Attention-based PCA

Rodrigo Maulen-Soto[♠] & Claire Boyer[♦♣]

May 11, 2026

[♠] Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, 75005, Paris, France
[♦] Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France
[♣] Institut Universitaire de France

Abstract

We study attention mechanisms through the lens of a canonical unsupervised problem: principal component analysis (PCA). We show that, when trained on Gaussian data, both softmax and linear attention layers learn parameters that align with the principal eigenvectors of the covariance matrix, thereby establishing a direct and explicit connection with PCA.

Our analysis covers both finite and infinite prompt regimes. In the infinite-prompt limit, we prove convergence to globally optimal solutions aligned with the leading spectral direction, while in the finite-prompt setting we show that the same behavior emerges up to sampling effects. We further extend the analysis to an in-context setting with spiked Wishart covariances, where attention successfully recovers the underlying signal direction.

These results demonstrate that attention inherently performs PCA-like computations under unsupervised objectives, providing a theoretical foundation for its representation-learning capabilities.

1 Introduction

Attention-based models (Bahdanau et al., 2015), in particular Transformers (Vaswani et al., 2017), have become central in modern machine learning, achieving state-of-the-art results in natural language processing (Devlin et al., 2018; Bubeck et al., 2023; Luong et al., 2015; Bahdanau et al., 2016) and computer vision (Dosovitskiy et al., 2020; Liu et al., 2021; Ramachandran et al., 2019). The core attention mechanism computes weighted combinations of token representations based on pairwise interactions, allowing the model to capture long-range dependencies without necessarily relying on fixed positional locality.

A full theoretical understanding of attention-based mechanisms remains incomplete, due to both the architectural complexity and the diversity of tasks they successfully address. A promising research direction toward bridging this gap is to identify key features of real-world problems and study minimal, canonical tasks that retain their core statistical structure, while remaining amenable to rigorous analysis. Notable recent efforts in this direction include Ahn et al. (2023); von Oswald et al. (2023); Yang et al. (2025); Zhang et al. (2024); Li et al. (2024, 2023). However, most existing work focuses on supervised settings, particularly in-context learning (von Oswald et al., 2023; Zhang et al., 2024; Garg et al., 2023; Li et al., 2023; Furuya et al., 2024), where the goal is to predict the output corresponding to a new query given a prompt consisting of input–output pairs. By contrast, only limited attention has been paid to unsupervised settings, with Maulen-Soto et al. (2025) providing one of the few studies exploring the behavior of attention layers in tasks such as clustering.

In this paper, we contribute to this line of research by studying attention layers in an unsupervised setting through the lens of Principal Component Analysis (PCA; Pearson, 1901; Hotelling, 1933; Jolliffe, 2002; Golub and Van Loan, 1996). PCA underlies many approaches to dimensionality reduction and feature extraction in statistical learning by relying on the estimation of principal components, i.e., the leading eigenvectors of the data covariance matrix. Understanding how attention layers can extract these spectral directions sheds light on the representation-learning capabilities of Transformers under unsupervised objectives.

Contributions. We consider simplified softmax and linear attention layers with rank-one attention parameters. The input tokens are assumed to be i.i.d. samples from a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$. Our main contributions can be then summarized as follows.

- **A tractable linear attention model.** As a warm-up, we provide an explicit analysis for a linear attention model. We show that attention parameters can be trained to recover the top eigenvector of the data covariance, building intuition for the mechanisms at play (full details are provided in the appendix).
- **Softmax attention: from infinite to finite prompts.** As for the softmax model, the training dynamics are more delicate to handle due to the nonlinearity of the softmax. We first analyze the tractable infinite-prompt limit, characterizing a global minimizer aligned with the top eigenvector, and then transfer this understanding to the finite-prompt regime. This requires refined concentration arguments, controlling both the optimization landscape (through concentration of critical points) and the training dynamics, ultimately showing concentration around the same solution up to sampling effects.
- **In-context learning under structured covariance models.** We extend the analysis to a distributional setting where the covariance follows a spiked Wishart model. In this setting, we show that attention recovers the spike direction both in finite and infinite prompt regimes.
- **Attention performs PCA.** Taken together, our results show that rank-one attention layers can be trained in an unsupervised manner to recover the principal component of the input tokens. Under this Gaussian setting, attention-based architectures are shown to detect the underlying structure of the data. This provides a clear theoretical connection between attention mechanism and classical spectral methods, positioning attention as an implicit, optimization-driven analogue of PCA.

Organization. Section 2 introduces the problem and studies the convergence of a simplified softmax attention layer in both finite and infinite prompt settings. Section 3 characterizes the distribution of the resulting encodings and their relation across both regimes. In Section 4, we extend the framework to an in-context PCA learning setting. The appendices gather results for a simplified linear attention model recovering the principal component, along with proofs of the main results, technical lemmas, and additional numerical experiments.

2 Training dynamics of a softmax attention layer

2.1 Rank-one softmax attention: model and risk functions

Setting. Let $\mathbb{X} = (X_1, \dots, X_L) \in \mathbb{R}^{d \times L}$ denote an input prompt made of Gaussian tokens where $X_\ell \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$, with $\Sigma \in \mathbb{R}^{d \times d}$ a symmetric and definite positive matrix. We consider a simplified softmax attention head acting on such a prompt, defined for $1 \leq \ell \leq L$, by

$$T_L^{\text{soft},\mu}(\mathbb{X})_\ell = \sum_{k=1}^{L} \text{softmax}(\lambda X_\ell^\top \mu\mu^\top X_k) X_k, \tag{1}$$

the softmax function being applied row-wise. In this simplified architecture, the vector $\mu \in \mathbb{R}^d$ denotes the only attention parameter. This formulation arises from standard architectures where the value matrix is taken to be the identity, and the key and query matrices reduce to row vectors. This induces the rank-one structure $K^\top Q = \mu\mu^\top$, which restricts the interaction mechanism of the attention layer and simplifies the analysis. We focus in the main text on the softmax attention head, whose theoretical analysis is more delicate due to the softmax nonlinearity, and provide analogous results for a simplified linear counterpart in the appendices.

In order to measure the quality of embedding performed by an attention layer, we consider the following theoretical population risk

$$\mathcal{R}_{\text{soft},L}(\mu) = \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{E}\left[\| X_\ell - T_L^{\text{soft},\mu}(\mathbb{X})_\ell \|^2 \right] = \mathbb{E}\left[\| X_1 - T_L^{\text{soft},\mu}(\mathbb{X})_1 \|^2 \right]. \tag{2}$$

The objective can be viewed as a reconstruction problem, in which the attention mechanism approximates a token X_1 as a combination of input tokens leveraging information from the entire prompt. In what follows, we assume for simplicity that training is performed through the population risk minimization. Although this idealization departs from the practical procedure, it does not affect the nature of our results, and an empirical counterpart could be handled similarly.

Measure-based formalism. To understand the training dynamics of a softmax attention layer, we lean on a measure-based formalism, see e.g. Boursier and Boyer (2025). A self-attention rank-one layer with attention parameter $\mu \in \mathbb{R}^d$ can be seen as an operator acting on measures:

$$T^{\lambda,\mu} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d,$$
$$(\nu, z) \mapsto T^{\lambda,\mu}[\nu](z) = \frac{\int_{\mathbb{R}^d} \exp(\lambda z^\top \mu\mu^\top z') z' \, d\nu(z')}{\int_{\mathbb{R}^d} \exp(\lambda z^\top \mu\mu^\top z') \, d\nu(z')},$$

so that when the prompt $\mathbb{X} = (X_1, \dots, X_L)$ is encoded by its associated empirical measure

$$\hat{\nu}_L = \frac{1}{L} \sum_{\ell=1}^{L} \delta_{X_\ell},$$

one exactly retrieves the softmax attention formula (1), i.e., $T^{\lambda,\mu}[\hat{\nu}_L](X_\ell) = T_L^{\mathrm{soft},\mu}(\mathbb{X})_\ell$.

