b quantifies the distortion of the pre-trained features induced by fine-tuning. *Pre-trained feature distortion* is defined as the mean squared error of the fine-tuned model on data generated by the pre-training teacher, after re-optimizing the output weights by linear probing. Evaluating the pre-training error after this re-optimization isolates the quality of the learned features: it measures how well the fine-tuned representation still captures the features learned during pre-training.

Full fine-tuning and LP-FT lead to comparable levels of feature distortion. While feature distortion under full fine-tuning is well documented, LP-FT is generally expected to mitigate this effect. As explained in Section 4, however, LP-FT only delays feature distortion rather than preventing it, which is precisely what we observe in our experiments. On the other hand, two-timescale fine-tuning induces almost no feature distortion, suggesting that the limiting dynamics of Proposition 1 accurately capture the fine-tuning dynamics even for finite initialization scales and learning-rate ratios. L2-SP also maintains low feature distortion, consistent with its explicit regularization toward the pre-trained hidden weights.

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing the writing, coding, and identifying relevant literature references. Generative AI tools were also used to explore some mathematical arguments. All mathematical proofs presented in the paper were developed, verified, and written by the human authors. Where AI tools provided useful suggestions, the authors critically evaluated, adapted, clarified, and improved them before incorporating the resulting arguments into the paper.

REPRODUCIBILITY STATEMENT

Experimental details are given in Appendix A and are extensive enough for the reader to reproduce our experiments. Moreover, the code is provided in the supplementary material.

REFERENCES

Nicolas Anguita, Francesco Locatello, Andrew M Saxe, Marco Mondelli, Flavia Mancini, Samuel Lippl, and Clémentine Carla Juliette Dominé. A theory of how pretraining shapes inductive bias in fine-tuning. In *Forty-third International Conference on Machine Learning*, 2026.

Raphaël Barboni, Gabriel Peyré, and François-Xavier Vialard. Ultra-fast feature learning for the training of two-layer neural networks in the two-timescale regime. *arXiv preprint arXiv:2504.18208*, 2025.

Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *Foundations of Computational Mathematics*, 25(5):1627–1710, 2025.

Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow part i: General properties and two-timescale learning. *Communications on Pure and Applied Mathematics*, 78(12):2354–2435, 2025.

Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188(1):19–51, 2021.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *Advances in Neural Information Processing Systems*, 35:20105–20118, 2022.

Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, 2011.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, volume 22, 2009.

Arif Kerem Dayi and Sitan Chen. Gradient dynamics for low-rank fine-tuning beyond kernels. *arXiv preprint arXiv:2411.15385*, 2024.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339, 2018.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Uijeong Jang, Jason D. Lee, and Ernest K. Ryu. LoRA training in the NTK regime has no spurious local minima. In *Forty-first International Conference on Machine Learning*, 2024.

Taj Jones-McCormick, Aukosh Jagannath, and Subhabrata Sen. Provable benefits of unsupervised pre-training and transfer learning via single-index models. In *Forty-second International Conference on Machine Learning*, 2025.

Junsu Kim, Jaeyeon Kim, and Ernest K. Ryu. LoRA training provably converges to a low-rank global minimum or it fails loudly (but it probably won't fail). In *Forty-second International Conference on Machine Learning*, 2025.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Clarissa Lauditi, Blake Bordelon, and Cengiz Pehlevan. Transfer learning in infinite width feature learning networks. In *International Conference on Learning Representations*, volume 2026, pp. 53982–54028, 2026.

Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. In *The Eleventh International Conference on Learning Representations*, 2023.

Samuel Lippl and Jack Lindsey. Inductive biases of multi-task learning and finetuning: multiple regimes of feature reuse. *Advances in Neural Information Processing Systems*, 37:118745–118776, 2024.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *International Conference on Machine Learning*, pp. 23610–23641. PMLR, 2023.

Pierre Marion and Raphaël Berthier. Leveraging the two-timescale regime to demonstrate convergence of neural networks. *Advances in Neural Information Processing Systems*, 36:64996–65029, 2023.

Matthew E Peters, Sebastian Ruder, and Noah A Smith. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pp. 7–14, 2019.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International conference on machine learning*, pp. 4433–4441. PMLR, 2018.

Gal Shachaf, Alon Brutzkus, and Amir Globerson. A theoretical analysis of fine-tuning with linear teachers. *Advances in Neural Information Processing Systems*, 34:15382–15394, 2021.

Berfin Simsek, Amire Bendjeddou, Wulfram Gerstner, and Johanni Brea. Should under-parameterized student networks copy or average teacher weights? *Advances in Neural Information Processing Systems*, 36:78028–78068, 2023.

Mahdi Soltanolkotabi. Learning relus via gradient descent, 2017a.

Mahdi Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization, 2017b.

Akiyoshi Tomihari and Issei Sato. Understanding linear probing then fine-tuning language models from ntk perspective. *Advances in Neural Information Processing Systems*, 37:139786–139822, 2024.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, NY, 1 edition, 1996. ISBN 978-1-4757-2545-2. doi: 10.1007/978-1-4757-2545-2.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. The power and limitation of pretraining-finetuning for linear regression under covariate shift. *Advances in Neural Information Processing Systems*, 35:33041–33053, 2022.

Weihang Xu and Simon Du. Over-parameterization exponentially slows down gradient descent for learning a single neuron. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 1155–1198. PMLR, 2023.

LI Xuhong, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International conference on machine learning*, pp. 2825–2834. PMLR, 2018.

Gilad Yehudai and Shamir Ohad. Learning a single neuron with gradient methods. In *Conference on Learning Theory*, pp. 3756–3786. PMLR, 2020.

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1524–1534. PMLR, 2019.

Yuanhe Zhang, Fanghui Liu, and Yudong Chen. LoRA-one: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently. In *Forty-second International Conference on Machine Learning*, 2025.

Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *International conference on machine learning*, pp. 4140–4149. PMLR, 2017.

Appendix

Table of Contents

A Additional experiments and experimental details **13**
A.1 Experimental details 13
A.2 Cosine similarities of learned features 14
A.3 Overparameterized student 15
A.4 Fine-tuning from real pre-training 15

B Proofs of Section 2 **16**
B.1 Proof of Lemma 1 16
B.2 Proof of Proposition 1 17

C Proofs of Section 3.1 **18**
C.1 Notations and preliminaries 18
C.2 Population dynamics 19
C.3 Key inequalities 21
C.4 Proof of Theorem 1 25

D Proof of Section 3.2 **26**
D.1 Additional notations and preliminaries 26
D.2 Empirical loss: expression and gradient 27
D.3 Concentration of key quantities 28
D.3.1 Uniform concentration bounds 28
D.3.2 Concentration near $w^\star_{m+1}$ 32
D.3.3 Proof of Proposition 2 38
D.3.4 Concentration at initialization 39
D.4 Proof of Theorem 2 41

E Additional lemmas **42**
E.1 Useful concentration bounds 42

A Additional experiments and experimental details

A.1 Experimental details

For the experimental setup of Section 5, we generate the ground truth features i.i.d. as $w_i^\star \mathcal{N}(0, \mathrm{I}_d)$ and both pre-training teacher (used to evaluate feature distortion) and fine-tuning teacher are two-layer ReLU networks, where the outer weights are also drawn i.i.d. as standard Gaussian. The m first features of the network are initialized as the ground truth features $w_i^\star$, while the $m+1$ feature is drawn as a random centered Gaussian of covariance $10^{-6}\mathrm{I}_d$. Output weights are also initialized at random, with unitary Gaussian for the m first ones, and a Gaussian of variance 10^{-6} for the last one.

All the models are trained via stochastic gradient descent (SGD) over batches of size 200. The learning rates are tuned among three or four values. For two-timescale, the learning rate associated to the output layer is 10^{-2}, and 10^{-5} for the hidden layer. The regularization strength of $L2-SP$ is chosen as the best value among a dozen tested values. The Python code is available in the supplementary material.

L2-SP. An important detail is that our implementation of L2-SP formulation does not regularize the squared distance between all model parameters and their pre-trained values. Instead, we only regularize the hidden parameters. This algorithmic choice is particularly well suited to the problem at hand and leads to substantially better performance for L2-SP.

This implementation choice is in fact crucial to the strong performance of L2-SP in our different experiments. When the new features are initialized close to $\mathbf{0}$, L2-SP can move them toward essentially any direction at little regularization cost. The corresponding output weights can then be scaled accordingly, allowing the model to add a new feature without incurring a significant regularization penalty.

Random initialization. We also trained the model from random initialization, corresponding to single-task learning without pre-trained weights. As expected, the model achieves nearly zero training loss but a much larger test loss, illustrating that the benefit of pre-training in our setting is statistical rather than purely optimization-based. The resulting test loss falls outside the range of the figures and is therefore not shown.

A.2 Cosine similarities of learned features

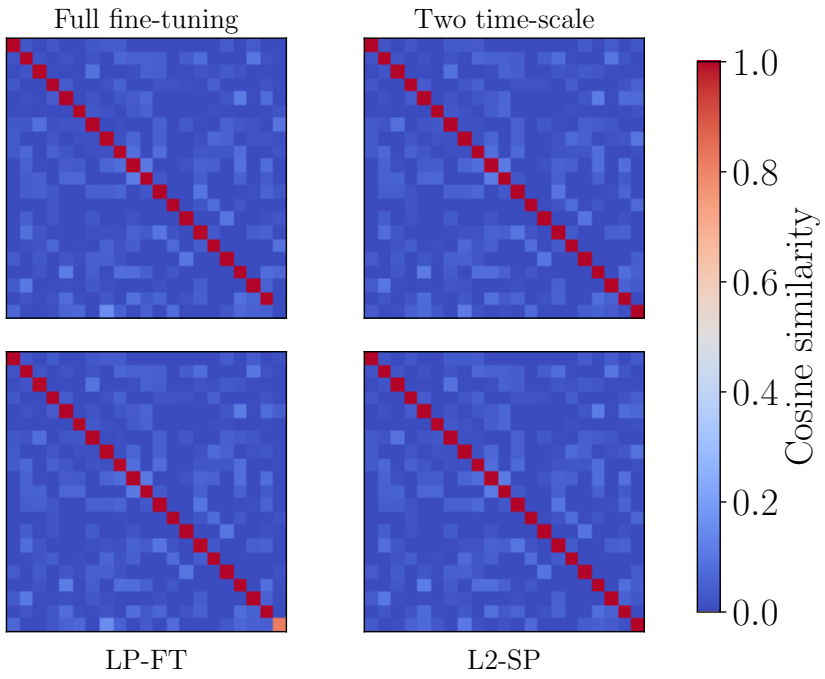


Figure 2: Cosine similarities between learned features and ground truth.

Figure 2 illustrates the cosine similarities between the learned features and ground truth ones. Precisely, each colormap represents a matrix C whose components are given by

$$C_{ij} = \frac{w_j^\top w_i^\star}{\|w_j\| \cdot \|w_i^\star\|},$$

where the w are the hidden weights of the represented model, after 5000 training epochs. The columns are also permuted here for clarity of the figure. The setting is the same as in Figures 1a and 1b. Interestingly, it confirms that after fine-tuning, both two-timescale and L2-SP perfectly learned the new feature (given by the last red square on the diagonal), while other algorithms might not have especially learned it.

We here truncate the values below 0 for a better visibility, as no value is much smaller than 0.

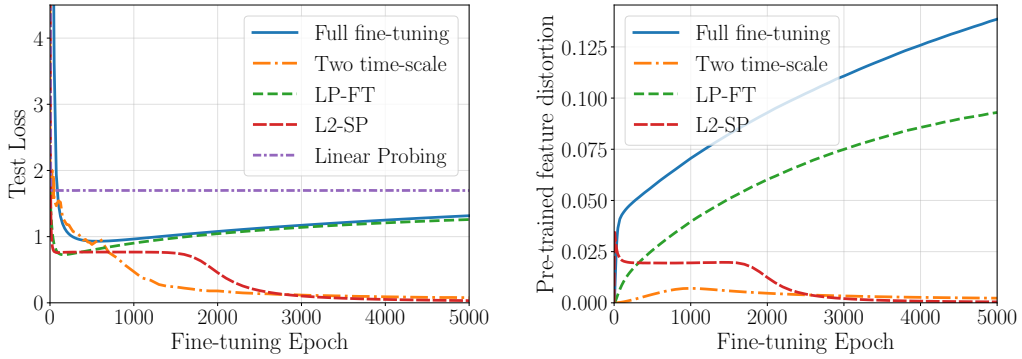
LP-FT has learned the new feature up to some extent. However, this is only true at the late stage of fine-tuning. If one had stopped the fine-tuning of LP-FT at epoch 160 (where its test loss is minimal in Figure 1a), this last diagonal square would also be blue, i.e., the new feature was not correctly learned at that time. If the model performance degrades although it learns this new feature between epoch 160 and 5000, it is thus because the pre-trained features are distorted in the meantime. The features do not appear as distorted in Figure 2 for both full fine-tuning and LP-FT. It is because

even a tiny distortion, i.e., having a cosine similarity of approximately 99.7% is sufficient enough to degrade the predictive performance, as observed in Figure 1b. Thus, this kind of colormap is not necessarily the best to visualise feature distortion, but is very helpful to see whether the new feature has somehow been learned by the model.

A.3 Overparameterized student

In this section, we run the same experiment as in Section 5, with the exception that the student network is overparameterized. To clarify, we consider the exact same 4000 fine-tuning data samples as in Section 5 (see details in Appendix A.1). However now, the learned model contains 100 neurons instead of 21: the first $m = 20$ neurons are again initialized by the ground truth weights $w_1^\star, \dots, w_m^\star$. And the remaining 80 neurons are, again, initialized as i.i.d. random centered Gaussian of variance 10^{-6}.

Figure 3 illustrates our findings in this setup. The empirical observations are very similar to the ones of Section 5, suggesting that our theoretical findings should also hold when the student network is overparameterized.



(a) Test loss.

(b) Distortion of pre-trained features.

Figure 3: Fine-tuning from pre-trained weights with an overparameterized model.

A.4 Fine-tuning from real pre-training

Until now, we considered fine-tuning from weights initialized according to Equation (1). As discussed above, it idealizes the weights obtained after pre-training on data generated by a teacher given by the m first ground truth parameters. This section empirically illustrates that similar fine-tuning behaviors happen, when fine-tuning on such a pre-trained model. Here is our empirical setup.

Model pre-training. We consider pre-training data of dimension $d = 500$, and represented by a teacher of $m_{\text{pre}} = 20$ neurons, $w_1^\star, \dots, w_{m_{\text{pre}}}^\star$ which are drawn i.i.d. as standard Gaussian. The pre-training data is then generated i.i.d. as follows:

$$x \sim \mathcal{N}(0, \mathrm{I}_d), \qquad y = \sum_{i=1}^{m_{\text{pre}}} \sigma(x^\top w_i^\star).$$

We then consider an overparameterized two-layer ReLU network of 100 neurons, initialized i.i.d. as centered Gaussian of variance 0.01. We then pre-train the model via online SGD with an initial learning rate of 10^{-3}, weight decay of 10^{-2}, batches of size 256 for $T = 10^6$ steps. To stabilize the pre-training dynamics, we also use the standard PyTorch cosine learning rate scheduler.

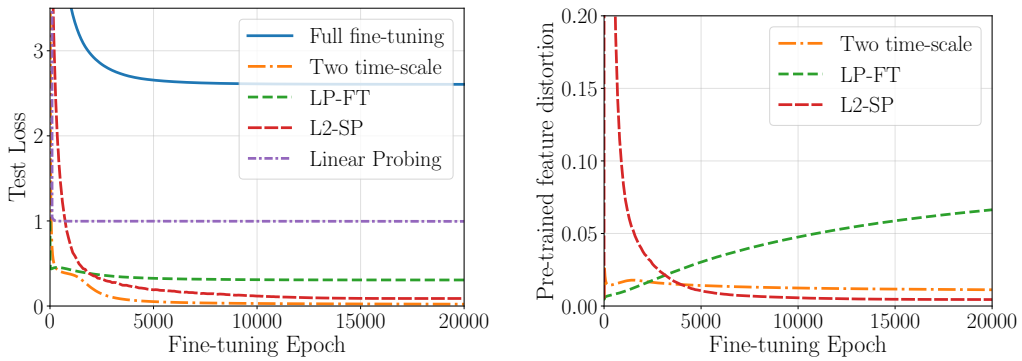
Fine-tuning. We then fine-tune the model on a new fine-tuning task defined as follows. We first generate a teacher, by drawing at random $m_{\text{fine}} = 10$ features among the pre-trained ones, and add an additional feature $w_{\text{fine}}^\star$ drawn at random following a standard Gaussian. More precisely, we let $\mathcal{I}$ be a subset of $\{1, \dots, m_{\text{pre}}\}$ of size m_{fine}—chosen uniformly at random—and the $n = 8000$ fine-tuning samples are drawn i.i.d. as

$$x_k \sim \mathcal{N}(0, \mathrm{I}_d), \qquad y_k = \sigma(x_k^\top w_{\text{fine}}^\star) + \sum_{i \in \mathcal{I}} \sigma(x_k^\top w_i^\star).$$

We then fine-tune the model, from the pre-trained weights obtained by the pre-training procedure described above, for the different fine-tuning algorithms, following the same optimization procedure as described in Appendix A.1.[2]

The fine-tuning procedure considered here is more challenging than the one in Section 5 for three reasons: 1) the initialization is obtained from an actually pre-trained model rather than from the idealized initialization considered in our theoretical analysis; 2) the model is overparameterized; 3) not all pre-trained features are relevant to the fine-tuning task.

The first point is the most challenging in practice. Overparameterization does not pose a significant difficulty, as shown in Appendix A.3, and is in fact useful here, as it allows the model to retain *void* features after pre-training. The third point simply considers a more general setting in which some of the features learned during pre-training are not reused by the fine-tuning task.



(a) Test loss.

(b) Distortion of pre-trained features.

Figure 4: Fine-tuning from real pre-trained weights (no idealized initialization).

Figure 4 illustrates the evolution of the test loss and feature distortion during fine-tuning in this setting. We first note that achieving good performance here requires a larger sample size—and more epochs—than in the idealized setting of Section 5. Thus, the theoretical setting does not directly translate to the practical setting considered here.

Nevertheless, the observations remain consistent with the main insights from our theoretical analysis. Two-timescale fine-tuning and L2-SP successfully recover the new feature while inducing little distortion of the pre-trained features—two-timescale even seems to perform better than L2-SP here. In contrast, both full fine-tuning and LP-FT suffer from feature distortion, although the effect is less pronounced for LP-FT.

Overall, the experiments in this section support the relevance of the idealized initialization scheme in Equation (1) as a model of practical fine-tuning, at least for the pre-training setups considered here.

B Proofs of Section 2

B.1 Proof of Lemma 1

By definition of the two timescale dynamics, $\nabla_a \mathcal{L}_n(a_n(W(t)), W(t)) = 0$, which can be rewritten for a any $i \in [m+1]$ as:

$$-\frac{1}{n}\sum_{k=1}^{n}(f_{a_n(W(t)),W(t)}(x_k) - y_k)\sigma(w_i(t)^\top x_k) = 0.$$

[2]We also use a cosine scheduler for L2-SP here for stability reasons.

Lemma 1 is then proven by computing the time derivative of $\|w_i(t)\|^2$. Indeed, we have a.e.

$$\frac{1}{2}\frac{\mathrm{d}\|w_i(t)\|^2}{\mathrm{d}t} = -a_n(W(t))_i \frac{1}{n}\sum_{k=1}^{n}(f_{a_n(W(t)),W(t)}(x_k) - y_k)\sigma(w_i(t)^\top x_k)$$
$$= 0.$$

So that $\|w_i(t)\|$ is constant over time. □

B.2 Proof of Proposition 1

The proof of Proposition 1 is split in two main parts. The first one ensures that we can apply Arzelà-Ascoli theorem, ensuring the existence of a limit function – up to extraction – W°. The second part then proves that this limit function is necessarily a solution of the limit process of Equation (4).

Lemma 3. *Suppose that Assumption 1 holds for some set $\mathcal{T}$. We then have the following:*

1. *$a_n(W)$ is uniquely defined for any $W \in \mathcal{T}$;*
2. *$W \mapsto a_n(W)$ is Lipschitz on $\mathcal{T}$.*

Proof. By definition, $a_n(W) \in \operatorname{argmin}_{a\in\mathbb{R}^{m+1}} \|a^\top \Sigma(W, \mathbf{X}) - Y\|^2$. This loss is convex and differentiable in a, so that it is equivalent to the following gradient condition:

$$\Sigma(W, \mathbf{X})\Sigma(W, \mathbf{X})^\top a_n(W) = \Sigma(W, \mathbf{X})Y.$$

In particular, if Assumption 1 holds for some set $\mathcal{T}$, $\Sigma(W, \mathbf{X})\Sigma(W, \mathbf{X})^\top$ is invertible, so that this equation admits a unique solution $a_n(W)$, given by

$$a_n(W) = \left(\Sigma(W, \mathbf{X})\Sigma(W, \mathbf{X})^\top\right)^{-1}\Sigma(W, \mathbf{X})Y.$$

Note that, $W \mapsto \Sigma(W, \mathbf{X})$ is Lipschitz. Moreover, it is bounded on the bounded set $\mathcal{T}$, so that $g : W \mapsto \Sigma(W, \mathbf{X})\Sigma(W, \mathbf{X})^\top$ is also Lipschitz on $\mathcal{T}$. By continuity and compactness, the eigenvalues of $g(W)$ are bounded away from 0 on $\mathcal{T}$. Thus, the matrix inversion is Lipschitz on $g(\mathcal{T})$, so that a_n is indeed Lipschitz on $\mathcal{T}$ by composition of Lipschitz functions, and multiplication with bounded, Lipschitz functions. □

Corollary 1. *Let $(\alpha_k)_{k\in\mathbb{N}}$ a positive sequence such that $\alpha_k \overset{k\to\infty}{\longrightarrow} 0$. Suppose that for k large enough, the trajectory $\left(\mathsf{W}^{\alpha_k}(\alpha_k^2 t)\right)_{t\in[0,T]}$ is included within some set $\mathcal{T}$, independent of k, satisfying Assumption 1. Then for k large enough, the sequence $\mathsf{W}^{\alpha_k}(\alpha_k^2\cdot)$ is uniformly Lipschitz.*

Proof. $\mathcal{T}$ is bounded and $a_n(W)$ is bounded on $\mathcal{T}$ thanks to Lemma 3, so that $\partial\mathcal{L}_n(a_n(W), W)$ is bounded on $\mathcal{T}$. Equation (3) then directly yields that the time derivative of $\mathsf{W}^{\alpha_k}(\alpha_k^2\cdot)$ is uniformly bounded almost everywhere on $[0, T]$ for large enough k. In consequence, $\mathsf{W}^{\alpha_k}(\alpha_k^2\cdot)$ is uniformly Lipschitz on the considered time interval. □

Thanks to Corollary 1, we can now apply the Arzelà-Ascoli theorem, for which we give a *simplified* version below, to the sequence $\left(\mathsf{W}^{\alpha_k}(\alpha_k^2 t)\right)_{t\in[0,T]}$.

Theorem 3 (Arzelá-Ascoli, see e.g., Brezis 2011, Theorem 4.25)**.** *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of uniformly bounded and equicontinuous functions from a compact K to $\mathbb{R}^d$. Then $(f_n)_{n\in\mathbb{N}}$ contains a uniformly convergent subsequence.*

In Theorem 3, a sequence of functions on K is said to be uniformly equicontinuous (for the distance d) if for every $\varepsilon > 0$, there exists $\delta > 0$, such that for any $n \in \mathbb{N}$ and $x, y \in K$ with $d(x, y) < \delta$,

$$\|f_n(x) - f_n(y)\| \leq \varepsilon.$$

Obviously, a sequence of uniformly Lipschitz functions is also equicontinuous. Thus, thanks to Corollary 1, the sequence of functions $\mathsf{W}^{\alpha_k}(\alpha_k^2\cdot))_k$ satisfies the conditions of Theorem 3, where the compact K is given by $[0, T]$. We can then extract a subsequence $\alpha_{\varphi(k)}$ such that, as $k \to \infty$,

$(\mathsf{W}^{\alpha_{\varphi(k)}}(\alpha_{j_k}^2 t))_{t\geq 0}$ converges uniformly towards $\mathsf{W}^\circ(t)$ on $[0,T]$. It now remains to show that $\mathsf{W}^\circ(t)$ is a solution of the differential inclusion given by Equation (4).

First note that, by Equation (3), for any $i \in [m]$, $\mathsf{w}_i^\alpha(\alpha^2 t)$ obviously converges to $\frac{w_i^\star}{\|w_i^\star\|}$ for any $t \in \mathbb{R}_+$ as $\alpha \to 0$. It directly implies the first line of Equation (4).

It remains to show the second line of Equation (4), i.e., the dynamics of w_{m+1}°. Denote for shortness in the remaining of this proof $V^k(t) = \mathsf{W}^{\alpha_{\varphi(k)}}(\alpha_{\varphi(k)}^2 t)$. Equation (3) rewrites for the $m+1$-th neuron:

$$\dot{v}_{m+1}^k(t) \in F_{m+1}(V^k(t)),$$

where $F_{m+1}(V) = -\partial_{v_{m+1}}\mathcal{L}(V)$. For k large enough, V^k is contained in $\mathcal{T}$ on $[0,T]$, so that there exists a bounded function f_k such that for any $k \in \mathbb{N}$ and $t \in [0,T]$,

$$v_{m+1}^k(t) = \breve{w} + \int_0^t f_k(s)\mathrm{d}s, \tag{7}$$

where, almost everywhere, $f_k(s) \in F_{m+1}(V^k(s))$. Importantly, f_k is uniformly bounded as soon as it is contained in$\mathcal{T}$. By Banach–Alaoglu theorem and separability of $L^1([0,T];\mathbb{R}^d)$, there exists $f \in L^\infty([0,T];\mathbb{R}^d)$ and a subsequence (f_{ℓ_k}) of (f_k) such that $f_{\ell_k} \rightharpoonup^* f$ in $L^\infty([0,T];\mathbb{R}^d)$ (see e.g., Brezis, 2011, Chapter 4), i.e., for any $\psi \in L^1([0,T];\mathbb{R}^d)$

$$\int_0^T f_{\ell_k}(t)\psi(t)\mathrm{d}t \longrightarrow \int_0^T f(t)\psi(t)\mathrm{d}t.$$

The weak-$\star$ convergence and Equation (7) directly imply that for any $t \in [0,T]$,

$$v_{m+1}^k(t) \underset{k\to\infty}{\longrightarrow} \breve{w} + \int_0^t f(s)\mathrm{d}s.$$

In consequence, $\mathsf{w}_{m+1}^\circ(t) = \breve{w} + \int_0^t f(s)\mathrm{d}s$. As f is bounded, w_{m+1}° is absolutely continuous (Lipschitz), and almost everywhere,

$$\dot{\mathsf{w}}_{m+1}^\circ(t) = f(t). \tag{8}$$

Since $L^2([0,T];\mathbb{R}^d) \subseteq L^1([0,T];\mathbb{R}^d)$, we also have $f_{\ell_k} \rightharpoonup f$ in $L^2([0,T];\mathbb{R}^d)$. By Mazur's lemma (see e.g., Brezis, 2011, Corollary 3.8), there exists a sequence $(g_k)_k$, of finite convex combinations of $(f_{\ell_k})_k$, more precisely,

$$g_k = \sum_{j\geq k} \lambda_j^{(k)} f_{\ell_j},$$

such that g_k converges strongly to f in $L^2([0,T];\mathbb{R}^d)$. Passing to a subsequence (Brezis, 2011, Theorem 4.9), we may therefore assume that $g_k(t)$ converges to $f(t)$ for almost any $t \in [0,T]$.

Moreover for any k, $f_k(t) \in F_{m+1}(V^k(t))$. By upper semi-continuity of the Clarke subdifferential, $V^k(t) \to \mathsf{W}^\circ(t)$ implies that for almost any $t \in [0,T]$,

$$d(f_k(t), F_{m+1}(\mathsf{W}^\circ(t))) \underset{k\to\infty}{\longrightarrow} 0,$$

where $d(x,S) = \inf_{y\in S}\|x-y\|_2$. By triangle inequality we thus also have $d(g_k(t), F_{m+1}(\mathsf{W}^\circ(t))) \underset{k\to\infty}{\longrightarrow} 0$. By closedness of the Clarke subdifferential and convergence of g_k to f, it implies that, almost everywhere, $f(t) \in F_{m+1}(\mathsf{W}^\circ(t))$, which combined with Equation (8) allows to conclude:

$$\dot{\mathsf{w}}_{m+1}^\circ(t) \in -\partial_{\mathsf{w}_{m+1}^\circ}\mathcal{L}(\mathsf{W}^\circ(t)).$$

□

C Proofs of Section 3.1

C.1 Notations and preliminaries

In this section, we denote by $\phi_v : \mathbb{R}^d \to \mathbb{R}_+$ the function defined as $\phi_v(x) = \sigma(v^\top x)$. We also consider the Hilbert space $L^2(\mathcal{N}(0,\mathrm{I}_d))$, defined by the scalar product, for any functions f, g from $\mathbb{R}^d$ to $\mathbb{R}$:

$$\langle f, g\rangle = \mathbf{E}_{x\sim\mathcal{N}(0,\mathrm{I}_d)}[f(x)g(x)].$$

We then define the linear operator $\Phi : \mathbb{R}^m \to L^2(\mathcal{N}(0, \mathrm{I}_d))$,

$$\Phi(a) = \sum_{k=1}^{m} a_k \phi_{w_k^\star}.$$

Its adjoint is then given by $\Phi^* : L^2(\mathcal{N}(0, \mathrm{I}_d)) \to \mathbb{R}^m$ such that its i-th coordinate is given by

$$\Phi^*(f)_i = \langle f, \phi_{w_i^\star} \rangle.$$

We also denote by $G = \Phi^*\Phi$ the $m \times m$ Gram matrix given for any $i, j \in [m]$ by $G_{ij} = \langle \phi_{w_i^\star}, \phi_{w_j^\star} \rangle$. Whenever it is invertible, we denote in the following by $\Pi = \mathrm{I}_{L^2(\mathcal{N}(0,\mathrm{I}_d))} - \Phi G^{-1} \Phi^*$ the orthogonal projection on $\{\phi_{w_1^\star}, \dots, \phi_{w_m^\star}\}^\perp$.

Let also $P_m \in \mathbb{R}^{d \times d}$ be the orthogonal projection on $\mathrm{Span}(w_1^\star, \dots, w_m^\star)$ and for any $w \in \mathbb{S}_{d-1}$, let $P_{w^\perp} \coloneqq \mathrm{I}_d - ww^\top$ be the projection on the orthogonal of w.

The trajectory of w_{m+1}° will be confined within some set $\mathcal{T}(r)$ defined for $r > 0$ as:

$$\mathcal{T}(r) \coloneqq \{w \in \mathbb{S}_{d-1} \mid \|P_m w\|_2 \leq r \text{ and } -\frac{1}{4} \leq w^\top w_{m+1}^\star\}.$$

Define also $\kappa(x) = \frac{\sqrt{1-x^2}+(\pi-\arccos(x))x}{2\pi}$, so that the arc-cosine kernel yields for any $u, v \in \mathbb{S}_{d-1}$, $\langle \phi_u, \phi_v \rangle = \kappa(u^\top v)$.

C.2 Population dynamics

Define in the following

$$\begin{aligned} H(w) &\coloneqq \kappa(w^\top w_{m+1}^\star) - \frac{1}{m+\pi-1} \sum_{i=1}^{m} \kappa(w^\top w_i^\star) \\ D(w) &\coloneqq \frac{1}{2} - \sum_{i,j \in [m]} (G^{-1})_{i,j} \kappa(w^\top w_i^\star) \kappa(w^\top w_j^\star). \end{aligned} \tag{9}$$

Recall Lemma 2 here.

Lemma 2. *Let Assumption 2 hold and $w_{m+1} \in \mathbb{S}_{d-1}$ such that $D(w_{m+1}) > 0$. Then, noting $W = [w_1^{\star\top}, \dots, w_m^{\star\top}, w_{m+1}^\top]$, $a(W)$ is uniquely defined and satisfies $a(W)_{m+1} = \frac{H(w_{m+1})}{D(w_{m+1})}$. Moreover, the loss and its subdifferential satisfy*

$$\mathcal{L}(a(W), W) = D(w_{m+1}^\star) - \frac{H(w_{m+1})^2}{D(w_{m+1})},$$

$$\nabla_{w_{m+1}} \mathcal{L}(a(W), W) = -\frac{H(w_{m+1})}{D(w_{m+1})} \left(\mathrm{I}_d - w_{m+1} w_{m+1}^\top\right) \Psi(w_{m+1}),$$

where $\Psi(w) \coloneqq 2\nabla H(w) - \frac{H(w)}{D(w)} \nabla D(w)$.

Proof. **Computation of $a(W)$.** By definition, $a(W)$ minimizes the optimization problem:

$$\min_{a \in \mathbb{R}^{m+1}} \| \sum_{k=1}^{m+1} a_k \phi_{w_k} - \sum_{k=1}^{m+1} \phi_{w_k^\star} \|^2_{L^2(0, \mathcal{N}(0, \mathrm{I}_d))}, \tag{10}$$

Note that for a given a_{m+1}, it admits a unique minimizer – thanks to the invertibility of the matrix G given below – on its first coordinates, given by

$$a_{1:m} = G^{-1} \Phi^* \left(\sum_{k=1}^{m+1} \phi_{w_k^\star} - a_{m+1} Z(w_{m+1}) \right).$$

G is indeed invertible under Assumption 2, since the arc-cosine kernel then directly gives: $G = \frac{1}{2\pi} \left((\pi - 1)\mathrm{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)$.

The optimization problem (10), then reaches the minimal value for a fixed a_{m+1}:

$$\left\| \Pi \left(a_{m+1} \phi_{w_{m+1}} - \phi_{w^\star_{m+1}} \right) \right\|^2_{L^2(\mathcal{N}(0,\mathrm{I}_d))},$$

which is exactly the squared norm of the projection on the orthogonal of $\{\phi_{w^\star_1} \phi_{w^\star_m}, \dots, \}$. We can then develop this term, so that,

$$\left\| \Pi \left(a_{m+1} \phi_{w_{m+1}} - \phi_{w^\star_{m+1}} \right) \right\|^2_{L^2(\mathcal{N}(0,\mathrm{I}_d))} = a^2_{m+1} \tilde{D}(w_{m+1}) - 2 a_{m+1} \tilde{H}(w_{m+1}) + \tilde{D}(w^\star_{m+1}), \tag{11}$$

where

$$\tilde{D}(w) = \|\Pi \phi_w\|^2_{L^2(\mathcal{N}(0,\mathrm{I}_d))} \quad \text{and} \quad \tilde{H}(w) = \langle \Pi \phi_w, \Pi \phi_{w^\star_{m+1}} \rangle_{L^2(\mathcal{N}(0,\mathrm{I}_d))}. \tag{12}$$

This value is then minimal for a unique value of a_{m+1} whenever $\tilde{D}(w_{m+1}) > 0$, which is given by

$$a_{m+1} = \frac{\tilde{H}(w_{m+1})}{\tilde{D}(w_{m+1})}.$$

Functions $\tilde{D}$ and $\tilde{H}$. To conclude on the value of $a_{m+1}(W)$, let us now show that $\tilde{D}$ and $\tilde{H}$ actually coincide with D and H, defined in Equation (9), under Assumption 2 on the sphere $\mathbb{S}_{d-1}$. Since $\Pi = \mathrm{I}_{L^2(\mathcal{N}(0,\mathrm{I}_d))} - \Phi G^{-1} \Phi^*$, we have for $w \in \mathbb{S}_{d-1}$,

$$\begin{aligned} \tilde{D}(w) &= \|\phi_w\|^2_{L^2(\mathcal{N}(0,\mathrm{I}_d))} - \Phi^*(\phi_w)^\top G^{-1} \Phi^*(\phi_w) \\ &= \frac{1}{2} - \Phi^*(\phi_w)^\top G^{-1} \Phi^*(\phi_w) \\ &= D(w). \end{aligned} \tag{13}$$

where we used the arc-cosine kernel to get that $\|\phi_w\|^2_{L^2(\mathcal{N}(0,\mathrm{I}_d))} = \kappa(1) = \frac{1}{2}$ and $\Phi^\star(\phi_w) = \left(\kappa(w^\top w_i^\star)\right)_{i \in [m]}$.

Note that the arc-cosine kernel also implies, with Assumption 2, that $G = \frac{1}{2\pi}\left((\pi - 1)\mathrm{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)$. For $\tilde{H}$, we can first use the Sherman-Morrison formula to get

$$\begin{aligned} G^{-1} &= 2\pi \left((\pi - 1)\mathrm{I}_m - \mathbf{1}_m \mathbf{1}_m^\top\right)^{-1} \\ &= \frac{2\pi}{\pi - 1} \mathrm{I}_m - \frac{2\pi}{(\pi - 1)^2} \frac{\mathbf{1}_m \mathbf{1}_m^\top}{1 + \frac{2\pi}{\pi - 1} m} \\ &= \frac{2\pi}{\pi - 1} \left(\mathrm{I}_m - \frac{\mathbf{1}_m \mathbf{1}_m^\top}{\pi + m - 1} \right). \end{aligned} \tag{14}$$

Then, thanks to Assumption 2 and the arc-cosine kernel, $\Phi^* \phi_{w^\star_{m+1}} = \frac{1}{2\pi} \mathbf{1}_m$, so that

$$\begin{aligned} G^{-1} \Phi^* \phi_{w^\star_{m+1}} &= \frac{1}{\pi - 1} \left(1 - \frac{m}{\pi + m - 1} \right) \mathbf{1}_m \\ &= \frac{1}{\pi + m - 1} \mathbf{1}_m. \end{aligned}$$

For any $w \in \mathbb{S}_{d-1}$,

$$\begin{aligned} \tilde{H}(w) &= \langle \Pi \phi_w, \Pi \phi_{w^\star_{m+1}} \rangle \\ &= \langle \phi_w, \phi_{w^\star_{m+1}} \rangle - \Phi^*(\phi_w)^\top G^{-1} \Phi^* \phi_{w^\star_{m+1}} \\ &= \langle \phi_w, \phi_{w^\star_{m+1}} \rangle - \frac{1}{\pi + m - 1} \Phi^*(\phi_w)^\top \mathbf{1}_m = H(w), \end{aligned}$$

which allows to conclude on the closed-formula of $a(W)_{m+1}$.

Population Loss. The value of $\mathcal{L}(a(W), W)$ is then given by the value of Equation (11) when $a_{m+1} = \frac{H(w)}{D(w)}$, which yields for any w_{m+1} such that $\widetilde{D}(w_{m+1}) > 0$

$$\mathcal{L}(a(W), W) = \left(\widetilde{D}(w^\star_{m+1}) - \frac{\widetilde{H}(w_{m+1})^2}{\widetilde{D}(w_{m+1})} \right). \tag{15}$$

This directly allows to conclude for the population loss value when $w_{m+1} \in \mathbb{S}_{d-1}$, thanks to the previous paragraph.