A key observation is that when the prompt length grows, the empirical attention operator converges to its infinite-prompt counterpart, i.e.,

$$T^{\lambda,\mu}[\hat{\nu}_L](z) \xrightarrow[L\to\infty]{a.s.} T^{\lambda,\mu}[\nu](z),$$

with $\hat{\nu}_L$ still the empirical measure associated with L i.i.d. tokens drawn according to ν. Moreover, when the token distribution ν is Gaussian, the infinite-prompt softmax attention becomes a linear operator. Specifically, it was shown in Boursier and Boyer (2025, Lemma 2.1) (see also Castin et al., 2025, Lemma 4.1), that if $\nu = \mathcal{N}(0, \Sigma)$ then,

$$T^{\lambda,\mu}[\nu](z) = \lambda \Sigma \mu\mu^\top z,$$

and consequently, the finite-prompt estimator converges almost surely to the linear operator

$$T_L^{\mathrm{soft},\mu}(\mathbb{X})_1 \xrightarrow[L\to\infty]{a.s.} T_\infty^{\mathrm{soft},\mu}(X_1) := \lambda \Sigma \mu\mu^\top X_1. \tag{3}$$

The same almost sure convergence holds for the corresponding gradient (w.r.t. μ) and Hessian of $T_L^{\mathrm{soft},\mu}$ (see Lemma 6). This result shows that in the large-prompt regime, the nonlinear softmax attention layer behaves effectively as a linear operator acting on the token distribution. This convergence enables transferring optimization analyses from the linear operator to softmax attention when the prompt length is sufficiently large. We therefore introduce the theoretical risk of the infinite-prompt layer as

$$\mathcal{R}_{\mathrm{soft},\infty}(\mu) = \mathbb{E}[\|X_1 - T_\infty^{\mathrm{soft},\mu}(X_1)\|^2],$$

which admits the closed form

$$\mathcal{R}_{\mathrm{soft},\infty}(\mu) = \mathrm{tr}(\Sigma) - 2\lambda b + \lambda^2 ab. \tag{4}$$

with $a = a(\mu) = \mu^\top \Sigma \mu$ and $b = b(\mu) = \mu^\top \Sigma^2 \mu$. This reduction will be key to the analysis of the optimization landscape for infinite-prompt architectures, and in particular for the characterization of the critical points, as conducted in the next section.

2.2 Optimization analysis for infinite prompts

We characterize in what follows the critical points of the infinite-prompt risk $\mathcal{R}_{\mathrm{soft},\infty}$.

Proposition 1 (Landscape of $\mathcal{R}_{\text{soft},\infty}$)**.** *Assume that the p.s.d. covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ has a simple spectrum, i.e., distinct eigenvalues $\sigma_1 > \cdots > \sigma_d > 0$, with u_j the associated unit eigenvectors. Then all critical points of $\mathcal{R}_{\text{soft},\infty}$ are nondegenerate, and*

$$\operatorname{crit}(\mathcal{R}_{\text{soft},\infty}) = \{0\} \cup \left\{ \pm \frac{1}{\sqrt{\lambda \sigma_j}}\, u_j : j = 1, \dots, d \right\}$$

Moreover,

1. *the point 0 is a strict local maximum;*
2. *the points $\pm \frac{1}{\sqrt{\lambda\sigma_j}} u_j$, for $j = 2, \dots, d$, are strict saddles;*
3. *the points $\pm \frac{1}{\sqrt{\lambda\sigma_1}} u_1$ are global minimizers of $\mathcal{R}_{\text{soft},\infty}$.*

Proposition 2 (Global minimization of $\mathcal{R}_{\text{soft},\infty}$)**.** *For almost every initialization $\mu_0 \in \mathbb{R}^d$, the solution of*

$$\begin{cases} \dot{\mu}_\infty(t) &= -\nabla \mathcal{R}_{\text{soft},\infty}(\mu_\infty(t)), \\ \mu_\infty(0) &= \mu_0. \end{cases} \tag{GF$_\infty$}$$

converges to $\pm \frac{1}{\sqrt{\lambda\, \sigma_1}}\, u_1$, with (σ_1, u_1) the leading eigenpair of Σ.

Proposition 2 shows that training a rank-one softmax attention layer with an infinite-length Gaussian prompt by minimizing the population risk drives the attention parameter toward the leading principal component of the covariance matrix (up to a sign). Note that all the results and arguments established for the gradient flow in this paper extend to the discrete-time setting, provided that gradient descent is used with a sufficiently small stepsize. We adopt the gradient flow formulation for notational simplicity.

Remark 1 (PCA with k components). With the knowledge of u_1, the second principal component u_2 can be obtained via the projected gradient flow

$$\begin{cases} \dot{\mu}_\infty(t) = -\mathrm{P}_{u_1^\perp}\big(\nabla \mathcal{R}_{\text{soft},\infty}(\mu_\infty(t))\big), \\ \mu_\infty(0) = \mu_0, \end{cases} \tag{PGF$_\infty$}$$

where $\mathrm{P}_{u_1^\perp} = I_d - u_1 u_1^\top$. By Proposition 1, the only non-degenerate critical points in $u_1^\perp$ that are not strict saddles are $\pm \frac{1}{\sqrt{\lambda\sigma_2}} u_2$, and thus, by the Stable Manifold Theorem (Shub, 1987, Theorem III.7), the flow converges to these points for a generic initialization. This procedure can be iterated: projecting onto $\operatorname{Span}(u_1, \dots, u_{k-1})^\perp$ yields u_k, allowing sequential learning of all eigenvectors through the attention mechanism.

To quantify the convergence rate of the gradient flow, we leverage Łojasiewicz inequalities, which relate the decay of the objective function to the norm of its gradient near critical points. Since $\mathcal{R}_{\text{soft},\infty}$ is analytic, this inequality holds locally around each critical point, which enables us to derive explicit convergence rates toward the principal component.

Definition 1 (Łojasiewicz inequality)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function and let $\mu^\star \in \mathbb{R}^d$ be a critical point of f. We say that f satisfies the Łojasiewicz inequality at $\mu^\star$ if there exist constants $C > 0$, $\alpha \in [0, 1)$, and a neighborhood U of $\mu^\star$ such that

$$\|\nabla f(\mu)\| \geq C\, |f(\mu) - f(\mu^\star)|^\alpha \quad \text{for all } \mu \in U.$$

We denote by $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ the smallest and largest eigenvalues of a matrix A.

Proposition 3 (Local convergence rate on infinite-prompt setting)**.** *The risk $\mathcal{R}_{\text{soft},\infty}$ satisfies Łojasiewicz inequality with exponent 1/2 at each critical point. Set $\mu^\star = \pm u_1/\sqrt{\lambda\sigma_1}$, then there exist constants $t_0 \geq 0$ and $s > 0$ such that for all $t \geq t_0$,*

$$\mathcal{R}_{\text{soft},\infty}(\mu_\infty(t)) - \mathcal{R}_{\text{soft},\infty}(\mu^\star) \leq \left(\mathcal{R}_{\text{soft},\infty}(\mu_\infty(t_0)) - \mathcal{R}_{\text{soft},\infty}(\mu^\star)\right) e^{-s(t-t_0)}.$$

Besides, for every $\varepsilon > 0$ small enough, there exists $t_0 > 0$ such that

$$\|\mu_\infty(t) - \mu^\star\| = \mathcal{O}\big(e^{-(\tilde{s}-\varepsilon)(t-t_0)}\big),$$

where

$$\tilde{s} = \sigma_{\min}(\nabla^2 \mathcal{R}_{\text{soft},\infty}(\mu^\star)) = 2\lambda \min\{\sigma_2(\sigma_1 - \sigma_2), \sigma_d(\sigma_1 - \sigma_d)\} > 0. \tag{5}$$

2.3 Analysis transfer to the finite-prompt case

In this section, we study how properties of the gradient flow (GF_∞) with infinite prompts can be transferred to the finite-prompt flow:

$$\begin{cases} \dot{\mu}_L(t) &= -\nabla_\mu \mathcal{R}_{\text{soft},L}(\mu_L(t)), \\ \mu_L(0) &= \mu_0. \end{cases} \tag{GF_L}$$

We begin by formalizing the relationship between the finite-prompt and infinite-prompt risks. Intuitively, as the prompt length L increases, the finite-prompt operator $T_L^{\text{soft},\mu}$ should approximate its population counterpart $T_\infty^{\text{soft},\mu}$, leading to convergence of the corresponding risk functions. The following result makes this precise by establishing uniform convergence of the risks and their derivatives up to second order on compact sets.

For $k \in \mathbb{N}$, $\nabla^k f$ denotes the k-th order derivative of f (in particular $\nabla^0 f = f$), when the codomain is not $\mathbb{R}$, we also write $D^k f$. For $x \in \mathbb{R}^d, \rho > 0$ $B(x,\rho) := \{x' \in \mathbb{R}^d : \|x' - x\| \leq \rho\}$ is the closed ball centered at x of radius ρ. We denote by $\|\cdot\|_F$ the Frobenius norm.

Proposition 4. *We have that:*

1. *For $k \in \{0,1,2\}$,*

$$\sup_{\mu \in B(0,\rho)} \mathbb{E}\Big[\|D_\mu^k T_L^{\text{soft},\mu}(\mathbb{X})_1 - D_\mu^k T_\infty^{\text{soft},\mu}(X_1)\|_F^2\Big] = \mathcal{O}(\psi_k(L)),$$

   where $\psi_k(L) = L^{-\epsilon_k}(1 + \ln L)^{1-\epsilon_k}$, and $\epsilon_k = \frac{1}{16(k+3)^2\lambda^2(\mu^\top \Sigma \mu)^2 + 1} \in (0,1)$.

2. *For $k \in \{0,1,2\}$, there exists a constant $C_k = C_k(\rho, \Sigma, \lambda)$, such that*

$$\sup_{\mu \in B(0,\rho)} \mathbb{E}\big[\|D_\mu^k T_\infty^{\text{soft},\mu}(X_1)\|_F^2\big] \leq C_k.$$

3. *Then for $k \in \{0,1,2\}$ there exists a constant $C = C(\rho, \Sigma, C_0, C_1, C_2)$ independent of L such that*

$$\sup_{\mu \in B(0,\rho)} \|\nabla^k \mathcal{R}_{\text{soft},L}(\mu) - \nabla^k \mathcal{R}_{\text{soft},\infty}(\mu)\|_F^2 \leq C \sum_{j=0}^{k} \psi_j(L).$$

   In particular,

$$\nabla^k \mathcal{R}_{\text{soft},L} \xrightarrow[L\to\infty]{} \nabla^k \mathcal{R}_{\text{soft},\infty} \quad \textit{uniformly on } B(0,\rho).$$

Assertion 1 is closely related to the concentration results of Boursier and Boyer (2025), which establish bounds for the output and its gradient and suggest similar behavior for the Hessian. However, our setting differs in two key ways. First, derivatives are taken with respect to μ rather than the product $K^\top Q$ of key and query matrices, which necessitates a specific analysis. Additionally, unlike prior work that treats the query token independently, we explicitly handle the autocorrelation induced by its inclusion in the prompt (see Appendix C.3 for more details).

Leveraging Proposition 4, we now turn to the finite-prompt setting. We begin by establishing a basic stability property of the infinite-prompt dynamics, namely that its trajectories remain bounded. We then show that this boundedness transfers to the finite-prompt dynamics for sufficiently large prompt lengths, ensuring that both flows remain confined to a common compact region over an infinite time horizon.

Lemma 1. *Since $\mathcal{R}_{\text{soft},\infty}$ is coercive, the solution trajectory μ_∞ of (GF_∞) is bounded for all $t \geq 0$. That is, that there exists $\rho \geq \|\mu_0\|$ such that*

$$\{\mu_\infty(t) : t \geq 0\} \subset \{\mu \in \mathbb{R}^d : \mathcal{R}_{\text{soft},\infty}(\mu) \leq \mathcal{R}_{\text{soft},\infty}(\mu_0)\} \subset B(0,\rho).$$

Proposition 5. *Let μ_L and μ_∞ be the solutions of (GF_L) and (GF_∞) respectively. Let $\rho > 0$ be such that $\mu_\infty(t) \in B(0,\rho)$ for all $t \geq 0$. Then, for every $\rho' > \rho$, there exists L' such that, for all $L \geq L'$, $\mu_L(t) \in B(0,\rho')$ for every $t \geq 0$. In particular, the trajectories μ_L are uniformly bounded in time for sufficiently large L.*

Having established uniform boundedness, we next analyze how closely the finite-prompt dynamics track their infinite-prompt counterpart. The following results show that, on any finite time horizon, the trajectories and their associated risk values converge uniformly as the prompt length increases.

Proposition 6. *Let μ_L, μ_∞ be the solutions of (GF_L) and (GF_∞) respectively. Then for every $T > 0$, μ_L converges uniformly to μ_∞ on $[0,T]$ as $L \to \infty$.*

As a consequence, we can also control the convergence of the corresponding risk values along the trajectories.

Proposition 7. *Let μ_L, μ_∞ be the solutions of (GF_L) and (GF_∞) respectively. Then $\mathcal{R}_{\text{soft},L}(\mu_L)$ converges uniformly to $\mathcal{R}_{\text{soft},\infty}(\mu_\infty)$ on $[0,T]$ as $L \to \infty$.*

Propositions 6 and 7 establish uniform convergence over any finite time horizon as $L \to \infty$. Our primary objective, however, is to characterize the long-time behavior of the finite-prompt trajectory. To this end, we can compare it by concentration arguments with the infinite-prompt dynamics, which, by Proposition 2, converges to a global minimizer aligned with the leading eigenvector.

Beyond this comparison, our setting also allows for a sharper result: a direct characterization of the landscape of the finite-prompt risk. This is particularly noteworthy, as identifying critical points of objectives involving a softmax is typically challenging (Dohmatob, 2025; Marion and Berthier, 2023; Duranthon et al., 2025; Boursier and Boyer, 2025; Maulen-Soto et al., 2025). The result is formalized in the following proposition.

Proposition 8. *Assume that the p.s.d. covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ has simple spectrum composed of the eigenvalues $\sigma_1 > \ldots > \sigma_d > 0$ with $(u_j)_{j=1}^d$ the associated unit eigenvectors. Consider a Gaussian finite prompt $X_1, \ldots, X_L \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\Sigma)$.*

Set $\rho > 0$ to be large enough so that $B(0,\rho)$ contains all critical points of $\mathcal{R}_{\text{soft},\infty}$ (see Proposition 1). Then there exists $L_0 \in \mathbb{N}$ such that for all $L \geq L_0$, the set of critical points of $\mathcal{R}_{\text{soft},L}$ contained in $B(0,\rho)$ is finite, nondegenerate, and

$$\text{crit}(\mathcal{R}_{\text{soft},L}) \cap B(0,\rho) = \{\mu^\star_{L,0}\} \cup \{\pm\mu^\star_{L,\sigma_j} : j = 1,\ldots,d\},$$

where

1. *The point $\mu^\star_{L,0}$ is a strict local maximum such that $\mu^\star_{L,0} \xrightarrow[L\to\infty]{} 0$;*
2. *The points $\pm\mu^\star_{L,\sigma_j}$, for $j = 2,\ldots,d$, are strict saddles such that $\mu^\star_{L,\sigma_j} \xrightarrow[L\to\infty]{} \frac{1}{\sqrt{\lambda\sigma_j}} u_j$;*
3. *The points $\pm\mu^\star_{L,\sigma_1}$ are strict local minima such that $\mu^\star_{L,\sigma_1} \xrightarrow[L\to\infty]{} \frac{1}{\sqrt{\lambda\sigma_1}} u_1$.*

We remark that it suffices to characterize the critical points of $\mathcal{R}_{\text{soft},L}$ within a sufficiently large bounded set. Indeed, we will study the dynamics induced by the gradient flow associated with $\mathcal{R}_{\text{soft},L}$, which corresponding trajectories remain bounded for L large enough (Proposition 5). As a consequence, only critical points contained in this bounded region are relevant for the analysis.

Proposition 9 (Local convergence rate on finite-prompt setting)**.** *Let L large enough and $\mu_L(t)$ be a solution of (GF_L) with a generic initialization μ_0. Then,*

$$\mu_L(t) \xrightarrow[t\to\infty]{} \mu^\star_L \in \{\pm\mu^\star_{L,\sigma_1}\}.$$

Moreover, there exist $t_0 \geq 0$ *and* $s > 0$ *such that for all* $t \geq t_0$,

$$\mathcal{R}_{\mathrm{soft},L}(\mu_L(t)) - \mathcal{R}_{\mathrm{soft},L}(\mu_L^\star) \leq (\mathcal{R}_{\mathrm{soft},L}(\mu_L(t_0)) - \mathcal{R}_{\mathrm{soft},L}(\mu_L^\star))\, e^{-s(t-t_0)}.$$

Besides, for every $\varepsilon > 0$ *small enough, there exists* $t_0 > 0$ *such that*

$$\|\mu_L(t) - \mu_L^\star\| = \mathcal{O}\big(e^{-(\tilde{s}_L - \varepsilon)(t-t_0)}\big),$$

with $\tilde{s}_L = \sigma_{\min}\big(\nabla^2 \mathcal{R}_{\mathrm{soft},L}(\mu_L^\star)\big) > 0$, *and* $\tilde{s}_L \xrightarrow[L\to\infty]{} \tilde{s}$, *where* $\tilde{s}$ *is defined in* (5).