Gradient of population loss. First note that $\widetilde{H}$, $\widetilde{D}$ and $\mathcal{L}$ are all differentiable by differentiability under the integral sign. Using the envelope theorem, it then comes whenever $\widetilde{D}(w_{m+1}) > 0$ from Equation (15) that

$$\nabla_{w_{m+1}}\mathcal{L}(a(W), W) = -\frac{\widetilde{H}(w_{m+1})}{\widetilde{D}(w_{m+1})}\left(2\nabla\widetilde{H}(w_{m+1}) - \frac{\widetilde{H}(w_{m+1})}{\widetilde{D}(w_{m+1})}\nabla\widetilde{D}(w_{m+1})\right). \tag{16}$$

Although $\widetilde{D}$ and D coincide on the sphere, their gradient does not (and similarly with $\widetilde{H}, H$). This equivalence is only true for their Riemannian gradient on the sphere, i.e.,

$$P_{w^\perp}\nabla\widetilde{D}(w) = P_{w^\perp}\nabla D(w) \quad \text{for any } w \in \mathbb{S}_{d-1}.$$

A similar equality holds for ∇H and $\nabla\widetilde{H}$. Now, thanks to Lemma 1 (or a similar argument in the population case), we also have $w_{m+1}^\top \nabla_{w_{m+1}}\mathcal{L}(a(W), W) = 0$, or said otherwise, $P_{w^\perp}\nabla_{w_{m+1}}\mathcal{L}(a(W), W) = \nabla_{w_{m+1}}\mathcal{L}(a(W), W)$. So that finally, Equation (16) yields for any $w_{m+1} \in \mathbb{S}_{d-1}$,

$$\begin{aligned}\nabla_{w_{m+1}}\mathcal{L}(a(W), W) &= -\frac{\widetilde{H}(w_{m+1})}{\widetilde{D}(w_{m+1})}\left(2P_{w^\perp}\nabla\widetilde{H}(w_{m+1}) - \frac{\widetilde{H}(w_{m+1})}{\widetilde{D}(w_{m+1})}P_{w^\perp}\nabla\widetilde{D}(w_{m+1})\right)\\ &= -\frac{H(w_{m+1})}{D(w_{m+1})}\left(2P_{w^\perp}\nabla H(w_{m+1}) - \frac{H(w_{m+1})}{D(w_{m+1})}P_{w^\perp}\nabla D(w_{m+1})\right).\end{aligned}$$

□

Corollary 2. *If W° follows the population dynamics given by Equation* (5) *and Assumption 2 holds, then for any $t \geq 0$:*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{w}_{m+1}^{\circ\top} w_{m+1}^\star &= \frac{H(\mathsf{w}_{m+1}^\circ)}{D(\mathsf{w}_{m+1}^\circ)} w_{m+1}^{\star\top} P_{\mathsf{w}_{m+1}^\circ(t)^\perp}\Psi(\mathsf{w}_{m+1}^\circ(t)),\\ \frac{\mathrm{d}}{\mathrm{d}t}\|P_m\mathsf{w}_{m+1}^\circ(t)\| &\leq \frac{H(\mathsf{w}_{m+1}^\circ)}{D(\mathsf{w}_{m+1}^\circ)}\|P_m P_{\mathsf{w}_{m+1}^\circ(t)^\perp}\Psi(\mathsf{w}_{m+1}^\circ(t))\|.\end{aligned}$$

Proof. The first point is a direct application of Lemma 2 to Equation (5). For the second point, $\|P_m\mathsf{w}_{m+1}^\circ(t)\|$ is absolutely continuous. So it is differentiable almost everywhere, and

$$\frac{\mathrm{d}}{\mathrm{d}t}\|P_m\mathsf{w}_{m+1}^\circ(t)\| \leq \left\|\frac{\mathrm{d}}{\mathrm{d}t}P_m\mathsf{w}_{m+1}^\circ(t)\right\|,$$

which then allows to conclude with Lemma 2. □

C.3 Key inequalities

Lemma 4. *Let Assumption 2 hold. There exist positive universal constants $\underline{D}, c, C, m_0$, such that if $m \geq m_0$, then for any $w \in \mathcal{T}(c)$:*

1. $\underline{D} \leq D(w) \leq \frac{1}{2}$;
2. $w_{m+1}^{\star\top} P_{w^\perp}\Psi(w) \geq c(1 - w^\top w_{m+1}^\star)$;
3. $\|P_m P_{w^\perp}\Psi(w)\| \leq C\left(\frac{1}{\sqrt{m}} + \|P_m w\|\right)$;

Moreover, if $\|P_m w\| \leq \frac{1}{4(\pi-1)\sqrt{m}}$ and $w^\top w_{m+1}^\star \geq -\frac{1}{4\pi m}$,

4. $H(w) \geq \frac{c}{m}$.

Proof. **Preliminaries.** By monotonicity of κ' for any $w \in \mathbb{S}_{d-1}$,

$$\begin{aligned}\|\Phi^*(\phi_w) - \frac{1}{2\pi}\mathbf{1}_m\|_2 &= \sqrt{\sum_{k=1}^m \left(\kappa(w^\top w_k^\star) - \kappa(0)\right)^2} \\ &\leq \max_k \kappa'(|w^\top w_k^\star|)\sqrt{\sum_{k=1}^m (w^\top w_k^\star)^2} \\ &\leq \kappa'(\|P_m w\|)\|P_m w\|_2 \end{aligned} \tag{17}$$

where we used the orthogonality property for the last line. Also, κ' is upper bounded by $1/2$ in general, which we might also use as a looser but sufficient bound in the proof.

In the following, we will also rely on the fact that G^{-1} is a positive definite matrix, with its largest eigenvalue equal to $\frac{2\pi}{\pi-1}$ thanks to Equation (14); and $G^{-1}\mathbf{1}_m = \frac{2\pi}{\pi+m-1}\mathbf{1}_m$. In consequence, we also have the bound for any $w \in \mathbb{S}_{d-1}$:

$$\begin{aligned}\|G^{-1}\Phi^*(\phi_w)\|_2 &\leq \frac{1}{2\pi}\|G^{-1}\mathbf{1}_m\|_2 + \|G^{-1}\|_{\text{op}}\,\|\Phi^*(\phi_w) - \frac{1}{2\pi}\mathbf{1}_m\|_2 \\ &\leq \frac{1}{\sqrt{m}} + \frac{\pi}{\pi-1}\|P_m w\|_2, \end{aligned} \tag{18}$$

By orthogonality of the features, another important observation is that

$$\|P_m w\|_2^2 \leq 1 - \left(w^\top w_{m+1}^\star\right)^2 \leq 2\left(1 - w^\top w_{m+1}^\star\right).$$

Let now $w \in \mathcal{T}(r)$ for some constant $r \in (0,1)$ to be fixed later.

1. Bound on D. The upper bound is direct by Equation (13). Define $D_0 = \frac{1}{2} - \frac{1}{(2\pi)^2}\mathbf{1}_m^\top G^{-1}\mathbf{1}_m$. Recall that the Sherman-Morrison formula of Equation (14) implies that

$$\begin{aligned} D_0 &= \frac{1}{2} - \frac{m}{2\pi(\pi+m-1)} \\ &= \frac{(\pi+m)(\pi-1)}{2\pi(\pi+m-1)} \geq \frac{\pi-1}{2\pi}. \end{aligned}$$

From there, observe that

$$\begin{aligned} D_0 - D(w) &= \Phi^*\phi_w^\top G^{-1}\Phi^*\phi_w - \frac{1}{(2\pi)^2}\mathbf{1}_m^\top G^{-1}\mathbf{1}_m \\ &= \left(\Phi^*\phi_w + \frac{1}{2\pi}\mathbf{1}_m\right)^\top G^{-1}\left(\Phi^*\phi_w - \frac{1}{2\pi}\mathbf{1}_m\right) \\ &= \frac{1}{\pi}\mathbf{1}_m^\top G^{-1}\left(\Phi^*\phi_w - \frac{1}{2\pi}\mathbf{1}_m\right) + \left\|\Phi^*\phi_w - \frac{1}{2\pi}\mathbf{1}_m\right\|_{G^{-1}}^2. \end{aligned} \tag{19}$$

so that finally, using the bound on $\left\|G^{-1}\right\|_{\text{op}}$ and the value of $\frac{1}{\pi}\mathbf{1}_m^\top G^{-1}$,

$$\begin{aligned} |D(w) - D_0| &\leq \frac{2}{\pi+m-1}\mathbf{1}_m^\top\left(\Phi^*\phi_w - \frac{1}{2\pi}\mathbf{1}_m\right) + \frac{2\pi}{\pi-1}\left\|\Phi^*\phi_w - \frac{1}{2\pi}\mathbf{1}_m\right\|_2^2 \\ &\leq \frac{2}{\sqrt{m}}\left\|\Phi^*\phi_w - \frac{1}{2\pi}\mathbf{1}_m\right\|_2 + \frac{2\pi}{\pi-1}\left\|\Phi^*\phi_w - \frac{1}{2\pi}\mathbf{1}_m\right\|_2^2 \\ &\leq \frac{1}{\sqrt{m}}\|P_m w\|_2 + \frac{\pi}{\pi-1}\|P_m w\|_2^2 \end{aligned} \tag{20}$$

where we used Equation (19) for the last line. By definition of $\mathcal{T}(r)$, $\|P_m w\|_2 \leq r$. By noting that $D_0 \geq \frac{\pi-1}{2\pi}$, we conclude for the first item when c is small enough.

2. Bound on $w_{m+1}^{\star\top} P_{w^\perp}\Psi(w)$. First recall that $\Psi(w) = 2\nabla H(w) - \frac{H(w)}{D(w)}\nabla D(w)$. The definitions of H in Equation (9), along with the function κ that allows to rewrite $\langle \phi_u, \phi_v\rangle = \kappa(u^\top v)$, yields the

gradient:

$$\nabla H(w) = \kappa'(w^\top w_{m+1}^\star) w_{m+1}^\star - \frac{1}{m+\pi-1}\sum_{k=1}^m \kappa'(w^\top w_k^\star) w_k^\star.$$

So that, using the orthogonality of the features,

$$w_{m+1}^{\star\top} P_{w^\perp} \nabla H(w) = \kappa'(w^\top w_{m+1}^\star)(1-(w^\top w_{m+1}^\star)^2) + \frac{w^\top w_{m+1}^\star}{m+\pi-1}(\sum_{k=1}^m \kappa'(w^\top w_k^\star) w^\top w_k^\star.$$

Similarly, we have by definition of D:

$$\nabla D(w) = -2\sum_{k=1}^m \left(G^{-1}\Phi^*(\phi_w)\right)_k \kappa'(w^\top w_k^\star) w_k^\star, \tag{21}$$

so that,

$$w_{m+1}^{\star\top} P_{w^\perp} \nabla D(w) = 2 w_{m+1}^{\star\top} w \sum_{k=1}^m \left(G^{-1}\Phi^*(\phi_w)\right)_k \kappa'(w^\top w_k^\star) w^\top w_k^\star.$$

We thus have

$$\begin{aligned} w_{m+1}^{\star\top} P_{w^\perp} \Psi(w) &= 2\kappa'(w^\top w_{m+1}^\star)\left(1-\left(w^\top w_{m+1}^\star\right)^2\right) \\ &\quad + 2w^\top w_{m+1}^\star \sum_{k=1}^m \kappa'(w^\top w_k^\star) w^\top w_k^\star \left(\frac{1}{m+\pi-1} - \frac{H(w)}{D(w)}(G^{-1}\Phi^*(\pi_w))_k\right) \\ &\geq 2\kappa'(w^\top w_{m+1}^\star)\left(1-\left(w^\top w_{m+1}^\star\right)^2\right) \\ &\quad - 2\max_k |\kappa'(w^\top w_k^\star)| \|P_m w\| \cdot \left\| \frac{1}{m+\pi-1}\mathbf{1}_m - \frac{H(w)}{D(w)} G^{-1}\Phi^*(\phi_w)\right\|, \end{aligned} \tag{22}$$

where the inequality is just Cauchy-Schwarz. We then decompose as follows

$$\begin{aligned} \left\| \frac{1}{m+\pi-1}\mathbf{1}_m - \frac{H(w)}{D(w)} G^{-1}\Phi^*(\phi_w)\right\|_2 &\leq \|G^{-1}\|_{\mathrm{op}} \left\| \frac{1}{m+\pi-1} G\mathbf{1}_m - \Phi^*(\phi_w)\right\| \\ &\quad + \left|1 - \frac{H(w)}{D(w)}\right| \|G^{-1}\Phi^*(\phi_w)\|. \end{aligned} \tag{23}$$

Recall that $\|G^{-1}\|_{\mathrm{op}} \leq \frac{2\pi}{\pi-1}$, $\|G^{-1}\Phi^*(\phi_w)\| \leq \frac{1}{\sqrt{m}} + \frac{\pi}{\pi-1}\|P_m w\|$, and by Equation (17)

$$\begin{aligned} \left\| \frac{1}{m+\pi-1} G\mathbf{1}_m - \Phi^*(\phi_w)\right\| &= \left\| \frac{1}{2\pi}\mathbf{1}_m - \Phi^*(\phi_w)\right\| \\ &\leq \kappa'(\|P_m w\|)\|P_m w\|. \end{aligned}$$

It thus remains to bound $\left|1 - \frac{H(w)}{D(w)}\right|$, or equivalently the difference $|D(w) - H(w)|$ as D is lower bounded by $\underline{D}$. It then comes, using the bounds above, that

$$\begin{aligned} |D(w) - H(w)| &= \left|\kappa(1) - \kappa(w^\top w_{m+1}^\star) + \Phi^*(\phi_w)^\top G^{-1}\left(\Phi^*(\phi_{w_{m+1}^\star}) - \Phi^*(\phi_w)\right)\right| \\ &\leq \max_{x\in[-1,1]} \kappa'(x)(1-w^\top w_{m+1}^\star) + \|G^{-1}\Phi^*(\phi_w)\| \cdot \|\Phi^*(\phi_{w_{m+1}^\star}) - \Phi^*(\phi_w)\| \\ &\leq \frac{1}{2}(1-w^\top w_{m+1}^\star) + \frac{1}{2}\|P_m w\|\left(\frac{1}{\sqrt{m}} + \frac{\pi}{\pi-1}\|P_m w\|\right). \end{aligned} \tag{24}$$

So wrapping things up back to Equation (22), and using the fact that $\|P_m w\|^2 \leq (1-(w^\top w_{m+1}^\star)^2)$, there is a universal constant C such that, as long as $D(w) \geq \underline{D}$,

$$\begin{aligned} \left\| \frac{1}{m+\pi-1}\mathbf{1}_m - \frac{H(w)}{D(w)} G^{-1}\Phi^*(\phi_w)\right\|_2 &\leq \kappa'(\|P_m w\|)\frac{2\pi}{\pi-1}\|P_m w\| \\ &\quad + C\left(1 - w^\top w_{m+1}^\star + \frac{\|P_m w\|}{m}\right). \end{aligned}$$

And plugging it back in Equation (22), again using that $\|P_m w\|^2 \leq 1 - (w^\top w^\star_{m+1})^2$ and that $1 - w^\top w^\star_{m+1} \leq \frac{4}{3}(1 - (w^\top w^\star_{m+1})^2)$ on the considered set then yields,

$$w^{\star\top}_{m+1} P_{w^\perp} \Psi(w) \geq 2 \left(\kappa'(w^\top w^\star_{m+1}) - \frac{2\pi}{\pi - 1} \kappa'(\|P_m w\|)^2 - \frac{2C}{3\underline{D}} \left(\|P_m w\| + \frac{1}{m} \right) \right) \cdot \left(1 - (w^\top w^\star_{m+1})^2\right)$$

Now note that $\kappa'(-\frac{1}{4}) \geq \frac{1}{4} > \frac{2\pi}{\pi-1}\kappa'(0)^2 = \frac{\pi}{8(\pi-1)}$, so we can choose universal positive constants m_0 large enough and c small enough, such that if $m \geq m_0$ and $w \in \mathcal{T}(c)$,

$$w^{\star\top}_{m+1} P_{w^\perp} \Psi(w) \geq c \left(1 - w^\top w^\star_{m+1}\right). \tag{25}$$

3. Bound on $\|P_m P_{w^\perp} \Psi(w)\|$. Similarly to 2., we have by orthogonality

$$P_m \nabla H(w) = -\frac{1}{m + \pi - 1} \sum_{k=1}^{m} \kappa'(w^\top w^\star_k) w^\star_k,$$

so that its norm is bounded by $\frac{1}{2\sqrt{m}}$. Moreover by orthogonality and thanks to Equation (18),

$$\begin{aligned} \|\nabla D(w)\| &\leq \|G^{-1} \Phi^*(\phi_w)\|_2 \\ &\leq \frac{1}{\sqrt{m}} + \frac{\pi}{\pi - 1} \|P_m w\|_2, \end{aligned} \tag{26}$$

where we used Equation (18). In consequence, $\|P_m \nabla D(w)\|_2 \leq \frac{1}{\sqrt{m}} + \frac{5}{2}\|P_m w\|_2$, so that

$$\begin{aligned} \|P_m P_{w^\perp} \Psi(w)\| &\leq \|P_m \Psi(w)\| + |w^\top \Psi(w)| \, \|P_m w\|_2 \\ &\leq \frac{1}{\sqrt{m}} + \frac{1}{\underline{D}} \left(\frac{1}{\sqrt{m}} + \frac{5}{2} \|P_m w\|_2 \right) + \|\Psi(w)\| \, \|P_m w\|_2. \end{aligned}$$

We then conclude by observing that $\|\Psi(w)\|_2 \leq \sqrt{2} + \frac{1}{\underline{D}} \|\nabla D(w)\|_2$ on $\mathcal{T}(c)$, so that we recover the final bound on $\|P_m P_{w^\perp} \Psi(w)\|$ for some constant C.

4. Bound on H. Lastly, κ is convex, so that for any $x \in [-1, 1]$,

$$\kappa(x) \geq \kappa(0) + \kappa'(0) x = \frac{1}{2\pi} + \frac{x}{4}.$$

Moreover it is below its chords, so for $x \in [0, 1]$,

$$\kappa(x) \leq \kappa(0) + (\kappa(1) - \kappa(0)) x = \frac{1}{2\pi} + \frac{\pi - 1}{2\pi} x.$$

Using these two inequalities, and the monotonicity of κ, it comes by definition of H that

$$\begin{aligned} H(w) &= \kappa(w^\top w^\star_{m+1}) - \frac{1}{m + \pi - 1} \sum_{k=1}^{m} \kappa(w^\top w^\star_k) \\ &\geq \frac{1}{2\pi} (1 - \frac{m}{m + \pi - 1}) + \frac{w^\top w^\star_{m+1}}{4} + \frac{\pi - 1}{2\pi(\pi + m - 1)} \sum_{k=1}^{m} |w^\top w^\star_k| \\ &\geq \frac{\pi - 1}{2\pi(m + \pi - 1)} - \frac{w^\top w^\star_{m+1}}{4m} - \frac{\pi - 1}{2\pi m} \|P_m w\|_1 \\ &\geq \frac{1}{4\pi m} (1 - \pi m \, w^\top w^\star_{m+1} - 2(\pi - 1)\sqrt{m} \|P_m w\|_2), \end{aligned}$$

which allows to conclude for the considered values of $\|P_m w\|$ and $w^\top w^\star_{m+1}$. □

Lemma 5. *Let $\delta \in (0, 1/2)$. If Assumption 2 holds, $\breve{w} \sim \mathrm{Unif}(\mathbb{S}_{d-1})$ and*

$$d \geq C m^2 \ln(1/\delta),$$

where C is a positive universal constant. Then with probability at least $1 - \delta$,

$$\|P_m \breve{w}\| \leq \frac{1}{4(\pi - 1)\sqrt{m}} \quad \textit{and} \quad \breve{w}^\top w^\star_{m+1} \geq -\frac{1}{4\pi \, m}.$$

Proof. **Tail of $\|P_m w\|_2$.** First note that by orthogonality and symmetry, $\mathbf{E}[\|P_m w\|_2^2] = \frac{m}{d}$. As $x \mapsto \|P_m x\|^2$ is 2-Lipschitz on the unit sphere, we can apply Theorem 5.1.4 of Vershynin (2018), so that there exists a universal constant $c > 0$, such that for any $t \geq 0$,

$$\mathbb{P}\left(\|P_m \breve{w}\| \geq \sqrt{t + \frac{m}{d}}\right) \leq \exp\left(-cdt^2\right).$$

Denote for shortness $r = \frac{1}{4(\pi-1)}$. Taking t such that $\sqrt{t + \frac{m}{d}} = \frac{r}{\sqrt{m}} - t \geq 0$ when $r \leq \frac{m}{\sqrt{d}}$. It then comes for $d \geq 2\frac{m^2}{r^2}$,

$$\begin{aligned}\mathbb{P}\left(\|P_m \breve{w}\| \geq \frac{r}{\sqrt{m}}\right) &\leq \exp\left(-cd(r^2/m - m/d)^2\right) \\ &\leq \exp\left(-cd\frac{r^4}{4m^2}\right)\end{aligned}$$

Thus, if $d \geq Cm^2 \ln(1/\delta)$ for a large enough universal constant C, $\mathbb{P}\left(\|P_m \breve{w}\| \geq \frac{r}{\sqrt{m}}\right) \leq \delta/2$.

Concentration of $\breve{w}^\top w^\star_{m+1}$. Similarly, $\mathbf{E}[\breve{w}^\top w^\star_{m+1}] = 0$, so that applying Theorem 5.1.4 of Vershynin (2018) again, for any $t \geq 0$:

$$\mathbb{P}\left(|\breve{w}^\top w^\star_{m+1}| \geq t\right) \leq \exp\left(-cdt^2\right).$$

In particular,

$$\mathbb{P}\left(|\breve{w}^\top w^\star_{m+1}| \geq \frac{1}{4\pi m}\right) \leq \exp\left(-cd\frac{1}{(4\pi)^2 m^2}\right).$$

Thus, if $d \geq Cm^2 \ln(1/\delta)$ for a large enough universal constant C, $\mathbb{P}\left(|\breve{w}^\top w^\star_{m+1}| \geq \frac{\eta}{m}\right) \leq \delta/2$, which concludes by taking a union bound on the two events above. □

C.4 Proof of Theorem 1

Consider the positive universal constants $C, c, \underline{D}$ and m_0 appearing in Lemma 4.