The finite-prompt dynamics converges to the first principal component, sharing the same local convergence behavior as the infinite-prompt limit, with an exponential rate governed by the local Hessian at the minimizer. Moreover, this rate converges to that of the infinite-prompt regime as $L \to \infty$. Numerical results in Figure 1 illustrate the alignment toward the leading eigenvector for both gradient descent on $\mathcal{R}_{\mathrm{soft},\infty}$ (assuming direct access to Σ) and stochastic gradient descent on $\mathcal{R}_{\mathrm{soft},L}$, the latter exhibiting slightly slower and noisier convergence.

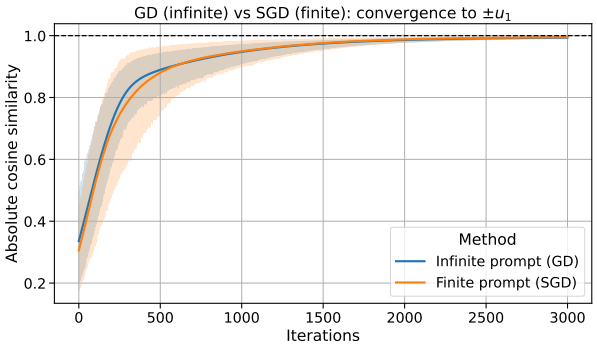


Figure 1: Alignment toward the principal eigenvector over iterations: SGD on $\mathcal{R}_{\mathrm{soft},L}$ ($L = 100$) vs GD on $\mathcal{R}_{\mathrm{soft},\infty}$

2.4 Connection of attention to Oja's flow

Interestingly, although our analysis is not motivated by classical results in online PCA, the infinite-prompt risk $\mathcal{R}_{\mathrm{soft},\infty}$ turns out to be closely related to the continuous-time limit of Oja's rule (Oja, 1982).

This connection emerges a posteriori from the structure of the gradient flow (GF_∞), which can be written explicitly as

$$\dot{\mu}_\infty = 4\lambda\Sigma^2\mu_\infty - 2\lambda^2\Big[(\mu_\infty^\top \Sigma^2 \mu_\infty)\,\Sigma\mu_\infty + (\mu_\infty^\top \Sigma \mu_\infty)\,\Sigma^2\mu_\infty\Big].$$

To make this link more transparent, we introduce the change of variables $w = \Sigma^{1/2}\mu_\infty$. In these coordinates, the dynamics take the form

$$\dot{w} = \Sigma\big[A(w)\,\Sigma w - B\,(w^\top \Sigma w)\,w\big], \quad A(w) := 2\lambda(2 - \lambda\, w^\top w),\ B := 2\lambda^2. \tag{6}$$

This dynamics can be then viewed as a variant of Oja's flow,

$$\dot{w} = \Sigma w - (w^\top \Sigma w)\, w, \tag{7}$$

in which the vector field is premultiplied by the invertible matrix Σ and modulated by the coefficients $A(w)$ and B. In particular, both flows share the same stationary points, namely the eigenvectors of Σ, and they will have the same nature (local minima, saddles, or maxima). Equation (7) arises in PCA as a continuous-time model for extracting the principal eigenvector of Σ. Its discrete counterpart, called Oja's rule, is a stochastic online approximation of the ODE (7). Overall, this shows that, in the infinite-prompt limit, the dynamics of the attention mechanism reduces, up to a linear change of variables, to a generalized version of Oja's flow, making explicit a connection between the attention mechanism and PCA that, to our knowledge, has not been discussed in the literature.

3 Distributional properties of the attention-based encoding

The measure-based perspective allows us to interpret an attention layer with an infinite-length prompt as an operator acting on the input distribution, thereby enabling a characterization of the resulting output distribution in the Gaussian setting. We begin by characterizing the output distribution when the layer has been trained, that is, when the attention parameter μ has converged.

Proposition 10 (Distribution of infinite-prompt attention operator)**.** *Let $\Sigma \in \mathbb{R}^{d\times d}$ be p.s.d. with (σ_1, u_1) its principal eigenpair. For $\mu \in \mathbb{R}^d$, define $\Gamma(\mu) := \lambda^2(\mu^\top \Sigma\mu)\,(\Sigma\mu)(\Sigma\mu)^\top$. When $X_1 \sim \mathcal{N}(0,\Sigma)$, then*

$$T_\infty^{\mathrm{soft},\mu}(X_1) = \lambda\Sigma\mu\mu^\top X_1 \sim \mathcal{N}(0,\Gamma(\mu)).$$

For $\lambda > 0$ and $\mu \neq 0$, the matrix $\Gamma(\mu)$ is p.s.d. of rank one. Moreover, for $\mu^\star = \frac{1}{\sqrt{\lambda\sigma_1}}\,u_1$, which is a global minimizer of $\mathcal{R}_{\mathrm{soft},\infty}$ (up to a sign) attained by gradient flow for generic initializations, one has

$$\Gamma(\mu^\star) = \sigma_1\, u_1 u_1^\top \qquad \text{and} \qquad T_\infty^{\mathrm{soft},\mu^\star}(X_1) \sim \mathcal{N}(0, \sigma_1 u_1 u_1^\top),$$

that is, the limiting distribution coincides with the law of the projection of $X \sim \mathcal{N}(0,\Sigma)$ onto the principal eigenspace, i.e., $\langle X, u_1\rangle u_1$.

This result highlights that, in the infinite-prompt limit, the attention mechanism effectively performs a rank-one projection of the input Gaussian distribution, and at optimality, recovers the principal eigendirection of Σ, thus aligning with the objective of PCA. This naturally raises the question of how this behavior extends beyond the idealized infinite-prompt regime, which we now investigate by analyzing the output distribution for finite-length prompts.

Corollary 1 (Wasserstein distance)**.** *Consider Gaussian input tokens $X_1, \dots, X_L \overset{i.i.d.}{\sim} \mathcal{N}(0,\Sigma)$. Then, the squared 2-Wasserstein distance between the distributions $\mathcal{L}(T_L^{\mathrm{soft},\mu}(\mathbb{X})_1)$ and $\mathcal{L}(T_\infty^{\mathrm{soft},\mu}(X_1))$ of the finite- and infinite-prompt architectures, both parameterized by μ, satisfies*

$$W_2^2\big(\mathcal{L}(T_L^{\mathrm{soft},\mu}(\mathbb{X})_1),\, \mathcal{L}(T_\infty^{\mathrm{soft},\mu}(X_1))\big) = \mathcal{O}\big(L^{-\epsilon}(1+\ln L)^{1-\epsilon}\big)\,,$$

where $\epsilon = \frac{1}{144\lambda^2(\mu^\top\Sigma\mu)^2+1} \in (0,1)$.

In particular, when $\mu^\star = \frac{u_1}{\sqrt{\lambda\sigma_1}}$ with (σ_1, u_1) the principal eigenpair of Σ, then

$$W_2^2\big(\mathcal{L}(T_L^{\mathrm{soft},\mu^\star}(\mathbb{X})_1),\, \mathcal{N}(0,\sigma_1 u_1 u_1^\top)\big) = \mathcal{O}\Big(L^{-\frac{1}{145}}(1+\ln L)^{\frac{144}{145}}\Big)\,.$$

Proposition 2 shows that minimizing the infinite-prompt risk $\mathcal{R}_{\mathrm{soft},\infty}$ recovers the principal components. Combined with the control of the deviation between the law of $T_L^{\mathrm{soft},\mu}(\mathbb{X})_1$ and its Gaussian limit $\mathcal{N}(0,\Gamma(\mu))$, as well as Proposition 9, this establishes that finite-prompt attention effectively performs an approximate PCA when training is performed through the minimization of $\mathcal{R}_{\mathrm{soft},L}$.

We also note that the convergence rate observed in practice is significantly faster than the theoretical bounds (see, e.g., Figure 5), indicating that these bounds may be conservative.

4 Toward spiked covariance models

In this section, we move beyond the fixed design considered so far and consider a setting in which the data remain Gaussian, but the (random) covariance structure depends on the prompt. This perspective places our analysis to some extent within the framework of in-context learning. We now make explicit the dependence on the covariance structure in the risks

$$\begin{cases} \mathcal{R}_{\mathrm{soft},L}^{(\Sigma)}(\mu) = \mathbb{E}_{X_1,\dots,X_L\sim\mathcal{N}(0,\Sigma)}\left[\|X_1 - T_L^{\mathrm{soft},\mu}(\mathbb{X})_1]\|^2\right] \\ \mathcal{R}_{\mathrm{soft},\infty}^{(\Sigma)}(\mu) = \mathbb{E}_{X\sim\mathcal{N}(0,\Sigma)}\left[\|X - T_\infty^{\mathrm{soft},\mu}(X)\|^2\right] \end{cases}$$

where we recall that the latter can be expressed in terms of the covariance matrix Σ:

$$\mathcal{R}_{\mathrm{soft},\infty}^{(\Sigma)}(\mu) = \mathrm{tr}(\Sigma) - 2\lambda\,\mu^\top\Sigma^2\mu + \lambda^2\,(\mu^\top\Sigma\mu)(\mu^\top\Sigma^2\mu).$$

This formulation naturally leads to "in-context learning" risks, obtained by averaging over the distribution $\mathcal{D}$ of covariance matrices

$$\mathcal{R}_L^{\mathrm{ICL}}(\mu) = \mathbb{E}_{\Sigma\sim\mathcal{D}}\big[\mathcal{R}_{\mathrm{soft},L}^{(\Sigma)}(\mu)\big] \quad \text{and} \quad \mathcal{R}_\infty^{\mathrm{ICL}}(\mu) = \mathbb{E}_{\Sigma\sim\mathcal{D}}\big[\mathcal{R}_{\mathrm{soft},\infty}^{(\Sigma)}(\mu)\big]. \tag{8}$$

Choice of $\mathcal{D}$. We consider covariance matrices drawn from a Wishart distribution $W_d(V, n)$ with scale matrix V and n degrees of freedom. This choice yields random covariance matrices that almost surely have a simple spectrum (i.e., distinct eigenvalues), while their expectation aligns with a spiked covariance model of the form

$$V = \xi^2 I_d + \theta\, vv^\top, \quad \text{for some } v \text{ such that } \|v\| = 1.$$

This mild random setting interpolates between the isotropic case ($\theta = 0$) and a structured anisotropic regime, providing a natural testbed to assess whether rank-one softmax attention layers can recover the latent direction v. Training is now modeled via the gradient flow of (8), corresponding to an idealized procedure based on the population risk rather than its empirical counterpart, with Gaussian prompts and prompt-dependent Wishart-distributed covariance matrices.

Infinite-prompt analysis. The function $\mathcal{R}_\infty^{\mathrm{ICL}}(\mu)$ can be rewritten depending only (see Lemma 13) on the norm of the attention parameter $r = \|\mu\|$ and on the angle $\alpha = \langle \mu, v\rangle$, between the attention parameter and the covariance latent direction, so that

$$\mathcal{R}_\infty^{\mathrm{ICL}}(\mu) = \tilde{\mathcal{R}}^{\mathrm{ICL}}(r^2, \alpha^2).$$

This reformulation enables a precise characterization of the critical points of the objective (see Propositions 18 and 19), and in particular shows that the only local (and thus global) minima are given by

$$\mu^\star = \pm \alpha^\star v, \tag{9}$$

where the scalar $\alpha^\star$ is defined in (22), and determines the optimal magnitude of the parameter along the signal direction v. In addition, $\mu = 0$ is a strict local maximum, and there are $2(d-1)$ saddle points corresponding to directions orthogonal to v.

Proposition 11 (Local convergence rate on ICL infinite-prompt setting)**.** *Consider the gradient flow*

$$\dot{\mu}(t) = -\nabla \mathcal{R}_\infty^{\mathrm{ICL}}(\mu(t)).$$

Then, for a generic initialization μ_0, the gradient flow satisfies $\mu(t) \xrightarrow[t\to\infty]{} \mu^\star$, where $\mu^\star$ is defined in (9). Besides for every $\varepsilon > 0$ small enough, there exists $t_0 > 0$ such that

$$\|\mu(t) - \mu^\star\| = \mathcal{O}(e^{-(\hat{s}-\varepsilon)(t-t_0)}),$$

with $\hat{s} = \sigma_{\min}\big(\nabla^2 \mathcal{R}_\infty^{\mathrm{ICL}}(\mu^\star)\big)$. When ξ^2 is small enough, we have $\hat{s} \sim 2\frac{\lambda n(n^2+5n+2)\theta\,\xi^2}{n+4}$, so the convergence rate scales proportionally to ξ^2 and θ, and quadratically with n.

We observe that a moderate level of isotropic noise ξ^2 can facilitate convergence. Moreover, the convergence rate increases linearly with the signal strength θ and quadratically with the degrees of freedom n.

Finite-prompt analysis. The transfer of the infinite-prompt analysis can be done to the finite-prompt one as we establish the uniform convergence on bounded sets of $\nabla^k \mathcal{R}_L^{\mathrm{ICL}}(\mu)$ to $\nabla^k \mathcal{R}_\infty^{\mathrm{ICL}}(\mu)$ for $k \in \{0, 1, 2\}$ as $L \to \infty$.

Proposition 12. *For $\rho > 0$, $k \in \{0, 1, 2\}$,*

$$\lim_{L\to\infty} \sup_{\mu \in B(0,\rho)} \|\nabla^k \mathcal{R}_L^{\mathrm{ICL}}(\mu) - \nabla^k \mathcal{R}_\infty^{\mathrm{ICL}}(\mu)\|_F^2 = 0.$$

Following the argument of Proposition 8, one proves that for L large enough, the critical points of $\mathcal{R}_L^{\mathrm{ICL}}$ in a compact set correspond to perturbations of those of $\mathcal{R}_\infty^{\mathrm{ICL}}$. In particular, in this compact set there is one local maximum near 0, two local minima near $\pm\alpha^\star v$, denoted $\pm\mu_{L,\|}^\star$, and possible saddle points located near the $2(d-1)$ orthogonal critical points of $\mathcal{R}_\infty^{\mathrm{ICL}}$ (see Proposition 20 for more details). As a consequence, by proceeding as before, we obtain the following convergence and rate result for the ICL objective.

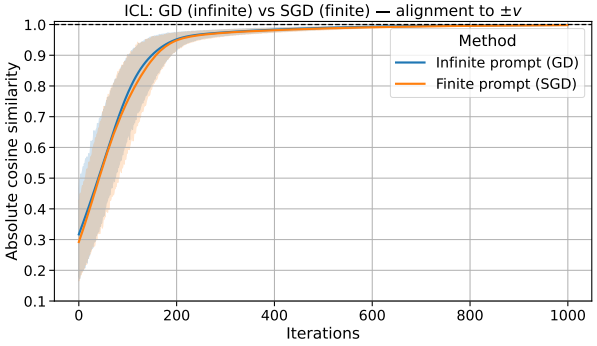


Figure 2: Alignment toward the signal direction v over iterations: SGD on $\mathcal{R}_L^{\mathrm{ICL}}$ ($L = 100$) vs GD on $\mathcal{R}_\infty^{\mathrm{ICL}}$

Proposition 13 (ICL finite-prompt convergence and rate)**.** *Let $\rho > 0$ be sufficiently large and $L \geq L_0$ as in Proposition 20. Consider the gradient flow*

$$\dot{\mu}(t) = -\nabla \mathcal{R}_L^{\mathrm{ICL}}(\mu(t)).$$

Then, for almost every initialization $\mu_0 \in B(0, \rho)$,

$$\mu(t) \xrightarrow[t\to\infty]{} \mu_L^\star \in \{\pm \mu_{L,\parallel}^\star\},$$

where $\mu_{L,\parallel}^\star \to \alpha^\star v$ as $L \to \infty$, with $\alpha^\star$ defined in (22). Moreover, there exist $t_0 \geq 0$ and $\hat{s}_L > 0$ such that for all $t \geq t_0$,

$$\mathcal{R}_L^{\mathrm{ICL}}(\mu(t)) - \mathcal{R}_L^{\mathrm{ICL}}(\mu_L^\star) \leq \big(\mathcal{R}_L^{\mathrm{ICL}}(\mu(t_0)) - \mathcal{R}_L^{\mathrm{ICL}}(\mu_L^\star)\big) e^{-\hat{s}_L (t - t_0)}.$$

Besides, for every $\varepsilon > 0$ small enough, there exists $t_0 > 0$ such that

$$\|\mu(t) - \mu_L^\star\| = \mathcal{O}\big(e^{-(\hat{s}_L - \varepsilon)(t - t_0)}\big),$$

with $\hat{s}_L = \sigma_{\min}\big(\nabla^2 \mathcal{R}_L^{\mathrm{ICL}}(\mu_L^\star)\big)$, and $\hat{s}_L \xrightarrow[L\to\infty]{} \hat{s}$, where $\hat{s}$ is defined in (16).