Lemma 5 yields that with probability at least $1 - \delta$,

$$\|P_m w_{m+1}(0)\| \leq \frac{1}{4(\pi-1)\sqrt{m}} \quad \text{and} \quad w_{m+1}(0)^\top w^\star_{m+1} \geq -\frac{1}{4\pi\, m}. \tag{27}$$

with the considered regime for d—for a large enough universal constant C. Let us assume that the random event of Equation (27) holds in the following.

Define now $\tau \coloneqq \inf\{t \geq 0 : w_{m+1}(t) \notin \mathcal{T}(c)\}$. Thanks to Lemma 4, we have for any $t \in [0, \tau]$, a positive constant $\underline{D}$ such that $D(w_{m+1}(t)) \in [\underline{D}, \frac{1}{2}]$. We then also have by the fourth point of Lemma 4 that $H(w_{m+1}(0)) \geq \frac{c}{m}$. Note that thanks to Lemma 2,

$$\mathcal{L}(a(W(t)), W(t)) = D(w^\star_{m+1}) - \frac{H(w_{m+1}(t))^2}{D(w_{m+1}(t))},$$

so that by chain rule, $\frac{H(w_{m+1}(t))^2}{D(w_{m+1}(t))}$ is increasing over time. In particular, $\frac{H(w_{m+1}(t))^2}{D(w_{m+1}(t))} \geq 2\frac{c^2}{m^2}$ on $[0, \tau]$. By continuity, $H(w_{m+1}(t))$ is of constant sign, so that, using the bound on D (and H)

$$\frac{H(w_{m+1}(t))}{D(w_{m+1}(t))} \in \left[\frac{Zc}{m}, \frac{1}{2\underline{D}}\right] \qquad \text{for any } t \in [0, \tau]. \tag{28}$$

We can thus reparametrize time, i.e., define

$$\gamma : s \mapsto \int_0^s \frac{D(w_{m+1}(u))}{H(w_{m+1}(u))} \mathrm{d}u \qquad \text{and} \qquad \widetilde{\mathsf{w}}(s) \coloneqq w_{m+1}(\gamma(s)).$$

The function γ is increasing and $2\underline{D}s \leq \gamma(s) \leq \frac{m}{2c}s$. By reparametrization, it comes for any $s \geq 0$:

$$\begin{aligned}\dot{\widetilde{\mathsf{w}}}(s) &= \frac{D(w_{m+1}(\gamma(s)))}{H(w_{m+1}(\gamma(s)))} \dot{w}_{m+1}(\gamma(s)) \\ &= P_{w_{m+1}(\gamma(s))^\perp} \Psi(w_{m+1}(\gamma(s))).\end{aligned} \tag{29}$$

Let[3] $\overline{s} = \gamma^{-1}(\tau)$. In particular, $\widetilde{\mathsf{w}}(s) \in \mathcal{T}(c)$ for any $s \in [0, \overline{s}]$. Thanks to Equation (29) and Lemma 4, for any $s \in [0, \overline{s}]$:

$$\frac{\mathrm{d}}{\mathrm{d}s} w_{m+1}^{\star\top} \widetilde{\mathsf{w}}(s) \geq c(1 - w_{m+1}^{\star\top} \widetilde{\mathsf{w}}(s)), \tag{30}$$

$$\frac{\mathrm{d}}{\mathrm{d}s} \|P_m \widetilde{\mathsf{w}}(s)\| \leq C_0 \left(\frac{1}{\sqrt{m}} + \|P_m w\| \right), \tag{31}$$

for some constant C_0. In particular, the first point implies that $(1 - w_{m+1}^{\star\top} \widetilde{\mathsf{w}}(s))$ is increasing over time, and by a Grönwall argument, for any $s \in [0, \overline{s}]$

$$1 - w_{m+1}^{\star\top} \widetilde{\mathsf{w}}(s) \leq \frac{5}{4} e^{-cs}. \tag{32}$$

Moreover the second point implies, again by a Grönwall argument that for any $s \in [0, \overline{s}]$:

$$\|P_m \widetilde{\mathsf{w}}(s)\| \leq \left(\|P_m \widetilde{\mathsf{w}}(0)\| + \frac{1}{\sqrt{m}} \right) e^{C_0 s}.$$

Moreover, $\|P_m \widetilde{\mathsf{w}}(s)\| \leq \sqrt{2(1 - w_{m+1}^{\star\top} \widetilde{\mathsf{w}}(s))}$, so that for any $s \in [0, \overline{s}]$:

$$\begin{aligned} \|P_m \widetilde{\mathsf{w}}(s)\| &\leq 2 \min \left(\frac{e^{C_0 s}}{\sqrt{m}}, e^{-cs/2} \right) \\ &= 2 \exp \left(- \frac{c/2}{C_0 + c/2} \ln(\sqrt{m}) \right). \end{aligned}$$

Thus, for a large enough choice of m_0, $\|P_m \widetilde{\mathsf{w}}(s)\| \leq \frac{c}{2}$ on $[0, \overline{s}]$, so that $\overline{s} = \infty$, i.e., $w_{m+1}(t) \in \mathcal{T}(c)$ for any $t \geq 0$.

Moreover, Equation (24) in the proof of Lemma 4 shows that for a small enough positive universal constant η, $H(w) \geq \frac{D}{2}$ for any $w \in \mathbb{S}_{d-1}$ such that $w^\top w_{m+1}^\star \geq 1 - \eta$. In consequence, letting now $s_\eta = \inf\{s \geq 0 \mid w_{m+1}^{\star\top} \widetilde{\mathsf{w}}(s) \geq 1 - \eta\}$, Equation (32) and the fact that $\gamma(s) \leq \frac{m}{2c} s$ imply that $\gamma(s_\eta) \leq C_0 m$ for a large enough universal constant C_0. Moreover, for any $s \geq s_\eta$,

$$\gamma'(s) = \frac{D(w_{m+1}(s))}{H(w_{m+1}(s))} \leq \frac{1}{\underline{D}}.$$

And in particular, for any $t \geq \gamma(s_\eta)$, Equation (30) now implies:

$$\begin{aligned} 1 - \mathsf{w}_{m+1}^\circ(t)^\top w_{m+1}^\star &= 1 - \widetilde{\mathsf{w}}(\gamma^{-1}(t))^\top w_{m+1}^\star \\ &\leq \eta \exp\left(-c(\gamma^{-1}(t) - s_\eta)\right) \\ &\leq \exp\left(-c\underline{D}(t - \gamma(s_\eta))\right), \end{aligned}$$

which concludes the proof. □

D Proof of Section 3.2

In this whole section, we consider Assumption 2 and W° a solution of the differential inclusion (4).

D.1 Additional notations and preliminaries

Besides the quantities introduced in Appendix C.1, we introduce here other useful notions.

For a random variable Y and $\alpha \in \{1, 2\}$, we use the standard notation $\|Y\|_{\psi_\alpha}$ for its α-Orlicz norm, which is defined by

$$\|Y\|_{\psi_\alpha} = \inf\{t > 0 \mid \mathbf{E}[\exp(|Y|^\alpha / t^\alpha)] \leq 2\}.$$

For a random vector X in $\mathbb{R}^d$, we also define its Orlicz norm as $\|X\|_{\psi_\alpha} = \sup_{a \in \mathbb{S}_{d-1}} \|a^\top X\|_{\psi_\alpha}$.

[3] Letting $\overline{s} = \infty$ if $\tau = \infty$.

We define for any $w \in \mathbb{S}_{d-1}$ the following empirical matrices/vectors, given by their component for any $i \in [n], j \in [m]$:

$$(F_n)_{ij} = \sigma(x_i^\top w_j^\star), \qquad Z_n(w)_i = \sigma(x_i^\top w),$$

so that $F_n \in \mathbb{R}^{n \times m}$ and $Z_n(w) \in \mathbb{R}^n$. In the whole proof, we will assume $F_n^\top F_n$ is invertible, which is guaranteed with high probability as soon as $n \gtrsim m$.

Assumption 3. *$F_n^\top F_n$ is invertible.*

Whenever Assumption 3 holds, we define the projection matrix $\Pi_n = \mathrm{I}_m - F_n \left(F_n^\top F_n\right)^{-1} F_n^\top$, which is the orthogonal projection onto $\{Z_n(w_1^\star), \dots, Z_n(w_m^\star)\}^\perp$. We also define the following key quantities of the dynamics for any $w \in \mathbb{S}_{d-1}$:

$$H_n(w) = \frac{1}{n} Z_n(w_{m+1}^\star)^\top \Pi_n Z_n(w), \qquad D_n(w) = \frac{1}{n} Z_n(w)^\top \Pi_n Z_n(w). \tag{33}$$

D.2 Empirical loss: expression and gradient

Lemma 6. *Let Assumption 3 hold and $w_{m+1} \in \mathbb{S}_{d-1}$ such that $D_n(w_{m+1}) > 0$. Then, noting $W = [w_1^{\star\top}, \dots, w_m^{\star\top}, w_{m+1}^\top]$, $a_n(W)$ is uniquely defined and satisfies*

$$\begin{cases} a_n(W)_{m+1} = \frac{H_n(w_{m+1})}{D_n(w_{m+1})} \\ a_n(W)_{1:m} = \mathbf{1}_m + (F_n^\top F_n)^{-1} \left(F_n^\top Z_n(w_{m+1}^\star) - \frac{H_n(w_{m+1})}{D_n(w_{m+1})} F_n^\top Z_n(w_{m+1}) \right). \end{cases}$$

Moreover, the loss and its subdifferential satisfy

$$\mathcal{L}_n(a_n(W), W) = \left(D_n(w_{m+1}^\star) - \frac{H_n(w_{m+1})^2}{D_n(w_{m+1})} \right),$$

$$\partial_{w_{m+1}} \mathcal{L}_n(a_n(W), W) \subseteq -\frac{1}{n} \frac{H_n(w_{m+1})}{D_n(w_{m+1})} J_n(w_{m+1})^\top \Pi_n \left(Z_n(w_{m+1}^\star) - \frac{H_n(w_{m+1})}{D_n(w_{m+1})} Z_n(w_{m+1}) \right),$$

where $J_n(w) = \prod_{i=1}^n J_n^{(i)}(w) \subset \mathbb{R}^{n \times d}$ and $J_n^{(i)}(w) = \partial\sigma(w^\top x_i) x_i \subset \mathbb{R}^d$ for each $i \in [n]$.

Proof. The proof here follows similar lines as the one of Lemma 2, so that we sometimes omit a few computations.

Computation of $a_n(W)$. By definition, $a_n(W)$ minimizes the optimization problem:

$$\min_{a \in \mathbb{R}^{m+1}} \|F_n a_{1:m} + a_{m+1} Z_n(w_{m+1}) - Y\|^2,$$

where $Y = F_n \mathbf{1}_m + Z_n(w_{m+1}^\star)$. For a given a_{m+1}, it admits a unique minimizer on its first coordinates, given by $a_{1:m} = (F_n^\top F_n)^{-1} F_n^\top (Y - a_{m+1} Z(w_{m+1}))$ and reaches the value:

$$\|\Pi_n \left(Z_n(w_{m+1}^\star) - a_{m+1} Z_n(w_{m+1}) \right)\|^2 = n \left(a_{m+1}^2 D_n(w_{m+1}) - 2 a_{m+1} H_n(w_{m+1}) + D_n(w_{m+1}^\star) \right).$$

This value is then minimal for a unique value of a_{m+1} whenever $D_n(w_{m+1}) > 0$, which is given by

$$a_{m+1} = \frac{H_n(w_{m+1})}{D_n(w_{m+1})}.$$

Empirical Loss. It then comes the value of $\mathcal{L}_n$ at this point:

$$\mathcal{L}_n(a_n(W), W) = \left(D_n(w_{m+1}^\star) - \frac{H_n(w_{m+1})^2}{D_n(w_{m+1})} \right).$$

Moreover, using classical properties of the Clarke subdifferential (see e.g., Bolte & Pauwels, 2021), it comes by definition that[4]

$$\begin{aligned} \partial_{w_{m+1}} \mathcal{L}_n(a_n(W), W) &\subseteq \frac{a_n(W)_{m+1}}{n} \sum_{i=1}^n (f_{a_n(W), W}(x_k) - y_k) \partial\sigma(w_{m+1}^\top x_k) x_k \\ &= \frac{a_n(W)_{m+1}}{n} J_n(w_{m+1})^\top \left(F_n a_n(W)_{1:m} + a_n(W)_{m+1} Z_n(w_{m+1}) - Y \right). \end{aligned}$$

[4] Using the envelope theorem is also possible here, but requires a more advanced version that extends to non-differentiable functions. We rely on simpler arguments here.

Moreover, as seen above by optimality of $a_n(W)$,

$$F_n a_n(W)_{1:m} + a_n(W)_{m+1} Z_n(w_{m+1}) - Y = a_n(W)_{m+1} \Pi_n Z_n(w_{m+1}) - \Pi_n Z_n(w^\star_{m+1}),$$

which then allows to conclude with the equality $a_n(W)_{m+1} = \frac{H_n(w_{m+1})}{D_n(w_{m+1})}$. □

In the following, we use for shortness for any $w \in \mathbb{S}_{d-1}$ the notation:

$$\Psi_n(w) \coloneqq \frac{1}{n} J_n(w)^\top \Pi_n \left(Z_n(w^\star_{m+1}) - \frac{H_n(w)}{D_n(w)} Z_n(w) \right),$$

so that $\partial_{w_{m+1}} \mathcal{L}_n(a_n(W), W) \subseteq -\frac{H_n(w_{m+1})}{D_n(w_{m+1})} \Psi_n(w_{m+1})$.

Corollary 3. *If Assumption 3 holds, then for almost any $t \geq 0$,*

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathsf{w}^{\circ\top}_{m+1} w^\star_{m+1} \in \frac{H_n(\mathsf{w}^\circ_{m+1})}{D_n(\mathsf{w}^\circ_{m+1})} w^{\star\top}_{m+1} P_{\mathsf{w}^\circ_{m+1}(t)^\perp} \Psi_n(\mathsf{w}^\circ_{m+1}(t)).$$

Moreover, we can also control the component of w°_{m+1} along the first m features for almost any $t \geq 0$:

$$\frac{\mathrm{d}}{\mathrm{d}t} \|P_m \mathsf{w}^\circ_{m+1}(t)\| \leq \frac{H_n(\mathsf{w}^\circ_{m+1})}{D_n(\mathsf{w}^\circ_{m+1})} \|P_m P_{\mathsf{w}^\circ_{m+1}(t)^\perp} \Psi_n(\mathsf{w}^\circ_{m+1}(t))\|.$$

Proof. The proof is similar to the one of Corollary 2. □

D.3 Concentration of key quantities

In this section, we show that the key quantities appearing in the dynamics concentrate towards their population versions, given in Appendix C. The main result of this section is given by Proposition 2 below.

Proposition 2. *Let $\varepsilon \in (0,1)$ and Assumption 2 holds. There exist universal positive constants $C, c, \underline{D}, \overline{D}, m_0$ such that if $m \geq m_0$ and $n \geq \frac{C}{\varepsilon^2} \ln(1/\varepsilon)^2 (d + \ln(1/\delta))$, then with probability at least $1-\delta$, Assumption 3 holds and for any $w \in \mathcal{T}(c)$:*

1. $\underline{D} \leq D_n(w) \leq \overline{D}$;
2. $\inf_{\psi \in \Psi_n(w)} w^{\star\top}_{m+1} P_{w^\perp} \psi \geq c(1 - w^\top w^\star_{m+1})$;
3. $\sup_{\psi \in \Psi_n(w)} \|P_m P_{w^\perp} \psi\| \leq C \left(\frac{1}{\sqrt{m}} + \|P_m w\| \right) + \varepsilon$.

The hardest item of Proposition 2 to prove is the second one. In that goal, we rely on distinguishing two cases:

- for w bounded away from $w^\star_{m+1}$, we rely on uniform concentration bounds of the different quantities of interest (see Appendix D.3.1). These bounds are also sufficient to yield the first and third items of Proposition 2.
- When w is close to $w^\star_{m+1}$, the term $(1 - w^\top w^\star_{m+1})$ vanishes, so that uniform concentration bounds fail at providing the desired bound. Instead, we rely on different bounding techniques (Appendix D.3.2), at the neighborhood of $w^\star_{m+1}$, using concentration bounds developed by Soltanolkotabi (2017a).

D.3.1 Uniform concentration bounds

To prove Proposition 2, we first rely on general concentrations of empirical processes given in Appendix E.1, which yields the uniform concentration of different intermediate quantities as stated in Lemma 7.

Lemma 7. *Consider Assumption 2. There exists a universal constant $C > 0$ such that for any $\varepsilon, \delta \in (0, 1/2)$, if*

$$n \geq \frac{C}{\varepsilon^2} \ln(1/\varepsilon)^2 (d + \ln(1/\delta)),$$

then with probability at least $1-\delta$, the following statements hold simultaneously:

1. $\|G^{-1/2}\left(\frac{1}{n}F_n^\top F_n\right)G^{-1/2} - \mathrm{I}_m\|_{\mathrm{op}} \leq \varepsilon$;
2. $\sup_{w,v\in\mathbb{S}_{d-1}} \left|\frac{1}{n}Z_n(w)^\top Z_n(v) - \kappa(w^\top v)\right| \leq \varepsilon$;
3. $\sup_{w\in\mathbb{S}_{d-1}} \left\|G^{-1/2}\left(\frac{1}{n}F_n^\top Z_n(w) - \Phi^*\phi_w\right)\right\|_2 \leq \varepsilon$;
4. $\sup_{w,v,u\in\mathbb{S}_{d-1}} \left|B_n(w,v,u) - B(w,v,u)\right| \leq \varepsilon$;
5. $\sup_{w,v\in\mathbb{S}_{d-1}} \left\|G^{-1/2}\left(b_n(w,v) - b(w,v)\right)\right\|_2 \leq \varepsilon$;

where $B_n, B : \mathbb{S}_{d-1}^3 \to \mathbb{R}$ and $b_n, b : \mathbb{S}_{d-1}^2 \to \mathbb{R}^m$ are defined as follows

$$B_n(w,v,u) \coloneqq \frac{1}{n}\sum_{k=1}^n \mathbb{1}_{w^\top x_k>0} v^\top x_k \sigma(u^\top x_k); \quad B(w,v,u) \coloneqq \mathbf{E}_{x\sim\mathcal{N}(0,\mathrm{I}_d)}[\mathbb{1}_{w^\top x>0} v^\top x\sigma(u^\top x)];$$
$$b_n(w,v)_i \coloneqq B_n(w,v,w_i^\star); \quad b(w,v)_i \coloneqq B(w,v,w_i^\star).$$

Proof. In the whole proof, we use the notations $\mathbb{S}_G \coloneqq \{a \in \mathbb{R}^d \mid G^{1/2}a \in \mathbb{S}_{m-1}\}$, $\mathcal{H}_w \coloneqq \{x \in \mathbb{R}^d \mid x^\top w \geq 0\}$ for any $w \in \mathbb{R}^d$, and we define the function:

$$h : \mathbb{R}^d \ni x \longmapsto \begin{pmatrix} \sigma(x^\top w_1^\star) \\ \vdots \\ \sigma(x^\top w_m^\star) \end{pmatrix} \in \mathbb{R}^m. \tag{34}$$

1. For the first item, it comes by definition of the mapping $\Phi : \mathbb{R}^m \to L^2(\mathcal{N}(0,\mathrm{I}_d))$ that

$$\begin{aligned}
\|G^{-1/2}\left(\frac{1}{n}F_n^\top F_n\right)G^{-1/2} - \mathrm{I}_m\|_{\mathrm{op}} &= \sup_{a\in\mathbb{S}_{m-1}} a^\top\left(G^{-1/2}\left(\frac{1}{n}F_n^\top F_n\right)G^{-1/2} - \mathrm{I}_m\right)a \\
&= \sup_{a\in\mathbb{S}_G} a^\top\left(\frac{1}{n}F_n^\top F_n - G\right)a \\
&= \sup_{a\in\mathbb{S}_G} a^\top \frac{1}{n}\sum_{k=1}^n h(x_k)h(x_k)^\top - \mathbf{E}[h(x)h(x)^\top]a \\
&\leq \|G^{-1}\|_{\mathrm{op}} \sup_{a\in\mathbb{S}_{d-1}} a^\top\left(\frac{1}{n}\sum_{k=1}^n h(x_k)h(x_k)^\top - \mathbf{E}[h(x)h(x)^\top]\right)a.
\end{aligned}$$

First, $\|G^{-1}\|_{\mathrm{op}} \leq \frac{2\pi}{\pi-1}$, thanks to Assumption 2. Item 1 is then a direct application of [5] Lemma 15, with $h_1 = h_2 = h$ and $\mathcal{A} = \{\mathbb{R}^d\}$.