Consequently, minimizing the finite-prompt ICL risk recovers an estimator that asymptotically aligns with the spike direction v. This provides an attention-based analogue of spiked PCA in which the latent direction is recovered through training dynamics rather than spectral decomposition. Numerical results in Figure 2 illustrate the alignment toward the signal direction for both gradient descent on $\mathcal{R}_\infty^{\mathrm{ICL}}$ (assuming direct access to $V = \xi^2 I_d + \theta v v^\top$) and stochastic gradient descent on $\mathcal{R}_L^{\mathrm{ICL}}$.

Code availability

Our code is available at
https://github.com/rodrigomaulen/Attention_based_PCA.

References

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=LziniAXEI9.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations*, 2015.

Dzmitry Bahdanau, Jan Chorowski, and Dmitriy Serdyuk. Neural machine translation by jointly learning to align and translate. *3rd International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4945–4949, 2016.

Etienne Boursier and Claire Boyer. Softmax as linear attention in the large-prompt regime: a measure-based perspective. *arXiv preprint: 2512.11784*, 2025.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: Composable transformations of Python+NumPy programs. https://github.com/google/jax, 2018.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Lee Peter, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv:2303.12712*, 2023.

Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers, 2025. URL https://arxiv.org/abs/2501.18322.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

Elvis Dohmatob. Understanding softmax attention layers:\\ exact mean-field analysis on a toy problem. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=GiqeRe1NsY.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, and Dirk Weissenborn. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2020.

Odilon Duranthon, Pierre Marion, Claire Boyer, Bruno Loureiro, and Lenka Zdeborová. Statistical advantage of softmax attention: Insights from single-location regression. *arXiv preprint arXiv:2509.21936*, 2025.

Takashi Furuya, Maarten V. de Hoop, and Gabriel Peyré. Transformers are universal in-context learners, 2024. URL https://arxiv.org/abs/2408.01367.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes, 2023. URL https://arxiv.org/abs/2208.01066.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, USA, 3rd edition, 1996. ISBN 0-8018-5414-8.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 498–520, 1933.

L. Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution. *Biometrika*, 12(1):134–139, 1918.

Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2 edition, 2002.

Steven G. Krantz and Harold R. Parks. *The implicit function theorem*. Birkhauser, 2013.

Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *COLT*, 2016.

Yingcong Li, M. Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning, 2023. URL https://arxiv.org/abs/2301.07067.

Zihao Li, Yuan Cao, Cheng Gao, Yihan He, Han Liu, Klusowkski Jason, Jianqing Fan, and Mengdi Wang. One-layer transformer provably learns one-nearest neighbor in context. *Advances in Neural Information Processing Systems*, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015.

Pierre Marion and Raphael Berthier. Leveraging the two timescale regime to demonstrate convergence of neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Rodrigo Maulen-Soto, Pierre Marion, and Claire Boyer. Attention-based clustering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=NRvxzOdSPU.

Erkki Oja. A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982. doi: 10.1007/BF00275687.

Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 2:1–2:12. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017. doi: 10.4230/LIPIcs.ITCS.2017.2.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2 (11):559–572, 1901.

Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada*, 2019.

Michael Shub. *Global Stability of Dynamical Systems*. Springer, New York, 1987.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30: 6000–6010, 2017.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174, 2023. URL https://arxiv.org/abs/2212.07677.

Hongru Yang, Zhangyang Wang, Jason D. Lee, and Yingbin Liang. Transformers provably learn two-mixture of linear classification via gradient flow. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=AuAj4vRPkv.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

Appendix

This appendix gathers additional results and technical details supporting the main text. Section A presents a linear attention model that similarly recovers the principal component. Sections B and C gather the proofs of the main results as well as the technical lemmas required throughout the paper. Finally, in Section D we present numerical experiments that illustrate and support our theoretical findings.

A Linear attention layer

In line with (Maulen-Soto et al., 2025), for $\mu \in \mathbb{R}^d$ and $X_1, \dots, X_L$ i.i.d. according to $\mathcal{N}(0, \Sigma)$, we introduce the linear attention operator given by

$$T_L^{\mathrm{lin},\mu}(\mathbb{X})_\ell = \frac{\lambda}{L} \sum_{k=1}^{L} (X_\ell^\top \mu \mu^\top X_k) X_k. \tag{10}$$

We first note that by the strong law of large numbers, for $X \sim \mathcal{N}(0, \Sigma)$,

$$\frac{1}{L} \sum_{k=1}^{L} \mu^\top X_k X_k \xrightarrow[L\to\infty]{\text{a.s.}} \mathbb{E}[\mu^\top X X] = \Sigma \mu.$$

Then,

$$T_L^{\mathrm{lin},\mu}(\mathbb{X})_1 = \frac{\lambda}{L} \sum_{k=1}^{L} (X_1^\top \mu \, \mu^\top X_k) \, X_k \xrightarrow[L\to\infty]{\text{a.s.}} \lambda \, \Sigma \mu \mu^\top X_1.$$

This shows that $T_L^{\mathrm{lin},\mu}(\mathbb{X})_1$ and $T_L^{\mathrm{soft},\mu}(\mathbb{X})_1$ share the same almost sure limit when $L \to \infty$, i.e.,

$$\lim_{L\to\infty} T_L^{\mathrm{lin},\mu}(\mathbb{X})_1 = \lim_{L\to\infty} T_L^{\mathrm{soft},\mu}(\mathbb{X})_1 = T_\infty^{\mathrm{soft},\mu}(X_1) = \lambda \Sigma \mu \mu^\top X_1, \quad a.s..$$

A.1 Risk

We define the associated population risk as

$$\mathcal{R}_{\mathrm{lin},L}(\mu) = \mathbb{E}\left[\|X_1 - T_L^{\mathrm{lin},\mu}(\mathbb{X})_1\|^2\right]. \tag{11}$$

We start by giving an expression of the risk as a function of the covariance matrix Σ of the input sequence.

Proposition 14. *Let $a(\mu) = \mu^T\Sigma\mu$ and $b(\mu) = \mu^T\Sigma^2\mu$. We have that*

$$\mathcal{R}_{\mathrm{lin},L}(\mu) = tr(\Sigma) - \frac{2\lambda}{L}tr(\Sigma)a - \frac{2\lambda(L+1)}{L}b + \lambda^2(L+2)tr(\Sigma)a^2 + \frac{\lambda^2(L+2)(L+3)}{L^2}ab. \tag{12}$$

Proof. The computation starts as follows

$$\begin{aligned}
\mathcal{R}_{\mathrm{lin},L}(\mu) &= \mathbb{E}\left[\|X_1 - \frac{\lambda}{L}\sum_{k=1}^{L} X_1^\top\mu\mu^\top X_k X_k\|^2\right] \\
&= \mathbb{E}\left[\|X_1 - \frac{\lambda}{L}\sum_{k=2}^{L} X_1^\top\mu\mu^\top X_k X_k - \frac{\lambda}{L}(X_1^\top\mu)^2 X_1\|^2\right] \\
&= \mathbb{E}\left[\|X_1 - \frac{\lambda}{L}\sum_{k=2}^{L} X_1^\top\mu\mu^\top X_k X_k\|^2\right] + \frac{\lambda^2}{L^2}\mathbb{E}\left[(X_1^\top\mu)^4\|X_1\|^2\right] - 2\frac{\lambda}{L}\mathbb{E}\left[(X_1^\top\mu)^2\|X_1\|^2\right] \\
&\quad + 2\frac{\lambda^2}{L^2}\sum_{k=2}^{L}\mathbb{E}\left[(X_1^\top\mu)^3\mu^\top X_k X_1^\top X_k\right]
\end{aligned}$$