The different assumptions are indeed satisfied here, as the VC dimension of $\mathcal{A}$ is 2 here and $h(X)$ is C-sub-Gaussian for some universal constant C, thanks to the independence of its coordinates.

2. Note that for any $w, v \in \mathbb{S}_{d-1}$,

$$\begin{aligned}
\left|\frac{1}{n}Z_n(w)^\top Z_n(v) - \kappa(w^\top v)\right| &= \left|\frac{1}{n}\sum_{k=1}^n \sigma(w^\top x_k)\sigma(v^\top x_k) - \mathbf{E}[\sigma(w^\top X)\sigma(v^\top X)]\right| \\
&= \left|\frac{1}{n}\sum_{k=1}^n (w^\top x_k)(v^\top x_k)\mathbb{1}_{x_k\in\mathcal{H}_w\cap\mathcal{H}_v} - \mathbf{E}[(w^\top X)(v^\top X)\mathbb{1}_{X\in\mathcal{H}_w\cap\mathcal{H}_v}]\right|.
\end{aligned}$$

This point is thus a direct application of Lemma 15 with $\mathcal{A} = \{\mathcal{H}_w \cap \mathcal{H}_v \mid v, w \in \mathbb{S}_{d-1}\}$ and $h_1 = h_2 = \mathrm{I}_d$. Classical VC bounds yield that $\mathcal{A}$ has VC dimension at most $2(d+2)$ (see e.g., van der Vaart & Wellner, 1996, Chapter 2.6, Exercise 14 and Lemma 2.6.17.ii.). The sub-Gaussian assumption is immediate.

[5] We could here use a simpler, covariance estimation bound.

3. Here, for any $w \in \mathbb{S}_{d-1}$, we can write:

$$\begin{aligned}\left\|G^{-1/2}\left(\frac{1}{n}F_n^\top Z_n(w) - \Phi^*\phi_w\right)\right\|_2 &= \sup_{a\in\mathbb{S}_{m-1}} a^\top G^{-1/2}\left(\frac{1}{n}F_n^\top Z_n(w) - \Phi^*\phi_w\right)\\ &= \sup_{a\in\mathbb{S}_G} a^\top\left(\frac{1}{n}F_n^\top Z_n(w) - \Phi^*\phi_w\right)\\ &= \sup_{a\in\mathbb{S}_G}\left|\frac{1}{n}\sum_{k=1}^n a^\top h(x_k)\sigma(w^\top x_k) - \mathbf{E}[a^\top h(X)\sigma(w^\top X)]\right|\\ &\le \|G^{-1/2}\|_{\mathrm{op}}\sup_{a\in\mathbb{S}_{d-1}}\left|\frac{1}{n}\sum_{k=1}^n a^\top h(x_k)\sigma(w^\top x_k) - \mathbf{E}[a^\top h(X)\sigma(w^\top X)]\right|\end{aligned}$$

Again, the third point is a direct application of Lemma 15 with $\mathcal{A} = \{\mathcal{H}_w \mid w \in \mathbb{S}_{d-1}\}$, $h_1 = h$ and $h_2 = \mathrm{I}_d$. Here again, a classical VC bound yields that $\mathcal{A}$ has VC dimension at most $d+2$ (van der Vaart & Wellner, 1996, Chapter 2.6, exercise 14).

4. It is again a direct application of Lemma 15, with $\mathcal{A} = \{\mathcal{H}_w \cap \mathcal{H}_u \mid u, w \in \mathbb{S}_{d-1}\}$ and $h_1 = h_2 = \mathrm{I}_d$.

5. Again note that for any $w, v \in \mathbb{S}_{d-1}$,

$$\begin{aligned}\left\|G^{-1/2}\left(b_n(w,v) - b(w,v)\right)\right\|_2 &= \sup_{a\in\mathbb{S}_G} a^\top\left(b_n(w,v) - b(w,v)\right)\\ &= \sup_{a\in\mathbb{S}_G}\left|\frac{1}{n}\sum_{k=1}^n \mathbb{1}_{w^\top x_k>0}a^\top h(x_k)x_k^\top v - \mathbf{E}[\mathbb{1}_{w^\top X>0}a^\top h(X)X^\top v]\right|\\ &\le \|G^{-1/2}\|_{\mathrm{op}}\sup_{a\in\mathbb{S}_{d-1}}\left|\frac{1}{n}\sum_{k=1}^n \mathbb{1}_{w^\top x_k>0}a^\top h(x_k)x_k^\top v - \mathbf{E}[\mathbb{1}_{w^\top X>0}a^\top h(X)X^\top v]\right|.\end{aligned}$$

It is again a consequence of Lemma 15 with $\mathcal{A} = \{\mathcal{H}_w \mid w \in \mathbb{S}_{d-1}\}$, $h_1 = h$ and $h_2 = \mathrm{I}_d$. □

Lemma 8. *Consider Assumption 2 and the items 1. to 5. of Lemma 7 all hold with $\varepsilon \in (0, 1/2]$. Then $F_n^\top F_n$ is invertible and there is a universal constant C such that for any $w \in \mathbb{S}_{d-1}$,*

1. $|D_n(w) - D(w)| \le C\varepsilon$;
2. $|H_n(w) - H(w)| \le C\varepsilon$;
3. $\sup_{J\in J_n(w)} \|\frac{1}{n}J^\top \Pi_n Z_n(w) - \frac{1}{2}\nabla\widetilde{D}(w)\| \le C\varepsilon$;
4. $\sup_{J\in J_n(w)} \|\frac{1}{n}J^\top \Pi_n Z_n(w^\star_{m+1}) - \nabla\widetilde{H}(w)\| \le C\varepsilon$;

where $\widetilde{D}$ and $\widetilde{H}$ are defined in Equation (12).

Proof. Let us write for this proof $G_n = \frac{1}{n}F_n^\top F_n$. The first item of Lemma 7 implies that

$$\frac{1}{2}G \preceq G_n \preceq \frac{3}{2}G.$$

In particular, G_n is invertible, and $\|G_n^{-1}\|_{\mathrm{op}} \le \frac{4\pi}{\pi-2}$. Moreover, the eigenvalues of $G^{-1/2}G_nG^{-1/2}$ are all within $[1-\varepsilon, 1+\varepsilon]$, so that

$$\|G^{1/2}G_n^{-1}G^{1/2} - \mathrm{I}_m\|_{\mathrm{op}} \le \frac{\varepsilon}{1-\varepsilon} \le 2\varepsilon. \tag{35}$$

Let us now prove the different items of Lemma 8. Again, we use in this proof the notation $\mathbb{S}_G = \{a \in \mathbb{R}^d \mid G^{1/2}a \in \mathbb{S}_{m-1}\}$.

1. By definition of both D_n and D, it comes for any $w \in \mathbb{S}_{d-1}$:

$$D_n(w) - D(w) = \left(\frac{1}{n}Z_n(w)^\top Z_n(w) - \frac{1}{2}\right) + \Phi^*\phi_w^\top G^{-1}\Phi^*\phi_w - \left(\frac{1}{n}F_n^\top Z_n(w)\right)^\top G_n^{-1}\left(\frac{1}{n}F_n^\top Z_n(w)\right).$$

The second item of Lemma 7 directly yields $|\frac{1}{n}Z_n(w)^\top Z_n(w) - \frac{1}{2}| \leq \varepsilon$. It remains to bound the second term. For that, denote here for shortness $u = \Phi^* \phi_w$ and $u_n = \frac{1}{n}F_n^\top Z_n(w)$. Note that the third item of Lemma 7 implies that $\|G^{-1/2}(u - u_n)\| \leq \varepsilon$. Moreover, it comes that

$$\begin{aligned}
\|G^{-1/2}u\|_2 &= \|G^{-1/2}\Phi^*\phi_w\| \\
&= \sup_{s \in \mathbb{S}_{m-1}} (G^{-1/2}s)^\top \Phi^* \phi_w \\
&= \sup_{s \in \mathbb{S}_{m-1}} \langle \Phi(G^{-1/2}s), \phi_w \rangle \\
&\leq \sup_{s \in \mathbb{S}_{m-1}} \|\Phi(G^{-1/2}s)\|_{L^2(\mathcal{N}(0,\mathrm{I}_d))} \|\phi_w\|_{L^2(\mathcal{N}(0,\mathrm{I}_d))}.
\end{aligned}$$

From there, observe that, by definition, $\|\Phi(G^{-1/2}s)\|^2_{L^2(\mathcal{N}(0,\mathrm{I}_d)} = \|s\|_2$ and $\|\phi_w\|_{L^2(\mathcal{N}(0,\mathrm{I}_d))} \leq \frac{1}{\sqrt{2}}$ thanks to the arc-cosine kernel. In consequence,

$$\|G^{-1/2}u\|_2 = \|G^{-1/2}\Phi^*\phi_w\|_2 \leq \frac{1}{\sqrt{2}}. \tag{36}$$

Using the third item of Lemma 7 and Equation (35) then yields

$$\begin{aligned}
|u^\top G^{-1} u - u_n^\top G_n^{-1} u_n| &\leq |u^\top G^{-1}(u - u_n)| + |(u - u_n)^\top G^{-1} u_n| + |u_n (G^{-1} - G_n^{-1}) u_n| \\
&\leq \|G^{-1/2}u\| \cdot \|G^{-1/2}(u - u_n)\| + \|G^{-1/2}u_n\| \cdot \|G^{-1/2}(u - u_n)\| \\
&\quad + \|G^{-1/2}u_n\|^2 \cdot \|\mathrm{I}_m - G^{1/2}G_n^{-1}G^{1/2}\|_{\mathrm{op}} \\
&\leq \frac{1}{\sqrt{2}}\varepsilon + (\varepsilon + \frac{1}{\sqrt{2}})\varepsilon + (\varepsilon + \frac{1}{\sqrt{2}})^2 2\varepsilon \\
&\leq 2(1 + \sqrt{2})\varepsilon.
\end{aligned}$$

So that finally, for any $w \in \mathbb{S}_{d-1}$, $|D_n(w) - D(w)| \leq (3 + 2\sqrt{2})\varepsilon$.

2. The proof of this item follows the exact same lines, since

$$\begin{aligned}
H_n(w) - H(w) = \frac{1}{n} Z_n(w^\star_{m+1})^\top Z_n(w) - \langle \phi_{w^\star_{m+1}}, \phi_w \rangle \\
+ \Phi^* \phi_{w^\star_{m+1}}^\top G^{-1} \Phi^* \phi_w - \left(\frac{1}{n}F_n^\top Z_n(w^\star_{m+1})\right)^\top G_n^{-1} \left(\frac{1}{n}F_n^\top Z_n(w)\right).
\end{aligned}$$

3. Let $w \in \mathbb{S}_{d-1}$ and assume first that the considered loss is differentiable at w, so that $J_n(w)$ reduces to a singleton, i.e., we can write by abuse of notation that $J_n(w) \in \mathbb{R}^{n \times d}$ with

$$J_n(w)_k = \mathbb{1}_{w^\top x_k > 0} x_k \qquad \text{for any } k \in [n].$$

In that case, note that for any $v \in \mathbb{S}_{d-1}$, and b_n defined as in Lemma 7

$$\frac{1}{n} v^\top J_n(w)^\top \Pi_n Z_n(w) = B_n(w, v, w) - b_n(w, v)^\top G_n^{-1} \left(\frac{1}{n}F_n^\top Z_n(w)\right).$$

The definition of $\widetilde{D}$ yields

$$\frac{1}{2} u^\top \nabla \widetilde{D}(w) = B(w, u, w) - b(w, u)^\top G^{-1} \Phi^* \phi_w.$$

In consequence, it comes

$$\begin{aligned}
\|\frac{1}{n} J_n(w)^\top \Pi_n Z_n(w) - \frac{1}{2}\nabla \widetilde{D}(w)\| &= \sup_{v \in \mathbb{S}_{d-1}} v^\top \left(\frac{1}{n} J_n(w)^\top \Pi_n Z_n(w) - \frac{1}{2}\nabla \widetilde{D}(w)\right) \\
&= B_n(w, v, w) - B(w, v, w) \\
&\quad + b(w, v)^\top G^{-1} \Phi^* \phi_w - b_n(w, v)^\top G_n^{-1} \left(\frac{1}{n}F_n^\top Z_n(w)\right).
\end{aligned}$$

Similarly to the first item of Lemma 8, we can bound these different terms. For that, first observe by a similar argument than above that $\|G^{-1/2}b(w,v)\| \leq \frac{1}{\sqrt{2}}$. Indeed,

$$\begin{aligned}\|G^{-1/2}b(w,v)\| &= \sup_{a\in\mathbb{S}_G} \mathbf{E}[\mathbb{1}_{w^\top x>0} v^\top x \Phi(a)(x)] \\ &\leq \sup_{a\in\mathbb{S}_G} \sqrt{\mathbf{E}[\mathbb{1}_{w^\top x>0}(v^\top x)^2]}\|\Phi(a)\|_{L^2(\mathcal{N}(0,\mathrm{I}_d)} \\ &= \frac{1}{\sqrt{2}},\end{aligned}$$

where we used that $\mathbf{E}[\mathbb{1}_{w^\top x>0}(v^\top x)^2] = \frac{1}{2}$ by symmetry, and again that $\|\Phi(a)\|_{L^2(\mathcal{N}(0,\mathrm{I}_d)} = 1$ for $a \in \mathbb{S}_G$.

We can now use the items 1, 3, 4 and 5 of Lemma 7, similarly to the first point, to conclude with a similar bound on $\|\frac{1}{n}J_n(w)^\top \Pi_n Z_n(w) - \frac{1}{2}\nabla\widetilde{D}(w)\|$.

We have thus shown for almost any $w \in \mathbb{S}_{d-1}$ (i.e., at differentiable points), the inequality $\|\frac{1}{n}J_n(w)^\top \Pi_n Z_n(w) - \frac{1}{2}\nabla\widetilde{D}(w)\| \leq C\varepsilon$. By convexity of the function $J \mapsto \|\frac{1}{n}J^\top \Pi_n Z_n(v) - \frac{1}{2}\nabla\widetilde{D}(v)\|$, continuity of $Z_n, \nabla\widetilde{D}$ and as $J_n(w)$ is the convex hull of limits $J_n(w_k)$ for differentiable points $w_k \to w$, the inequality actually holds for any $w \in \mathbb{S}_{d-1}$.

4. The proof of this item follows the exact same lines as the third item, since at any differentiability point w:

$$\begin{aligned}\|\frac{1}{n}J_n(w)^\top \Pi_n Z_n(w^\star_{m+1}) - \nabla\widetilde{H}(w)\| = \sup_{v\in\mathbb{S}_{d-1}} &B_n(w,v,w^\star_{m+1}) - B(w,v,w^\star_{m+1}) \\ &+ b(w,v)^\top G^{-1}\Phi^*\phi_{w^\star_{m+1}} - b_n(w,v)^\top G_n^{-1}\left(\frac{1}{n}F_n^\top Z_n(w^\star_{m+1})\right).\end{aligned}$$

□

D.3.2 Concentration near $w^\star_{m+1}$

Lemma 9. *Let $\delta \in (0,1/2)$ and Assumption 2 holds. There exist positive universal constants c and C such that, if $n \geq C(d+\ln(1/\delta))$, then with probability at least $1-\delta$, the following hold simultaneously:*

1. $\|\mathbf{X}\|_{\mathrm{op}} \leq C\sqrt{n}$;
2. $\frac{1}{n}\|\Pi_n Z_n(w^\star_{m+1})\|^2 \geq c$;
3. $\frac{1}{n}\|P_{n,w^\star_{m+1}} J_n(w^\star_{m+1})q\|^2 \geq c\|q\|^2$ *for any* $q \in \{w^\star_{m+1}\}^\perp$;

where $P_{n,w^\star_{m+1}} \coloneqq \Pi_n - \frac{\Pi_n Z_n(w^\star_{m+1}) Z_n(w^\star_{m+1})^\top \Pi_n}{Z_n(w^\star_{m+1})^\top \Pi_n Z_n(w^\star_{m+1})}$ is the orthogonal projection on the orthogonal of the column space of $[F_n, Z_n(w^\star_{m+1})]$; and $\mathbf{X} = [x_1^\top, \dots, x_n^\top]^\top \in \mathbb{R}^{n\times d}$ is the data matrix.

Note that $J_n(w^\star_{m+1})$ is almost surely a singleton, so that, by abuse of notation, point 3. is stated for the single element in $J_n(w^\star_{m+1})$.

Proof. **1. Bound on $\|X\|_{\mathrm{op}}$.** This is a classical concentration bound for random matrices (see e.g., Vershynin, 2018, Theorem 4.4.5).

3. Bound on $\|P_{n,w^\star_{m+1}} J_n(w^\star_{m+1})q\|$. Let $(u_1, \dots, u_{d-1})$ be an orthonormal basis of $\{w^\star_{m+1}\}^\perp$ and define $U = [u_1^\top, \dots, u_{d-1}^\top] \in \mathbb{R}^{d\times(d-1)}$. Without loss of generality, we can even assume that the first m vectors of the basis are exactly $(w^\star_1, \dots, w^\star_m)$, i.e., $u_k = w^\star_k$ for $k \leq m$. Now define for the

remaining of the proof the function:

$$h : \mathbb{R}^d \ni x \longmapsto \begin{pmatrix} \sigma(x^\top w_1^\star) \\ \vdots \\ \sigma(x^\top w_m^\star) \\ \sigma(x^\top w_{m+1}^\star) \\ \mathbb{1}_{x^\top w_{m+1}^\star > 0} U^\top x \end{pmatrix} \in \mathbb{R}^{m+d}.$$

First note that as $X \sim \mathcal{N}(0, \mathrm{I}_d)$, $h(X)$ is C-sub-Gaussian for some universal constant C. Indeed, for any $a \in \mathbb{R}^{m+d}$,

$$\|a^\top h(X)\|_{\psi_2} \leq \|a_{1:m+1}^\top h(X)_{1:m+1}\|_{\psi_2} + \|a_{m+2:m+d}^\top h(X)_{m+2:m+d}\|_{\psi_2}.$$

Moreover, thanks to Assumption 2, the different coordinates of $h(X)_{1:m+1}$ are independent and identically distributed[6], so that $\|a_{1:m+1}^\top h(X)_{1:m+1}\|_{\psi_2} \leq C\|a_{1:m+1}\|$ for some universal constant. The same reasoning holds for the second term, so that $\|a^\top h(X)\|_{\psi_2} \leq C\|a\|$.

Moreover, the arc-cosine kernel with Assumption 2 provides the following block structure:

$$M \coloneqq \mathbf{E}[h(X)h(X)^\top] = \left(\begin{array}{c|c} \frac{\pi-1}{2\pi}\mathrm{I}_{m+1} + \frac{1}{2\pi}\mathbf{1}_{m+1}\mathbf{1}_{m+1}^\top & \frac{1}{4}\widetilde{I}_m \\ \hline \frac{1}{4}\widetilde{I}_m^\top & \frac{1}{2}\mathrm{I}_{d-1} \end{array}\right),$$

$$\text{where} \qquad \widetilde{I}_{ij} = \begin{cases} 1 \text{ if } i = j \leq m \\ 0 \text{ otherwise.} \end{cases}$$

In particular, it implies that for any $a \in \mathbb{R}_{m+d}$:

$$\begin{aligned} a^\top M a &\geq \frac{\pi-1}{2\pi}\|a_{1:m+1}\|^2 + \frac{1}{2}\|a_{m+2:m+d}\|^2 - \frac{1}{2}\|a_{1:m+1}\|\|a_{m+2:m+d}\| \\ &\geq \left(\frac{\pi-1}{2\pi} - \frac{1}{4}\right)\|a_{1:m+1}\|^2 + \left(\frac{1}{2} - \frac{1}{4}\right)\|a_{m+2:m+d}\|^2 \\ &\geq 0.09\|a\|^2, \end{aligned}$$

i.e. the smallest eigenvalue of M is larger than 0.09—which is a positive universal constant.