And using Proposition 17, we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\|X_1 - \frac{\lambda}{L}\sum_{k=2}^{L} X_1^\top\mu\mu^\top X_k X_k\|^2\right] \\
&= \mathbb{E}\left[\|X_1\|^2\right] - 2\frac{\lambda}{L}\sum_{k=2}^{L}\mathbb{E}\left[X_1^\top\mu\mu^\top X_k X_1^\top X_k\right] + \frac{\lambda^2}{L^2}\sum_{k=2}^{L}\mathbb{E}[(X_1^\top\mu)^2]\mathbb{E}[(X_k^\top\mu)^2\|X_k\|^2] \\
&\quad + 2\frac{\lambda^2}{L^2}\sum_{2\le k<j\le L}\mathbb{E}[(X_1^\top\mu)^2]\mathbb{E}[X_k^\top\mu X_j^\top\mu X_k^\top X_j] \\
&= \mathrm{tr}[\Sigma] - 2\lambda(L-1)b + \frac{\lambda^2}{L^2}(L-1)[\mathrm{tr}(\Sigma)a^2 + 2ab] + \frac{\lambda^2}{L^2}(L-1)(L-2)ab,
\end{aligned}$$

where we used

$$\mathbb{E}\left[(X_1^\top\mu)^4\|X_1\|^2\right] = 3\mathrm{tr}(\Sigma)a^2 + 12ab,$$

$$\mathbb{E}\left[(X_1^\top\mu)^2\|X_1\|^2\right] = \mathrm{tr}(\Sigma)a + 2b,$$

$$\mathbb{E}\left[(X_1^\top\mu)^3\mu^\top X_2 X_1^\top X_2\right] = 3ab.$$

Finally,

$$\begin{aligned}
\mathcal{R}_{\mathrm{lin},L}(\mu) &= \mathrm{tr}[\Sigma] - 2\frac{\lambda}{L}(L-1)b + \frac{\lambda^2}{L^2}(L-1)[\mathrm{tr}(\Sigma)a^2 + 2ab] + \frac{\lambda^2}{L^2}(L-1)(L-2)ab \\
&\quad + \frac{\lambda^2}{L^2}(3\mathrm{tr}(\Sigma)a^2 + 12ab) - 2\frac{\lambda}{L}(\mathrm{tr}(\Sigma)a + 2b) + 6\frac{\lambda^2}{L^2}(L-1)ab \\
&= \mathrm{tr}(\Sigma) - 2\frac{\lambda}{L}\mathrm{tr}(\Sigma)a - 2\frac{\lambda}{L}(L+1)b + \frac{\lambda^2}{L^2}(L+2)\mathrm{tr}(\Sigma)a^2 + \frac{\lambda^2}{L^2}(L+2)(L+3)ab.
\end{aligned}$$

□

A.2 Optimization landscape analysis

We first express the gradient of $\nabla\mathcal{R}_{\text{lin},L}$ with respect to μ, (using $\nabla a = 2\Sigma\mu$ and $\nabla b = 2\Sigma^2\mu$) as

$$\begin{aligned}\nabla\mathcal{R}_{\text{lin},L}(\mu) &= -4\frac{\lambda}{L}\text{tr}(\Sigma)\Sigma\mu - 4\frac{\lambda}{L}(L+1)\Sigma^2\mu + 4\frac{\lambda^2}{L^2}(L+2)\text{tr}(\Sigma)a\Sigma\mu \\ &\quad + 2\frac{\lambda^2}{L^2}(L+2)(L+3)\Big(b\Sigma\mu + a\Sigma^2\mu\Big), \\ &= \alpha(\mu)\Sigma\mu + \beta(\mu)\Sigma^2\mu,\end{aligned}$$

where $\alpha(\mu) = -4\frac{\lambda}{L}\text{tr}(\Sigma) + 4\frac{\lambda^2}{L^2}(L+2)\text{tr}(\Sigma)a(\mu) + 2\frac{\lambda^2}{L^2}(L+2)(L+3)b(\mu)$, and $\beta(\mu) = -4\frac{\lambda}{L}(L+1) + 2\frac{\lambda^2}{L^2}(L+2)(L+3)a(\mu)$. Setting $\nabla\mathcal{R}_{\text{lin},L}(\mu) = 0$ leads to

$$\alpha(\mu)\Sigma\mu + \beta(\mu)\Sigma^2\mu = 0.$$

We first note that $\mu = 0$ is a critical point. Now consider the case $\mu \neq 0$, multiplying by Σ^{-1} (since Σ is invertible) gives

$$\alpha(\mu)\mu + \beta(\mu)\Sigma\mu = 0. \tag{13}$$

This equation implies that $\Sigma\mu$ is aligned with the direction of μ, i.e., $\Sigma\mu = -\frac{\alpha(\mu)}{\beta(\mu)}\mu$, whenever $\beta(\mu) \neq 0$. Assume by contradiction that $\beta(\mu) = 0$, then $\alpha(\mu)$ would be fixed accordingly, and the critical point condition would further impose $\alpha(\mu) = 0$, which would in turn fix $b(\mu)$ to a negative value, which is contradictory with its definition. Hence, for any non-zero critical point, one must have $\beta(\mu) \neq 0$. It follows that μ must be an eigenvector of Σ with eigenvalue $-\frac{\alpha(\mu)}{\beta(\mu)}$.

It follows that the critical points are given by $\mu = 0$ and points of the form $\gamma_i u_i$ for $i = 1, \ldots, d$ with u_i are unit eigenvectors of Σ with associated eigenvalue $\sigma_i > 0$, and $\gamma_i \neq 0$. Plugging $\mu = \gamma_i u_i$ into the equation (13) and simplifying we get the condition

$$-\text{tr}(\Sigma) - (L+1)\sigma_i + \frac{\lambda}{L}(L+2)\text{tr}(\Sigma)\gamma_i^2\sigma_i + \frac{\lambda}{L}(L+2)(L+3)\gamma_i^2\sigma_i^2 = 0.$$

Solving for γ_i, we obtain

$$\gamma_i = \pm\sqrt{\frac{\text{tr}(\Sigma) + (L+1)\sigma_i}{\frac{\lambda}{L}\sigma_i(L+2)(\text{tr}(\Sigma) + (L+3)\sigma_i)}}.$$

Proposition 15 (Characterization of critical points)**.** *Let $\Sigma \in \mathbb{R}^{d\times d}$ be symmetric positive definite with simple spectrum, and let $(\sigma_i, u_i)_{i=1}^d$ be its eigenpairs, where u_i are unit eigenvectors and $\sigma_1 > \ldots > \sigma_d$. For $i \in \{1, \ldots, d\}$, define*

$$\mu_i^{\pm} = \pm\gamma_i^\star u_i, \qquad \gamma_i^\star = \sqrt{\frac{tr(\Sigma) + (L+1)\sigma_i}{\frac{\lambda}{L}\sigma_i(L+2)(tr(\Sigma) + (L+3)\sigma_i)}}.$$

Then $\text{crit}(\mathcal{R}_{\text{lin},L}) = \{0\} \cup \{\mu_i^{\pm} : i = 1, \ldots, d\}$. We have that $\nabla^2\mathcal{R}_{\text{lin},L}(0)$ is negative definite, thus 0 is a local maxima. Moreover, the Hessian $\nabla^2\mathcal{R}_{\text{lin},L}(\mu_i^{\pm})$ is diagonal in the eigenbasis of Σ. The eigenvalue of $\nabla^2\mathcal{R}_{\text{lin},L}(\mu_i^{\pm})$ associated with the eigenvector u_j is given by

$$\begin{cases} 8\frac{\lambda}{L}\sigma_i(tr(\Sigma) + (L+1)\sigma_i), \text{ if } j = i, \\ 2\frac{\lambda}{L}\sigma_j(\sigma_i - \sigma_j)\frac{(L-1)tr(\Sigma) + (L+1)(L+3)\sigma_i}{tr(\Sigma) + (L+3)\sigma_i}, \text{ if } j \neq i. \end{cases}$$

In particular, if $i = 1$, the Hessian $\nabla^2\mathcal{R}_{\text{lin},L}(\mu_1^{\pm})$ is positive definite. If $i > 1$, the Hessian $\nabla^2\mathcal{R}_{\text{lin},L}(\mu_i)$ has both positive and negative eigenvalues. Consequently, $\mu_1^{\pm} = \pm\alpha_1^\star u_1$ are global minimizers of $\mathcal{R}_{\text{lin},L}$, whereas $\mu_i^{\pm} = \pm\gamma_i^\star u_i$ is a strict saddle point for every $i > 1$.