From the above sub-Gaussian property and smallest eigenvalue bound on M, it thus implies that there exists a universal constant C, such that for any $a \in \mathbb{R}_{m+d}$, $\|a^\top h(X)\|_{\psi_2}^2 \leq Ca^\top Ma$. A typical covariance concentration bound (see e.g., Vershynin, 2018, Theorem 4.7.1 and Exercise 4.7.3) implies if $n \geq C(m + d + \ln(1/\delta))$ that with probability at least $1 - \delta$,

$$\left\| \frac{1}{n}\sum_{i=1}^n M^{-1/2} h(x_i)h(x_i)^\top M^{-1/2} - \mathrm{I}_{m+d} \right\|_{\mathrm{op}} \leq \frac{1}{2}.$$

Assume for the remaining of the proof that this event holds[7]. In particular, it implies for any $a \in \mathbb{R}_{m+d}$:

$$\frac{1}{n}\sum_{i=1}^n (a^\top h(x_i))^2 \geq \frac{1}{2}a^\top Ma \geq c\|a\|^2, \tag{37}$$

[6] And each coordinate is 1-sub-Gaussian.

[7] By Assumption 2, $m \leq d$.

for some universal constant $c > 0$. Let $q \in \{w^\star_{m+1}\}^\perp$. By definition of U, we can consider $s \in \mathbb{R}^{d-1}$ such that $Us = q$ and $\|s\| = \|q\|$. From there, observe that

$$\begin{aligned}
\frac{1}{n}\|P_{n,w^\star_{m+1}} J_n(w^\star_{m+1})q\|^2 &= \inf_{\alpha\in\mathbb{R}^{m+1}} \frac{1}{n}\left\| J_n(w^\star_{m+1})q - \sum_{k=1}^{m+1} \alpha_k Z_n(w^\star_k)\right\|^2 \\
&= \inf_{\alpha\in\mathbb{R}^{m+1}} \frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}_{x_i^\top w^\star_{m+1}>0} x_i^\top q - \alpha^\top[\sigma(x_i^\top w^\star_1),\dots,\sigma(x_i^\top w^\star_{m+1})]\right)^2 \\
&= \inf_{\alpha\in\mathbb{R}^{m+1}} \frac{1}{n}\sum_{i=1}^n (h(x_i)^\top[-\alpha, s])^2 \\
&\geq c \inf_{\alpha\in\mathbb{R}^{m+1}} \|[-\alpha, s]\|^2 \\
&\geq c\|s\|^2.
\end{aligned}$$

By definition of s, it finally yields the third point of Lemma 9.

2. Bound on $\|\Pi_n Z_n(w^\star_{m+1})\|$. Again, assuming that Equation (37) holds, it comes

$$\begin{aligned}
\frac{1}{n}\|\Pi_n Z_n(w^\star_{m+1})\|^2 &= \inf_{\alpha\in\mathbb{R}^m} \frac{1}{n}\|Z_n(w^\star_{m+1}) - \sum_{k=1}^m \alpha_k Z_n(w^\star_k)\|^2 \\
&= \inf_{\alpha\in\mathbb{R}^m} \frac{1}{n}\sum_{i=1}^n \left(h(x_i)^\top[-\alpha, 1, \mathbf{0}_{d-1}]\right)^2 \\
&\geq c \inf_{\alpha\in\mathbb{R}^m} \|[-\alpha, 1, \mathbf{0}_{d-1}]\|^2 \\
&\geq c.
\end{aligned}$$

□

Lemma 10. *Let $\delta \in (0, 1/2)$ and $\varepsilon \in (0,1)$. There exist positive constants η_ε and C_ε such that, if $n \geq C_\varepsilon(d + \ln(1/\delta))$, then with probability at least $1-\delta$, uniformly for all $w \in \mathbb{S}_{d-1}$ such that $1 - w^\top w^\star_{m+1} \leq \eta_\varepsilon$,*

1. $\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(x_i^\top w)(x_i^\top w^\star_{m+1})\leq 0}\left(x_i^\top P_{w^\perp} w^\star_{m+1}\right)^2 \leq \varepsilon\|P_{w^\perp} w^\star_{m+1}\|^2$;
2. $\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(x_i^\top w)(x_i^\top w^\star_{m+1})\leq 0}\left(x_i^\top w^\star_{m+1}\right)^2 \leq \varepsilon\|P_{w^\perp} w^\star_{m+1}\|^2$..

Proof. Let $\eta \in (0, 1/2)$ and $\varepsilon_1 \in (0, 1/2)$ to be fixed later. Lemma 16 (which is adapted from Soltanolkotabi, 2017a, Lemma 5.5) yields that, if $n \geq \frac{C}{\varepsilon_1}(d + \ln(1/\delta))$ for some universal constant $C \geq 1$, then with probability at least $1-\delta$, uniformly over all $w \in \mathbb{S}_{d-1}$ such that $\|P_{w^{\star\perp}_{m+1}} w\| \leq \sqrt{2\eta}$,

$$\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\frac{1}{2}|x_i^\top w^\star_{m+1}|\leq |x_i^\top P_{w^{\star\perp}_{m+1}} w|}\left(x_i^\top P_{w^{\star\perp}_{m+1}} w\right)^2 \leq C(\varepsilon_1 + \sqrt{\eta})\|P_{w^{\star\perp}_{m+1}} w\|^2. \tag{38}$$

Assume in the following of the proof that Equation (38) holds for any $w \in \mathbb{S}_{d-1}$ such that $\|P_{w^{\star\perp}_{m+1}} w\| \leq \sqrt{2\eta}$.

Let now $w \in \mathbb{S}_{d-1}$ with $w^\top w^\star_{m+1} \geq 1-\eta$. First notice that $\|P_{w^{\star\perp}_{m+1}} w\| \leq \sqrt{2\eta}$. Moreover, decomposing $w^\star_{m+1} = (w^\top w^\star_{m+1}) w^\star_{m+1} + P_{w^{\star\perp}_{m+1}} w$, observe that for any $x \in \mathbb{R}^d$

$$\begin{aligned}
\mathbb{1}_{(x^\top w)(x^\top w^\star_{m+1})\leq 0} &\leq \mathbb{1}_{(x^\top w^\star_{m+1})^2 w^\top w^\star_{m+1} \leq -(x^\top w^\star_{m+1})(x^\top P_{w^{\star\perp}_{m+1}} w)} \\
&\leq \mathbb{1}_{(w^\top w^\star_{m+1})|x^\top w^\star_{m+1}|\leq |x^\top P_{w^{\star\perp}_{m+1}} w|} \\
&\leq \mathbb{1}_{\frac{1}{2}|x^\top w^\star_{m+1}|\leq |x^\top P_{w^{\star\perp}_{m+1}} w|}.
\end{aligned} \tag{39}$$

In particular, it yields for any $i \in [n]$

$$\mathbb{1}_{(x_i^\top w)(x_i^\top w^\star_{m+1})\le 0}\left(x_i^\top w^\star_{m+1}\right)^2 \le \mathbb{1}_{\frac{1}{2}|x_i^\top w^\star_{m+1}|\le |x_i^\top P_{w^{\star\perp}_{m+1}} w|}\left(x_i^\top w^\star_{m+1}\right)^2$$
$$\le 4\,\mathbb{1}_{\frac{1}{2}|x_i^\top w^\star_{m+1}|\le |x_i^\top P_{w^{\star\perp}_{m+1}} w|}\left(x_i^\top P_{w^{\star\perp}_{m+1}} w\right)^2,$$

which yields, thanks to Equation (38), the second inequality in Lemma 10 when choosing $\eta \le \left(\frac{\varepsilon}{8C}\right)^2$ and $\varepsilon_1 \le \frac{\varepsilon}{8C}$.

Moreover, note that

$$\begin{aligned} P_{w^\perp} w^\star_{m+1} &= w^\star_{m+1} - (w^\top w^\star_{m+1}) w \\ &= w^\star_{m+1} - (w^\top w^\star_{m+1})\left(P_{w^{\star\perp}_{m+1}} w + (w^\top w^\star_{m+1}) w^\star_{m+1}\right) \\ &= (1-(w^\top w^\star_{m+1})^2) w^\star_{m+1} - (w^\top w^\star_{m+1}) P_{w^{\star\perp}_{m+1}} w. \end{aligned}$$

So that whenever $\frac{1}{2}|x^\top w^\star_{m+1}| \le |x^\top P_{w^{\star\perp}_{m+1}} w|$, we have $|x^\top P_{w^\perp} w^\star_{m+1}| \le 3|x^\top P_{w^{\star\perp}_{m+1}} w|$. Along with Equation (39), this yields for any $i \in [n]$

$$\left(x_i^\top P_{w^\perp} w^\star_{m+1}\right)^2 \mathbb{1}_{(x_i^\top w)(x_i^\top w^\star_{m+1})\le 0} \le 9\left(x_i^\top P_{w^{\star\perp}_{m+1}} w\right)^2 \mathbb{1}_{\frac{1}{2}|x_i^\top w^\star_{m+1}|\le |x_i^\top P_{w^{\star\perp}_{m+1}} w|}.$$

The first inequality of Lemma 10 is then also a consequence of Equation (38), when taking $\eta = \left(\frac{\varepsilon}{18C}\right)^2$ and $\varepsilon_1 = \frac{\varepsilon}{18C}$. □

Corollary 4. *Assume that the concentration events of both Lemmas 9 and 10 hold with parameters $\varepsilon, \eta_\varepsilon, c$ and C. We then have for any $w \in \mathbb{S}_{d-1}$ such that $1 - w^\top w^\star_{m+1} \le \min\left(\eta_\varepsilon, \frac{1}{2}\right)$*

1. $\frac{1}{n}\left\|P_{n,w^\star_{m+1}} J_n(w^\star_{m+1}) P_{w^\perp} w^\star_{m+1}\right\|^2 \ge \frac{3c}{8}\left(1 - w^\top w^\star_{m+1}\right)$;
2. $\|P_{n,w} - P_{n,w^\star_{m+1}}\|_{\mathrm{op}} \le 4\sqrt{2}\,\frac{\sqrt{2c}+C}{c}\,C\sqrt{\left(1 - w^\top w^\star_{m+1}\right)}$;

and for any $J \in J_n(w)$:

3. $\frac{1}{n}\|J P_{w^\perp} w^\star_{m+1}\|^2 \le 2C^2\left(1 - w^\top w^\star_{m+1}\right)$;
4. $\frac{1}{n}\left\|\left(J - J_n(w^\star_{m+1})\right) P_{w^\perp} w^\star_{m+1}\right\|^2 \le 2\varepsilon\left(1 - w^\top w^\star_{m+1}\right)$;
5. $\frac{1}{n}\left\|Z_n(w^\star_{m+1}) - w^\top w^\star_{m+1} Z_n(w) - J P_{w^\perp} w^\star_{m+1}\right\|^2 \le 2\varepsilon\left(1 - w^\top w^\star_{m+1}\right)$.

Proof. In the whole proof, let $w \in \mathbb{S}_{d-1}$ be such that $1 - w^\top w^\star_{m+1} \le \min\left(\eta_\varepsilon, \frac{1}{2}\right)$ and denote $u = P_{w^\perp} w^\star_{m+1}$ for shortness. We first recall the useful identity:

$$u = P_{w^\perp} w^\star_{m+1} = \left(1 - (w^\top w^\star_{m+1})^2\right) w^\star_{m+1} - (w^\top w^\star_{m+1}) P_{w^{\star\perp}_{m+1}} w. \tag{40}$$

In the remaining, we will also be using the fact that $\|u\|^2 = 1 - (w^\top w^\star_{m+1})^2 \le 2(1 - w^\top w^\star_{m+1})$.

1. By definition, note that $J_n(w^\star_{m+1}) w^\star_{m+1} = Z_n(w^\star_{m+1})$, so that $P_{n,w^\star_{m+1}} J_n(w^\star_{m+1}) w^\star_{m+1} = \mathbf{0}$. In consequence, using Equation (40)

$$\begin{aligned} \frac{1}{n}\left\|P_{n,w^\star_{m+1}} J_n(w^\star_{m+1}) u\right\|^2 &= (w^\top w^\star_{m+1})^2 \frac{1}{n}\left\|P_{n,w^\star_{m+1}} J_n(w^\star_{m+1}) P_{w^{\star\perp}_{m+1}} w\right\|^2 \\ &\ge c(w^\top w^\star_{m+1})^2 \|P_{w^{\star\perp}_{m+1}} w\|^2 \\ &\ge c(w^\top w^\star_{m+1})^2\left(1 - (w^\top w^\star_{m+1})^2\right) \\ &\ge \frac{3c}{8}\left(1 - w^\top w^\star_{m+1}\right), \end{aligned}$$

where the second line is due to item 3. of Lemma 9.

2. For shortness of notation, denote here the vectors $a = \frac{1}{\sqrt{n}}\Pi_n Z_n(w^\star_{m+1})$ and $b = \frac{1}{\sqrt{n}}\Pi_n Z_n(w)$. By definition, $P_{n,w} - P_{n,w^\star_{m+1}} = \frac{aa^\top}{\|a\|^2} - \frac{bb^\top}{\|b\|^2}$. In consequence,

$$\begin{aligned}
\|P_{n,w} - P_{n,w^\star_{m+1}}\|_{\mathrm{op}} &= \left\| \frac{aa^\top}{\|a\|^2} - \frac{bb^\top}{\|b\|^2} \right\|_{\mathrm{op}} \\
&\leq \frac{1}{\|a\|^2}\left(\|aa^\top - bb^\top\|_{\mathrm{op}}\right) + \|bb^\top\|_{\mathrm{op}}\left(\frac{1}{\|b\|^2} - \frac{1}{\|a\|^2}\right) \\
&\leq \frac{1}{\|a\|^2}\left(\|ba^\top - bb^\top\|_{\mathrm{op}} + \|aa^\top - ba^\top\|_{\mathrm{op}}\right) + (1 - \frac{\|b\|^2}{\|a\|^2}) \\
&\leq 2\frac{\|a\| + \|b\|}{\|a\|^2}\|a - b\| \leq \frac{4\|a\| + 2\|a - b\|}{\|a\|^2}\|a - b\|. \qquad (41)
\end{aligned}$$

Note that, thanks to item 2. of Lemma 9, $\|a\|^2 \geq c$. Moreover,

$$\begin{aligned}
\|a - b\|^2 &= \frac{1}{n}\|Z_n(w) - Z_n(w^\star_{m+1})\|^2 \\
&\leq \frac{1}{n}\|\mathbf{X}\|^2_{\mathrm{op}}\,\|w - w^\star_{m+1}\|^2 \\
&\leq 2C^2(1 - w^\top w^\star_{m+1}).
\end{aligned}$$

Plugging this into Equation (41) concludes item 3.

3. It is a simple consequence of the fact that $\|J\|_{\mathrm{op}} \leq \|\mathbf{X}\|_{\mathrm{op}}$ for any $J \in J_n(w)$, along with item 1. of Lemma 9.

4. It is a direct application of the first item of Lemma 10.

5. It is a consequence of the following inequality for any $J \in J_n(w)$ and $i \in [n]$:

$$0 \leq \left(Z_n(w^\star_{m+1}) - w^\top w^\star_{m+1} Z_n(w) - Jq\right)_i \leq |x_i^\top w^\star_{m+1}|\mathbb{1}_{(x_i^\top w)(x_i^\top w^\star_{m+1})\leq 0},$$

where the right inequality is an equality whenever $x_i^\top w \neq 0$. Item 5. then directly follows from the second item of Lemma 10. □

We rely on the following decomposition.

Lemma 11. *For any $w \in \mathbb{R}^d$ such that $D_n(w) > 0$, and noting $u = P_{w^\perp} w^\star_{m+1}$,*

$$u^\top \Psi_n(w) = \left\{ \frac{1}{n}\left\|P_{n,w}Ju\right\|^2 + \frac{1}{n}\left(Ju\right)^\top P_{n,w}\left(Z_n(w^\star_{m+1}) - w^\top w^\star_{m+1} Z_n(w) - Ju\right) \mid J \in J_n(w) \right\},$$

where $P_{n,w} \coloneqq \Pi_n - \frac{\Pi_n Z_n(w) Z_n(w)^\top \Pi_n}{Z_n(w)^\top \Pi_n Z_n(w)}$ is the orthogonal projection on the orthogonal of the column space of $[F_n, Z_n(w)]$.

Proof. Recall that $\Psi_n = \left\{ \frac{1}{n} J^\top \Pi_n \left(Z_n(w^\star_{m+1}) - \frac{H_n(w)}{D_n(w)} Z_n(w) \right) \mid J \in J_n(w) \right\}$.

It then suffices to show that for any $J \in J_n(w)$

$$\begin{aligned}
\Pi_n\left(Z_n(w^\star_{m+1}) - \frac{H_n(w)}{D_n(w)}Z_n(w)\right) &= P_{n,w}Ju + P_{n,w}\left(Z_n(w^\star_{m+1}) - w^\top w^\star_{m+1} Z_n(w) - Ju\right) \\
&= P_{n,w}\left(Z_n(w^\star_{m+1}) - w^\top w^\star_{m+1} Z_n(w)\right). \qquad (42)
\end{aligned}$$

Let us prove Equation (42) above—which does not depend on J anymore. For that, recall the observation from the proof of Lemma 6 (computation of $a_n(W)$), that

$$\frac{H_n(w_{m+1})}{D_n(w_{m+1})} = \arg\min_{a \in \mathbb{R}} \left\|\Pi_n\left(Z_n(w^\star_{m+1}) - a Z_n(w_{m+1})\right)\right\|^2.$$

Equivalently, $\Pi_n \left(Z_n(w^\star_{m+1}) - \frac{H_n(w)}{D_n(w)} Z_n(w) \right)$ is the orthogonal projection of $\Pi_n Z_n(w^\star_{m+1})$ onto $\Pi_n Z_n(w)^\perp$, i.e.,

$$\begin{aligned}\Pi_n \left(Z_n(w^\star_{m+1}) - \frac{H_n(w)}{D_n(w)} Z_n(w) \right) &= \left(\mathrm{I}_n - \frac{\Pi_n Z_n(w) Z_n(w)^\top \Pi_n}{Z_n(w)^\top \Pi_n Z_n(w)} \right) \Pi_n Z_n(w^\star_{m+1}) \\ &= P_{n,w} Z_n(w^\star_{m+1}).\end{aligned}$$

Since $P_{n,w} Z_n(w) = \mathbf{0}$ by definition of $P_{n,w}$, it directly implies Equation (42) and thus Lemma 11. □

Although this decomposition might not seem helpful at first hand, it is actually very useful thanks to the concentration bounds of Corollary 4, which allow to control both terms appearing in the decomposition of Lemma 11 when close to $w^\star_{m+1}$. The first term then dominates the second term, ensuring local convergence of $\mathsf{w}^\circ_{m+1}(t)$ towards $w^\star_{m+1}$.

Lemma 12. *Let $\delta \in (0, 1/2)$ and Assumption 2 holds. There exist positive universal constants η_0, c and C such that if $n \geq C\left(d + \ln(1/\delta)\right)$, then with probability at least $1-\delta$, for any $w \in \mathbb{S}_{d-1}$ with $1 - w^\top w^\star_{m+1} \leq \eta_0$:*

$$\inf_{\psi \in \Psi_n(w)} w^{\star\top}_{m+1} P_{w^\perp} \psi \geq c \left(1 - w^\top w^\star_{m+1}\right). \tag{43}$$

Proof. Let $\varepsilon \in (0,1)$ to be fixed later. If $n \geq C_\varepsilon(d + m + \ln(1/\delta))$, then with probability at least $1-\delta$, the five concentration bounds of Corollary 4 all hold uniformly on any $w \in \mathbb{S}_{d-1}$ such that $1 - w^\top w^\star_{m+1} \leq \eta_\varepsilon$. From now on, assume these five items hold, and let $w \in \mathbb{S}_{d-1}$ such that $1 - w^\top w^\star_{m+1} \leq \min(\eta_\varepsilon, \frac{1}{2})$.

Thanks to Lemma 11, noting $u = P_{w^\perp} w^\star_{m+1}$ for shortness,

$$\inf_{\psi \in \Psi_n(w)} u^\top \psi \geq \inf_{J \in J_n(w)} \frac{1}{n} \|P_{n,w} J u\|^2 - \sup_{J \in J_n(w)} \frac{1}{n} \|Ju\| \|Z_n(w^\star_{m+1}) - w^\top w^\star_{m+1} Z_n(w) - Ju\|. \tag{44}$$

Items 3. and 5. of Corollary 4 directly imply that

$$\sup_{J \in J_n(w)} \frac{1}{n} \|Ju\| \|Z_n(w^\star_{m+1}) - w^\top w^\star_{m+1} Z_n(w) - Ju\| \leq 2C\sqrt{\varepsilon}(1 - w^\top w^\star_{m+1}). \tag{45}$$

Moreover for the first term, note that for any $J \in J_n(w)$:

$$\begin{aligned}\frac{1}{n}\|P_{n,w} J u\|^2 &= \frac{1}{n}\|P_{n,w^\star_{m+1}} J_n(w^\star_{m+1}) u\|^2 + \frac{1}{n}\left(\|P_{n,w} J u\|^2 - \|P_{n,w^\star_{m+1}} J_n(w^\star_{m+1}) u\|^2 \right) \\ &\geq \frac{3c}{8}(1 - w^\top w^\star_{m+1}) - \frac{1}{n}\left(\|P_{n,w} J u\| + \|P_{n,w^\star_{m+1}} J_n(w^\star_{m+1}) u\| \right) \left\| P_{n,w^\star_{m+1}} J_n(w^\star_{m+1}) u - P_{n,w} J u \right\|,\end{aligned}$$

thanks to item 1. of Corollary 4. From there, first note that

$$\begin{aligned}\frac{1}{\sqrt{n}}\left(\|P_{n,w} J u\| + \|P_{n,w^\star_{m+1}} J_n(w^\star_{m+1}) u\| \right) &\leq \frac{2}{\sqrt{n}}\|Ju\| + \frac{1}{\sqrt{n}}\|(J - J_n(w^\star_{m+1}))u\| \\ &\leq \left(2\sqrt{2}C + \sqrt{2\varepsilon}\right)\sqrt{1 - w^\top w^\star_{m+1}},\end{aligned}$$

thanks to items 3. and 4. of Corollary 4. Finally, we have by triangle inequality

$$\begin{aligned}\frac{1}{\sqrt{n}}\left\| P_{n,w^\star_{m+1}} J_n(w^\star_{m+1}) u - P_{n,w} J u \right\| &\leq \frac{1}{\sqrt{n}}\left\| P_{n,w^\star_{m+1}} J_n(w^\star_{m+1}) u - P_{n,w^\star_{m+1}} J u \right\| + \frac{1}{\sqrt{n}}\left\| (P_{n,w^\star_{m+1}} - P_{n,w}) J u \right\| \\ &\leq \frac{1}{\sqrt{n}}\|(J - J_n(w^\star_{m+1}))u\| + \frac{1}{\sqrt{n}}\|Ju\| \|P_{n,w^\star_{m+1}} - P_{n,w}\|_{\mathrm{op}} \\ &\leq \sqrt{2\varepsilon}\sqrt{1 - w^\top w^\star_{m+1}} + 8\frac{2c + C}{c} C^{3/2}(1 - w^\top w^\star_{m+1}),\end{aligned}$$

thanks to items 2., 3. and 4. of Corollary 4. Plugging everything back together, it finally yields that:

$$\inf_{J\in J_n(w)} \frac{1}{n}\|P_{n,w}Ju\|^2 \geq \left(\frac{3c}{8} - \left(2\sqrt{2}C + \sqrt{2\varepsilon}\right)\cdot\left(\sqrt{2\varepsilon} + 8\frac{2c+C}{c}C^{3/2}\sqrt{1-w^\top w^\star_{m+1}}\right)\right) \cdot (1-w^\top w^\star_{m+1}).$$

Thus, for small enough universal constants ε and η_0, it comes that for any $w \in \mathbb{S}_{d-1}$ such that $1-w^\top w^\star_{m+1} \leq \eta_0$,

$$\inf_{J\in J_n(w)} \frac{1}{n}\|P_{n,w}Ju\|^2 \geq \frac{c}{4}(1-w^\top w^\star_{m+1}),$$

and simultaneously, thanks to Equation (45):

$$\sup_{J\in J_n(w)} \frac{1}{n}\|Ju\|\|Z_n(w^\star_{m+1}) - w^\top w^\star_{m+1} Z_n(w) - Ju\| \leq \frac{c}{8}(1-w^\top w^\star_{m+1}).$$

It then allows to conclude, thanks to the decomposition of Equation (44). □

D.3.3 Proof of Proposition 2

Let us first restate Proposition 2.

Proposition 2. *Let $\varepsilon \in (0,1)$ and Assumption 2 holds. There exist universal positive constants $C, c, \underline{D}, \overline{D}, m_0$ such that if $m \geq m_0$ and $n \geq \frac{C}{\varepsilon^2}\ln(1/\varepsilon)^2\,(d+\ln(1/\delta))$, then with probability at least $1-\delta$, Assumption 3 holds and for any $w \in \mathcal{T}(c)$:*

1. $\underline{D} \leq D_n(w) \leq \overline{D}$;
2. $\inf_{\psi\in\Psi_n(w)} w^{\star\top}_{m+1} P_{w^\perp}\psi \geq c(1-w^\top w^\star_{m+1})$;
3. $\sup_{\psi\in\Psi_n(w)} \|P_m P_{w^\perp}\psi\| \leq C\left(\frac{1}{\sqrt{m}} + \|P_m w\|\right) + \varepsilon$.

Proof. Let $\eta \in (0,1)$ to be fixed later. First recall the identity

$$\Psi_n(w) = \frac{1}{n}J_n(w)^\top \Pi_n\left(Z_n(w^\star) - \frac{H_n(w)}{D_n(w)}Z_n(w)\right).$$

If $n \geq \frac{C}{\varepsilon^2}\ln(1/\varepsilon)^2\,(d+\ln(1/\delta))$ for a large enough constant C, Lemma 8 directly implies that for any $w \in \mathbb{S}_{d-1}$, with probability at least $1-\delta$, uniformly on $\mathbb{S}_{d-1}$:

1. $|D_n(w) - D(w)| \leq C_0\eta$;
2. $\inf_{\psi\in\Psi_n(w)} \|\psi - \nabla\tilde{H}(w) - \frac{H(w)}{2D(w)}\nabla\tilde{D}(w)\| \leq C_0\min(\eta,\varepsilon)$,

for some universal constant C_0. Assume in the following that these different concentrations hold.

Since $P_{w^\perp}\nabla\widetilde{D}(w) = P_{w^\perp}\nabla D(w)$—and similarly for $\widetilde{H}$ and H—it then implies, thanks to Lemmas 4 that for a small enough choice of constants η and c that for any $w \in \mathcal{T}(c)$,

1. $\underline{D} \leq D_n(w) \leq \overline{D}$;
2. $\inf_{\psi\in\Psi_n(w)} w^{\star\top}_{m+1} P_{w^\perp}\psi \geq c(1-w^\top w^\star_{m+1}) - C\eta$;
3. $\sup_{\psi\in\Psi_n(w)} \|P_m P_{w^\perp}\psi\| \leq C\left(\frac{1}{\sqrt{m}} + \|P_m w\|\right) + \varepsilon$.

It already implies items 1. and 3. of Proposition 2–assuming η will be taken of constant order. For item 2., it also proves it when sufficiently far from $w^\star_{m+1}$. Indeed, we have for any $w \in \mathcal{T}(c)$ such that $(1-w^\top w^\star_{m+1}) \geq \frac{C\eta}{2c}$,

$$\inf_{\psi\in\Psi_n(w)} w^{\star\top}_{m+1} P_{w^\perp}\psi \geq \frac{c}{2}(1-w^\top w^\star_{m+1}). \tag{46}$$

For the neighborhood of $w^\star_{m+1}$, Lemma 12 directly implies that for a small enough universal constant η, then with probability at least $1-\delta$, for any $w \in \mathcal{T}(c)$ such that $(1 - w^\top w^\star_{m+1}) \leq \frac{C\eta}{2c}$:

$$\inf_{\psi \in \Psi_n(w)} w^{\star\top}_{m+1} P_{w^\perp} \psi \geq c\left(1 - w^\top w^\star_{m+1}\right). \tag{47}$$

This concludes the proof. □

D.3.4 Concentration at initialization

The population loss also relied on a lower bound of H of order $\frac{1}{m}$ on the trajectory. We also need such a lower bound in the empirical case. However, we cannot rely on a uniform concentration over $\mathbb{S}_{d-1}$, as we would get an error term $\Theta(1)$, which is larger than $1/m$ and thus yields no guarantee on the sign of H_n on the points of interest. Instead, we rely on a single point concentration at the initialization $\breve{w}$, which is chosen independently of the training data.

An important detail that can be noted from the proof of Lemma 7 is that for its first item, we only need a sample complexity $n \geq \frac{C}{\varepsilon^2}(m + \ln(1/\delta))$, since we only relied on the use of Lemma 15 with parameters $p_1 = p_2 = m$ and $\mathcal{A} = \{\mathbb{R}^d\}$ for that point. Using the same proof technique, we can thus also prove the following pointwise concentration, which yields tighter bounds at the initialization point of the dynamics.

Lemma 13. *Consider Assumption 2. There exists a universal constant $C > 0$ such that for any $\varepsilon, \delta \in (0, 1/2)$ and $w, v \in \mathbb{R}^d$, if $n \geq \frac{C}{\varepsilon^2}(m + \ln(1/\delta))$, then with probability at least $1-\delta$, the following statements hold simultaneously:*

1. $\|G^{-1/2}\left(\frac{1}{n}F_n^\top F_n\right)G^{-1/2} - \mathrm{I}_m\|_{\mathrm{op}} \leq \varepsilon$;
2. $\left\|G^{-1/2}\left(\frac{1}{n}F_n^\top Z_n(w) - \Phi^*\phi_w\right)\right\|_2 \leq \varepsilon$.

If $n \geq \frac{C}{\varepsilon^2}\ln(1/\delta)$, then with probability at least $1-\delta$, the following statements hold simultaneously:

3. $\left|\frac{1}{n}Z_n(v)^\top Z_n(w) - \kappa(w^\top v)\right| \leq \varepsilon$;
4. $\left|\Phi^*(\phi_v)^\top G^{-1/2}\left(\frac{1}{n}G^{-1/2}F_n^\top F_n G^{-1/2} - \mathrm{I}_m\right)G^{-1/2}\Phi^*\phi_w\right| \leq \varepsilon$;
5. $\left|\Phi^*(\phi_v)^\top G^{-1}\left(\frac{1}{n}F_n^\top Z_n(w) - \Phi^*\phi_w\right)\right| \leq \varepsilon$.

The difference with Lemma 7 here is that the concentrations are pointwise in w and v, while Lemma 7 is stated uniformly over all w and v, thus requiring a larger sample complexity.

Proof. Let the function $h : \mathbb{R}^d \to \mathbb{R}^m$ be defined as in Equation (34). This proof follows similar arguments to the one of Lemma 7, where we only use Lemma 15 in the particular case $\mathcal{A} = \{\mathbb{R}^d\}$, leading to improved bounds.

1. Exactly as the first item of Lemma 7, it is a consequence of Lemma 15 with $h_1 = h_2 = h$ and $\mathcal{A} = \{\mathbb{R}^d\}$.

2. It is a consequence of Lemma 15 with $h_1 = h$, $h_2 = \phi_w$ and $\mathcal{A} = \{\mathbb{R}^d\}$.

3. It is a consequence of Lemma 15 with $h_1 = \phi_v$, $h_2 = \phi_w$ and $\mathcal{A} = \{\mathbb{R}^d\}$—which is actually Bernstein's inequality.

4. Denote $a = G^{-1/2}\Phi^*\phi_w$ and $b = G^{-1/2}\Phi^*\phi_v$. It comes that $\|a\|_2 \leq \frac{1}{\sqrt{2}}$ thanks to Equation (36). Also $\|G^{-1/2}a\|_2 \leq \sqrt{\frac{\pi}{\pi-1}}$, using the bound on $\|G^{-1}\|_{\mathrm{op}}$. We now derive similarly to the proof of item 1. of Lemma 7:

$$b^\top\left(\frac{1}{n}G^{-1/2}F_n^\top F_n G^{-1/2} - \mathrm{I}_m\right)a = \frac{1}{n}\sum_{k=1}^n (G^{-1/2}a)^\top h(x_k)h(x_k)^\top (G^{-1/2}b) \\ - \mathbf{E}[(G^{-1/2}a)^\top h(X)h(X)^\top (G^{-1/2}b)].$$

Bernstein's inequality then allows to conclude.

5. A similar Bernstein's inequality argument allows to conclude here. □

Corollary 5. *Let $\breve{w}$ be drawn uniformly at random on $\mathbb{S}_{d-1}$, independently from the training data $(x_1, \dots, x_n)$. Then there exist positive universal constants c, C such that for any $\delta \in (0, 1/2)$, if*

$$\min(n, d) \geq Cm^2 \ln(1/\delta),$$

then with probability at least $1 - \delta$, the following statements hold simultaneously:

1. $\|P_m \breve{w}\| \leq \frac{1}{4(\pi-1)\sqrt{m}}$;
2. $\breve{w}^\top w^\star_{m+1} \geq -\frac{1}{4\pi\, m}$;
3. $H_n(\breve{w}) \geq \frac{c}{m}$.

Proof. First, the condition on d along with Lemma 5 guarantees that with probability at least $1 - \delta$, the first two items of Corollary 5 hold.

For the third point, by independence between $\breve{w}$ and the training data, we can derive concentration bounds conditionally on $\breve{w}$. We thus assume in the following $\breve{w}$ fixed such that $\|P_m \breve{w}\| \leq \frac{1}{4(\pi-1)\sqrt{m}}$ and $\breve{w}^\top w^\star_{m+1} \geq -\frac{1}{4\pi\, m}$. First note that, thanks to Lemma 4, $H(\breve{w}) \geq \frac{c}{m}$ for some positive constant c.

From there, define for the remaining of the proof:

$$A_n = G^{-1/2} \left(\frac{1}{n} F_n^\top F_n \right) G^{-1/2}$$
$$u = G^{-1/2} \Phi^* \phi_{\breve{w}}; \qquad u_n = \frac{1}{n} G^{-1/2} F_n^\top Z_n(\breve{w})$$
$$v = G^{-1/2} \Phi^* \phi_{w^\star_{m+1}}; \qquad v_n = \frac{1}{n} G^{-1/2} F_n^\top Z_n(w^\star_{m+1}).$$

Assume now that $n \geq C \max\left(\frac{1}{\varepsilon^2}, \frac{m}{\varepsilon}\right) \ln(1/\delta)$ for some universal constant C and $\varepsilon \in (0, 1/2)$. Lemma 13 applied to the two fixed vectors $\breve{w}$ and $w^\star_{m+1}$ implies with probability at least $1 - \delta$:

1. $\max\left(\|A_n - \mathrm{I}_m\|_{\mathrm{op}}, \|u - u_n\|_2, \|v - v_n\|_2\right) \leq \sqrt{\varepsilon}$;
2. $\left| \frac{1}{n} Z_n(w^\star_{m+1})^\top Z_n(\breve{w}) - \kappa(\breve{w}^\top w^\star_{m+1}) \right| \leq \varepsilon$;
3. $\max\left(|v^\top (A_n - \mathrm{I}_m) u|, |(v - v_n)^\top u|, |v^\top (u - u_n)| \right) \leq \varepsilon$.

By definition, we have

$$H_n(\breve{w}) - H(\breve{w}) = \frac{1}{n} Z_n(w^\star_{m+1})^\top Z_n(\breve{w}) - \kappa(\breve{w}^\top w^\star_{m+1}) + v^\top u - v_n^\top A_n^{-1} u_n.$$

The first difference is directly bounded by ε, thanks to the second item above. For the second difference, a technical development yields

$$\begin{aligned} v^\top u - v_n^\top A_n^{-1} u_n = {} & (v - v_n)^\top u + v^\top (\mathrm{I}_m - A_n^{-1}) u - (v - v_n)^\top (\mathrm{I}_m - A_n^{-1}) u \\ & + v^\top (u - u_n) - (v - v_n)^\top (u - u_n) - v_n^\top (\mathrm{I}_m - A_n^{-1})(u - u_n). \end{aligned}$$

Now using the identity $\mathrm{I}_m - A_n^{-1} = A_n - \mathrm{I}_m - (A_n - \mathrm{I}_m)^2 A_n^{-1}$ for the last term, along with the fact that $\max(\|v\|, \|u\|) \leq \frac{1}{\sqrt{2}}$ and $\|A_n^{-1}\|_{\mathrm{op}} \leq 2$, it comes with the items 1 and 3 above that:

$$|v^\top u - v_n^\top A_n^{-1} u_n| \leq (5 + 2\sqrt{2})\varepsilon.$$

Finally, we thus have $H_n(\breve{w}) \geq H(\breve{w}) - (6 + 2\sqrt{2})\varepsilon$. Since $H(\breve{w}) \geq \frac{c}{m}$ for some positive constant c, choosing $\varepsilon = \frac{c_0}{m}$ for a small enough universal constant $c_0 > 0$ then allows to conclude. □

D.4 Proof of Theorem 2

Consider the positive universal constants $C, c, \underline{D}, \overline{D}$ and m_0 appearing in Proposition 2 and Corollary 5.

Let $\varepsilon \in (0,1)$ a small universal constant to be fixed later. Under the stated regime in n and d, we have thanks to Corollary 5, with probability at least $1-\delta/2$, that the following events simultaneously hold for some positive constant c:

1. $\|P_m \mathsf{w}^\circ_{m+1}(0)\| \leq \frac{1}{4(\pi-1)\sqrt{m}}$;
2. $\mathsf{w}^\circ_{m+1}(0)^\top w^\star_{m+1} \geq -\frac{1}{4\pi\, m}$;
3. $H_n(\mathsf{w}^\circ_{m+1}(0)) \geq \frac{c}{m}$.

Moreover, thanks to Proposition 2, the following events also hold with probability at least $1-\delta/2$ for any $w \in \mathcal{T}(c)$:

3. $\underline{D} \leq D_n(w) \leq \overline{D}$;
4. $\inf_{\psi \in \Psi_n(w)} w^{\star\top}_{m+1} P_{w^\perp} \psi \geq c(1 - w^\top w^\star_{m+1})$;
5. $\sup_{\psi \in \Psi_n(w)} \|P_m P_{w^\perp} \psi\| \leq C_0 \left(\frac{1}{\sqrt{m}} + \|P_m w\| \right) + \varepsilon$.

Assume these five random events all hold. From that point, the remaining of the proof follows exactly the same lines as in the population case, given in Appendix C.4. The chain rule still applies (thanks to Bolte & Pauwels, 2021), so that $\frac{H_n(\mathsf{w}^\circ_{m+1}(t))^2}{D_n(\mathsf{w}^\circ_{m+1}(t))}$ increases over time. From there, the reparametrized time variable given by

$$\gamma : s \mapsto \int_0^s \frac{D_n(\mathsf{w}^\circ_{m+1}(u))}{H_n(\mathsf{w}^\circ_{m+1}(u))} \mathrm{d}u \qquad \text{and} \qquad \widetilde{\mathsf{w}}(s) \coloneqq \mathsf{w}^\circ_{m+1}(\gamma(s))$$

satisfies almost anywhere, as long as $\widetilde{\mathsf{w}}(s) \in \mathcal{T}(c)$, that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}s} w^{\star\top}_{m+1} \widetilde{\mathsf{w}}(s) &\geq c(1 - w^{\star\top}_{m+1} \widetilde{\mathsf{w}}(s)) \\ \frac{\mathrm{d}}{\mathrm{d}s} \|P_m \widetilde{\mathsf{w}}(s)\| &\leq C_0 \left(\frac{1}{\sqrt{m}} + \|P_m w\| \right) + \varepsilon. \end{aligned} \tag{48}$$

Defining $\overline{s} \coloneqq \inf \{ s \geq 0 : \widetilde{\mathsf{w}}(s) \notin \mathcal{T}(c) \}$, $(1 - w^{\star\top}_{m+1} \widetilde{\mathsf{w}}(s))$ is increasing on $[0, \overline{s}]$ and by a Grönwall argument, for any $s \in [0, \overline{s}]$

$$1 - w^{\star\top}_{m+1} \widetilde{\mathsf{w}}(s) \leq \frac{5}{4} e^{-cs}. \tag{49}$$

Moreover the second point implies, with $\|P_m \widetilde{\mathsf{w}}(s)\| \leq \sqrt{2(1 - w^{\star\top}_{m+1} \widetilde{\mathsf{w}}(s))}$, that for any $s \in [0, \overline{s}]$:

$$\begin{aligned} \|P_m \widetilde{\mathsf{w}}(s)\| &\leq 2 \min \left(\left(\frac{1}{\sqrt{m}} + \frac{\varepsilon}{C_0} \right) e^{C_0 s}, e^{-cs/2} \right) \\ &= 2 \exp \left(\frac{c/2}{C_0 + c/2} \ln \left(\frac{1}{\sqrt{m}} + \frac{\varepsilon}{C_0} \right) \right). \end{aligned}$$

Thus, for a large enough choice of m_0 and small enough, constant, choice of ε, $\|P_m \widetilde{\mathsf{w}}(s)\| \leq \frac{c}{2}$ on $[0, \overline{s}]$, so that $\overline{s} = \infty$, i.e., $w_{m+1}(t) \in \mathcal{T}(c)$ for any $t \geq 0$.

Similarly to the proof of Theorem 1, we can again conclude that for a small enough constant η and $s_\eta = \inf\{ s \geq 0 \mid w^{\star\top}_{m+1} \widetilde{\mathsf{w}}(s) \geq 1 - \eta \}$,

$$\gamma'(s) = \frac{D_n(w_{m+1}(s))}{H_n(w_{m+1}(s))} \leq \frac{\overline{D}}{2\underline{D}} \quad \text{for any } s \geq s_\eta,$$

and $\gamma(s_\eta) \leq C_0 m$ for a large enough universal constant C_0. In particular, for any $t \geq \gamma(s_\eta)$, Equation (48) then implies:

$$1 - \mathsf{w}^\circ_{m+1}(t)^\top w^\star_{m+1} \leq \exp \left(-\frac{2c\underline{D}}{\overline{D}} (t - \gamma(s_\eta)) \right),$$

which concludes on the first part of Theorem 2.

Letting now $\mathcal{T} = \{[w_1^{\star\top}, \dots, w_m^{\star\top}, w_{m+1}^\top] \mid w_{m+1} \in \mathcal{T}(c)\}$, the trajectory W° is obviously confined to $\mathcal{T}$. Moreover, $\mathcal{T}$ satisfies Assumption 1: it is obviously compact. Moreover, for any $W \in \mathcal{T}$,

$$\Sigma(W, \mathbf{X}) = [F_n, Z_n(w)^\top]^\top.$$

Note that from Lemma 8, under the same random event, $F_n^\top F_n$ is invertible and $D_n(w) > 0$ for any $w \in \mathcal{T}(c)$. Those two conditions imply that the family $(Z_n(w_1^\star), \dots, Z_n(w_m^\star), Z_n(w))$ is linearly independent for any $w \in \mathcal{T}(c)$ or, equivalently, the matrix $[F_n, Z_n(w)^\top]^\top$ is full rank. In other words, $\lambda_{\min}\left(\Sigma(W, \mathbf{X})\Sigma(W, \mathbf{X})^\top\right) > 0$ for any $W \in \mathcal{T}$. By compactness and continuity of the considered function, this directly yields that $\mathcal{T}$ satisfies Assumption 1.

□

E Additional lemmas

Lemma 14. *Consider $T \in \mathbb{R}_+$ and any positive sequence $(\alpha_k)_{k\in\mathbb{N}}$ such that $\alpha_k \overset{k\to\infty}{\longrightarrow} 0$, and corresponding solutions W^{α_k} of Equation (3) such that $\|w_i(0)\| = \begin{cases} \|w_i^\star\| & \text{for } i \in [m] \\ \alpha & \text{otherwise} \end{cases}$.*

Suppose all solutions $\mathsf{W}^\circ(t)$ of Equation (4) are contained within some set $\mathcal{T}$ satisfying Assumption 1 on $[0, T]$. Then there exists an augmented set $\mathcal{T}_+$ satisfying Assumption 1, such that for k large enough, the trajectory $\left(\mathsf{W}^{\alpha_k}(\alpha_k^2 t)\right)_{t\in[0,T]}$ is included within $\mathcal{T}_+$.

Proof. Observe that by continuity of $W \mapsto \lambda_{\min}\left(\frac{1}{n}\Sigma(W, \mathbf{X})\Sigma(W, \mathbf{X})^\top\right)$ and compactness, we can choose a small enough $\varepsilon > 0$ such that the set $\mathcal{T}_+ = \mathcal{T} + \overline{B}(0, \varepsilon)$ also satisfies Assumption 1. From there, define for any k, $\tau_k = \inf\left\{t \geq 0 \mid \mathsf{W}^{\alpha_k}(\alpha_k^2 t) \notin \mathcal{T}_+\right\}$. Since $\mathsf{W}^{\alpha_k}(0) = \mathsf{W}^\circ(0) \in \mathcal{T}$ by definition, $\tau_k \geq 0$. Then consider $\tau = \liminf_k \tau_k \geq 0$ and assume without loss of generality – by extraction – that $\tau = \lim_k \tau_k$.

Assume here that $\tau \leq T$ and define $\tau_\delta = \max(0, \tau - \delta)$ for some $\delta > 0$. For large enough k, $|\tau_k - \tau| \leq \delta$, so that we can apply Proposition 1 on $[0, \tau_\delta]$; i.e., up to extraction, $\mathsf{W}^{\alpha_k}(\alpha_k^2 \cdot)$ converges uniformly on $[0, \tau_\delta]$ to some solution W° of Equation (4). In particular, for k large enough, $\mathsf{W}^{\alpha_k}(\alpha_k^2 t) \in \mathcal{T} + B(0, \varepsilon/3)$ for any $t \in [0, \tau_\delta]$. Moreover, thanks to Corollary 1, $\mathsf{W}^{\alpha_k}(\alpha_k^2 \cdot)$ is uniformly Lipschitz[8] on $[0, \max(\tau_k, \tau_\delta)]$. In particular, if we denote L the Lipschitz constant and take $\delta < \varepsilon/(3L)$, the condition $\mathsf{W}^{\alpha_k}(\alpha_k^2 \tau_\delta) \in \mathcal{T} + B(0, \varepsilon/3)$ implies that for k large enough, $\mathsf{W}^{\alpha_k}(\alpha_k^2 \tau_k) \in \mathcal{T} + B(0, \frac{2\varepsilon}{3})$, which contradicts the definition of τ_k. By contradiction, we here showed that $\tau > T$, i.e., $\liminf_k \tau_k > T$, which directly implies Lemma 14. □

E.1 Useful concentration bounds

The different uniform concentration bounds of Appendix D.3.1 rely on the following key concentration lemma for empirical processes, which derives from classical chaining and VC subgraph classes arguments (see van der Vaart & Wellner, 1996, Chapter 2).

For a collection of subsets $\mathcal{A}$ and a family of functions $\mathcal{F}$, we denote their VC dimension and VC subgraph dimension by $\mathrm{VC}(\mathcal{A})$ and $\mathrm{VC}(\mathcal{F})$, respectively, following the standard definitions of van der Vaart & Wellner (1996, Chapter 2.6).

Lemma 15. *Let $\mathcal{F}$ be a family of functions from $\mathbb{R}^d$ to $\mathbb{R}$, defined as*

$$\mathcal{F} = \left\{x \mapsto (u^\top h_1(x))(h_2(x)^\top v)\mathbb{1}_{x\in A} \mid u \in \mathbb{S}_{p_1-1}, v \in \mathbb{S}_{p_2-1}, A \in \mathcal{A}\right\}$$

where $\mathcal{A}$ is a VC class of subsets of $\mathbb{R}^d$ and $h_1 : \mathbb{R}^d \to \mathbb{R}^{p_1}$, $h_2 : \mathbb{R}^d \to \mathbb{R}^{p_2}$ satisfy: $\max(\|h_1(X)\|_{\psi_2}, \|h_2(X)\|_{\psi_2}) \leq C_0$ for some universal constant C_0 when $X \sim \mathcal{N}(0, \mathrm{I}_d)$.

[8] Corollary 1 is stated for a fixed interval $[0, T]$ for simplicity, but it can directly be extended to allow for different time intervals of the type $[0, \tau_k]$, as long as the weights are in $\mathcal{T}_+$.

Then, there is a universal constant C such that for any $\varepsilon, \delta \in (0, 1/2)$, if $n \geq C\varepsilon^{-2}\ln(1/\varepsilon)^2\left(\mathrm{VC}(\mathcal{A}) + p_1 + p_2 + \ln(\frac{1}{\delta})\right)$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbf{E}\left[f(X)\right]\right| \geq \varepsilon\right) \leq \delta,$$

where the probability is over $X_1, \ldots, X_n$ drawn i.i.d. from $\mathcal{N}(0, \mathrm{I}_d)$.

Moreover, if $\mathcal{A} = \{\mathbb{R}^d\}$, $n \geq C\varepsilon^{-2}\left(p_1 + p_2 + \ln(\frac{1}{\delta})\right)$ samples are sufficient to reach the same concentration bound.

Proof. We first fix $u \in \mathbb{S}_{p_1-1}, v \in \mathbb{S}_{p_2-1}$ and denote for convenience $F(x) = (u^\top h_1(x))(h_2(x)^\top v)$, $\mathcal{F}_{u,v} = \{x \mapsto F(x)\mathbb{1}_{x\in A} \mid A \in \mathcal{A}\}$.

Bound on expected supremum. Theorems 2.14.1 and 2.6.7 of van der Vaart & Wellner (1996) directly yield that for some universal constant C,

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}_{u,v}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbf{E}[f(X)]\right|\right] \leq \frac{C}{\sqrt{n}}\|F\|_{L^2(\mathcal{N}(0,\mathrm{I}_d))} \cdot \int_0^1 \sqrt{1 + \ln\left(C \cdot \mathrm{VC}(\mathcal{F}_{u,v})(16e)^{\mathrm{VC}(\mathcal{F}_{u,v})}\left(\frac{1}{\varepsilon}\right)^{2\mathrm{VC}(\mathcal{F}_{u,v})-2}\right)}\mathrm{d}\varepsilon.$$

Since $\max(\|h_1(X)\|_{\psi_2}, \|h_1(X)\|_{\psi_2}) \leq C_0$, $\|F(X)\|_{\psi_1} \leq C_0^2$. In particular, its second moment, i.e., $\|F\|_{L^2(\mathcal{N}(0,\mathrm{I}_d))}$, is of constant order. A simple Cauchy-Schwarz argument then yields that for some universal constant C,

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}_{u,v}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbf{E}[f(X)]\right|\right] \leq C\sqrt{\frac{\mathrm{VC}(\mathcal{F}_{u,v}) + 1}{n}}.$$

Moreover, by definition of $\mathcal{F}_{u,v}$, $\mathrm{VC}(\mathcal{F}_{u,v}) \leq \mathrm{VC}(\mathcal{A})$ (see e.g., van der Vaart & Wellner, 1996, Lemma 2.6.18), so that

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}_{u,v}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbf{E}[f(X)]\right|\right] \leq C\sqrt{\frac{\mathrm{VC}(\mathcal{A}) + 1}{n}}.$$

Truncated decomposition. Let $M \geq 0$ to be fixed later. Define $\mathcal{F}_M = \{x \mapsto f(x)\mathbb{1}_{F(x)\in[-M,M]} \mid f \in \mathcal{F}_{u,v}\}$ and

$$\mu_{n,M} \coloneqq \mathbf{E}\left[\sup_{f\in\mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbf{E}[f(X)]\right|\right].$$

The exact same argument as above first yields that $\mu_{n,M} \leq C\sqrt{\frac{\mathrm{VC}(\mathcal{A})+1}{n}}$ for some constant C. Moreover,

$$\begin{aligned}\sup_{f\in\mathcal{F}_{u,v}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbf{E}[f(X)]\right| &\leq \sup_{f\in\mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbf{E}[f(X)]\right| + \frac{1}{n}\sum_{i=1}^{n}|F(X_i)|\mathbb{1}_{|F(X_i)|\geq M} \\ &\quad + \mathbf{E}[|F(X)|\mathbb{1}_{|F(X)|\geq M}] \\ &\leq \sup_{f\in\mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbf{E}[f(X)]\right| + 2\mathbf{E}[|F(X)|\mathbb{1}_{|F(X)|\geq M}] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}|F(X_i)|\mathbb{1}_{|F(X_i)|\geq M} - \mathbf{E}[|F(X)|\mathbb{1}_{|F(X)|\geq M}].\end{aligned}$$

Let us now bound each of these terms with high probability. First, Theorem 2.14.25 of van der Vaart & Wellner (1996) yields for any $t \geq 0$ that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbf{E}[f(X)]\right| \geq C(\mu_{n,M} + t)\right) \leq \exp\left(-Cn\min\left(\frac{t^2}{M^2}, \frac{t}{M}\right)\right).$$

Taking $t = M\sqrt{\frac{\ln(1/\delta)}{Cn}}$ and $n \geq \frac{\ln(1/\delta)}{C}$—so that $\frac{t}{M} \leq 1$—it yields that with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbf{E}[f(X)]\right| \leq C\sqrt{\frac{\mathrm{VC}(\mathcal{A})+1}{n}} + CM\sqrt{\frac{\ln(1/\delta)}{n}} \tag{50}$$

for some universal constant C.

For the second term, it directly comes from the sub-exponential tail of $F(X)$, i.e.,

$$\begin{aligned}\mathbf{E}[|F(X)|\mathbb{1}_{|F(X)|\geq M}] &\leq \sqrt{\mathbf{E}[F(X)^2]}\cdot\sqrt{\mathbb{P}(|F(X)|\geq M)} \\ &\leq Ce^{-cM},\end{aligned} \tag{51}$$

for positive universal constants C and c. Lastly for the third term, note that $|F(X_i)|\mathbb{1}_{|F(X_i)|\geq M} - \mathbf{E}[|F(X)|\mathbb{1}_{|F(X)|\geq M}]$ is also C sub-exponential for some constant C. So that by classical Bernstein inequality (Vershynin, 2018, Theorem 2.8.1) and if $n \geq C\ln(1/\delta)$, with probability at least $1-\delta$:

$$\frac{1}{n}\sum_{i=1}^n |F(X_i)|\mathbb{1}_{|F(X_i)|\geq M} - \mathbf{E}[|F(X)|\mathbb{1}_{|F(X)|\geq M}] \leq \sqrt{\frac{\ln(1/\delta)}{n}}. \tag{52}$$

Gathering altogether Equations (50) to (52) with $M = C\ln(1/\varepsilon)$ for a large enough universal constant C finally yields that if $n \geq C\ln(1/\delta)$, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}_{u,v}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbf{E}[f(X)]\right| \leq \frac{1}{2}\varepsilon + C\sqrt{\frac{\mathrm{VC}(\mathcal{A})+1}{n}} + C\ln(1/\varepsilon)\sqrt{\frac{\ln(1/\delta)}{n}}. \tag{53}$$

Covering net argument. Now we have proven Equation (53) for a fixed choice of vectors u, v, it remains to conclude by a covering net argument. For that, consider $\mathcal{N}$ a $\frac{1}{4}$-net of $\mathbb{S}_{p_1-1}\times\mathbb{S}_{p_2-1}$, with cardinality smaller than $9^{p_1+p_2}$. From Equation (53), a simple union bound then yields that if $n \geq C\ln(9^{p_1+p_2}/\delta)$ with probability at least $1-\delta$,

$$\sup_{(u,v)\in\mathcal{N}}\sup_{f\in\mathcal{F}_{u,v}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbf{E}[f(X)]\right| \leq \frac{1}{2}\varepsilon + C\sqrt{\frac{\mathrm{VC}(\mathcal{A})+1}{n}} + C\ln(1/\varepsilon)\sqrt{\frac{\ln(9^{p_1+p_2}/\delta)}{n}}.$$

In particular, if $n \geq C\varepsilon^{-2}\ln(1/\varepsilon)^2(\mathrm{VC}(\mathcal{A}) + p_1 + p_2 + \ln(1/\delta))$ for a large enough universal constant C, then with probability at least $1-\delta$,

$$\sup_{(u,v)\in\mathcal{N}}\sup_{f\in\mathcal{F}_{u,v}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbf{E}[f(X)]\right| \leq \varepsilon. \tag{54}$$

Moreover, writing for shortness $M_A = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i\in A}h_1(X_i)h_2(X_i)^\top - \mathbf{E}[\mathbb{1}_{X\in A}h_1(X)h_2(X)^\top]$ for any $A \in \mathcal{A}$,

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbf{E}[f(X)]\right| = \sup_{A\in\mathcal{A}}\sup_{\substack{u\in\mathbb{S}_{p_1-1}\\ v\in\mathbb{S}_{p_2-1}}} u^\top M_A v.$$

Fix $A \in \mathcal{A}$ and let $(u_A, v_A) \in \mathbb{S}_{p_1-1}\times\mathbb{S}_{p_2-1}$ be such that $\sup_{\substack{u\in\mathbb{S}_{p_1-1}\\ v\in\mathbb{S}_{p_2-1}}} u^\top M_A v = u_A^\top M_A v_A$. Letting (u', v') be such that $\|(u_A, v_A) - (u', v')\|_2 \leq \frac{1}{4}$, it then comes that

$$\begin{aligned}u'^\top M_A v' &= u_A^\top M_A v_A - (u_A - u')^\top M_A v_A - u'^\top M_A(v_A - v') \\ &\geq (1 - \|u_A - u'\| - \|v_A - v'\|)\sup_{\substack{u\in\mathbb{S}_{p_1-1}\\ v\in\mathbb{S}_{p_2-1}}} u^\top M_A v \\ &\geq \frac{1}{2}\sup_{\substack{u\in\mathbb{S}_{p_1-1}\\ v\in\mathbb{S}_{p_2-1}}} u^\top M_A v.\end{aligned}$$

In particular, this implies that

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i)-\mathbf{E}[f(X)]\right|\leq 2\sup_{A\in\mathcal{A}}\sup_{(u,v)\in\mathcal{N}}u^\top M_A v,$$

which concludes the proof of Lemma 15, thanks to Equation (54).

Case $\mathcal{A}=\{\mathbb{R}^d\}$. This particular case is much simpler, as $\mathcal{F}_{u,v}$ is now a single function. In consequence, instead of Equation (53), a simple Bernstein inequality yields that when $n\geq C\ln(1/\delta)$, with probability at least $1-\delta$

$$\sup_{f\in\mathcal{F}_{u,v}}\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i)-\mathbf{E}[f(X)]\right|\leq C\sqrt{\frac{\ln(1/\delta)}{n}}.$$

The covering argument then allows to conclude similarly, requiring only a sample complexity $n\geq C\varepsilon^{-2}(p_1+p_2+\ln(1/\delta))$. □

Of note, we could probably improve—or even get rid of—the term $\ln(1/\varepsilon)^2$ in the required sample complexity of Lemma 15, since Theorem 2.14.25 of van der Vaart & Wellner (1996) actually yields a probability bounded as $\exp\left(-Cn\min\left(\frac{t^2}{\sigma^2_{\mathcal{F}_M}},\frac{t}{M}\right)\right)$ for some variance term $\sigma^2_{\mathcal{F}_M}$. We use the obvious bound $\sigma^2_{\mathcal{F}_M}\leq M^2$ here, but we believe a tighter bound is possible. However, it is sufficient for our purpose, since we will use Lemma 15 with ε of constant order.

Lemma 16 (adapted from Soltanolkotabi, 2017a, Lemma 5.5)**.** *Let $w_0^\star$ be a fixed vector in $\mathbb{S}_{d-1}$ and $\varepsilon_1,\varepsilon_2,\delta\in(0,1/2)$. There exists a positive universal constant C, such that if $n\geq\frac{C}{\varepsilon_1}(d+\ln(1/\delta))$, then with probability at least $1-\delta$,*

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i^\top P_{w_0^\star{}^\perp}w\right)^2\mathbb{1}_{\frac{1}{2}|x_i^\top w_0^\star|\leq|x_i^\top P_{w_0^\star{}^\perp}w|}\leq C\left(\varepsilon_2+\varepsilon_1\right)\|P_{w_0^\star{}^\perp}w\|^2,$$

uniformly over all $w\in\mathbb{R}^d$ such that $\|P_{w_0^\star{}^\perp}w\|\leq\varepsilon_2$.

Proof. This is actually a direct application[9] of Soltanolkotabi (2017a, Lemma 5.5) with, following their notations,

- $\mathcal{C}=\{w_0^\star\}^\perp$;
- $\varepsilon=\frac{\varepsilon_2}{\frac{1}{2}+\varepsilon_2}$;
- $\delta=\sqrt{\varepsilon_1}$;
- $w^*=\left(\frac{1}{2}+\varepsilon_2\right)w_0^\star$.

□

[9]Note that there appear to be two typos in the statement of Lemma 5.5 in Soltanolkotabi (2017a): the term $\sqrt{\frac{21}{20}}\varepsilon$ should read $\sqrt{\frac{21}{20}\varepsilon}$, and the sample-size requirement should be $n\geq c\delta^{-2}\omega^2(\mathcal{C}\cap\mathbb{S}_{d-1})$. The first typo appears to be inherited from a similar typo in the proof on page 30 of Soltanolkotabi (2017b). Moreover, the sample complexity required in Soltanolkotabi (2017b) is indeed $n\geq c\delta^{-2}\omega^2(\mathcal{C}\cap\mathbb{S}_{d-1})$.