Proof. We compute the Hessian of $\mathcal{R}_{\text{lin},L}$ at a generic μ,

$$\begin{aligned}\nabla^2\mathcal{R}_{\text{lin},L}(\mu) = &\Big(-4\frac{\lambda}{L}\text{tr}(\Sigma) + 4\frac{\lambda^2}{L^2}(L+2)\text{tr}(\Sigma)a(\mu) + 2\frac{\lambda^2}{L^2}(L+2)(L+3)b(\mu)\Big)\Sigma\\ &+\Big(-4\frac{\lambda}{L}(L+1) + 2\frac{\lambda^2}{L^2}(L+2)(L+3)a(\mu)\Big)\Sigma^2 + 8\frac{\lambda^2}{L^2}(L+2)\text{tr}(\Sigma)\Sigma\mu(\Sigma\mu)^\top\\ &+4\frac{\lambda^2}{L^2}(L+2)(L+3)\Big(\Sigma\mu(\Sigma^2\mu)^\top + \Sigma^2\mu(\Sigma\mu)^\top\Big).\end{aligned}$$

Then

$$\nabla^2\mathcal{R}_{\text{lin},L}(0) = -4\frac{\lambda}{L}\Sigma(\text{tr}(\Sigma) + (L+1)\Sigma).$$

And evaluating at $\mu_i^\pm = \pm\gamma_i^\star u_i$ we obtain

$$\begin{aligned}\nabla^2\mathcal{R}_{\text{lin},L}(\mu_i^\pm) =&\Big(-4\frac{\lambda}{L}\text{tr}(\Sigma) + 4\frac{\lambda^2}{L^2}(L+2)\text{tr}(\Sigma)(\gamma_i^\star)^2\sigma_i + 2\frac{\lambda^2}{L^2}(L+2)(L+3)(\gamma_i^\star)^2\sigma_i^2\Big)\Sigma\\ &+\Big(-4\frac{\lambda}{L}(L+1) + 2\frac{\lambda^2}{L^2}(L+2)(L+3)(\gamma_i^\star)^2\sigma_i\Big)\Sigma^2\\ &+8\frac{\lambda^2}{L^2}(L+2)(\gamma_i^\star)^2\sigma_i^2\Big(\text{tr}(\Sigma) + (L+3)\sigma_i\Big)u_iu_i^\top.\end{aligned}$$

Consequently,

$$\begin{aligned}\nabla^2\mathcal{R}_{\text{lin},L}(\mu_i^\pm)u_i =&\Big(-4\frac{\lambda}{L}\text{tr}(\Sigma) + 4\frac{\lambda^2}{L^2}(L+2)\text{tr}(\Sigma)(\gamma_i^\star)^2\sigma_i + 2\frac{\lambda^2}{L^2}(L+2)(L+3)(\gamma_i^\star)^2\sigma_i^2\Big)\sigma_iu_i\\ &+\Big(-4\frac{\lambda}{L}(L+1) + 2\frac{\lambda^2}{L^2}(L+2)(L+3)(\gamma_i^\star)^2\sigma_i\Big)\sigma_i^2u_i\\ &+8\frac{\lambda^2}{L^2}(L+2)(\gamma_i^\star)^2\sigma_i^2\Big(\text{tr}(\Sigma) + (L+3)\sigma_i\Big)u_i.\\ =\;& 8\frac{\lambda}{L}\sigma_i(\text{tr}(\Sigma) + (L+1)\sigma_i)u_i,\end{aligned}$$

where the last equality comes from replacing $(\gamma_i^\star)^2$ into the expression. Besides, for $j \neq i$

$$\begin{aligned}\nabla^2\mathcal{R}_{\text{lin},L}(\mu_i^\pm)u_j =&\Big(-4\frac{\lambda}{L}\text{tr}(\Sigma) + 4\frac{\lambda^2}{L^2}(L+2)\text{tr}(\Sigma)(\gamma_i^\star)^2\sigma_i + 2\frac{\lambda^2}{L^2}(L+2)(L+3)(\gamma_i^\star)^2\sigma_i^2\Big)\sigma_ju_j\\ &+\Big(-4\frac{\lambda}{L}(L+1) + 2\frac{\lambda^2}{L^2}(L+2)(L+3)(\gamma_i^\star)^2\sigma_i\Big)\sigma_j^2u_j\\ =\;& 2\frac{\lambda}{L}\sigma_j(\sigma_i-\sigma_j)\frac{(L-1)\text{tr}(\Sigma) + (L+1)(L+3)\sigma_i}{\text{tr}(\Sigma) + (L+3)\sigma_i}u_j,\end{aligned}$$

□

Proposition 16. *The function $\mathcal{R}_{\text{lin},L}$ is coercive and locally Lipschitz. Hence, the gradient flow dynamic*

$$\dot{\mu}(t) = -\nabla\mathcal{R}_{\text{lin},L}(\mu(t)),$$

converges to a critical point of $\mathcal{R}_{\text{lin},L}$. Furthermore, for almost every initialization, the sequence converges to one of the two global minimizers, $\mu^\star = \pm\alpha_1^\star u_1$, where u_1 is the normalized principal eigenvector of Σ, and

$$\alpha_1^\star = \sqrt{\frac{tr(\Sigma) + (L+1)\sigma_1}{\frac{\lambda}{L}\sigma_1(L+2)(tr(\Sigma) + (L+3)\sigma_1)}}.$$

Normalizing the limit point $\mu^\star$ recovers the principal eigenvector u_1 up to a sign. Besides,

$$\sigma_1 = \quad -\frac{\alpha(\mu^\star)}{\beta(\mu^\star)} = -\frac{-2tr(\Sigma) + 2\frac{\lambda}{L}(L+2)tr(\Sigma)\mu^\star\Sigma\mu^\star + \frac{\lambda}{L}(L+2)(L+3)\mu^\star\Sigma^2\mu^\star}{-2(L+1) + \frac{\lambda}{L}(L+2)(L+3)\mu^\star\Sigma\mu^\star}.$$

Proof. Since $\mathcal{R}_{\text{lin},L}$ is a polynomial in μ, it is C^∞ and locally Lipschitz. Moreover, $\mathcal{R}_{\text{lin},L}$ is coercive, this implies that the sublevel set $\mathcal{S} = \{\mu : \mathcal{R}_{\text{lin},L}(\mu) \leq \mathcal{R}_{\text{lin},L}(\mu^0)\}$ is compact for any initialization μ^0. The gradient flow trajectory $\mathcal{R}_{\text{lin},L}(\mu(t))$ is non-increasing and $\|\nabla\mathcal{R}_{\text{lin},L}(\mu(t))\| \to 0$. Furthermore, since $\mathcal{R}_{\text{lin},L}$ is polynomial, it satisfies the Łojasiewicz inequality at every critical point, thus $\mu(t)$ converges to the set of critical points given by Proposition 15, namely $\text{crit}(\mathcal{R}_{\text{lin},L}) = \{0\} \cup \{\mu_i^\pm : i = 1, \dots, d\}$.

Using the classification from Proposition 15, the Stable Manifold Theorem (Shub, 1987, Theorem III.7) states that the set of initial conditions converging to an unstable fixed point of a C^2 diffeomorphism has Lebesgue measure zero. Consequently, for almost all initializations, the sequence cannot converge to 0 or any μ_i with $i > 1$. Since the sequence is bounded and must converge to a critical point, it converges to the stable minimizer $\mu_1^\pm = \pm\alpha_1^\star u_1$. □

Remark 2. As in Remark 1, once the leading eigenvector u_1 has been recovered, the second principal component u_2 can be obtained by constraining the dynamics to the orthogonal subspace $u_1^\perp$. To this end, we consider the projected gradient flow

$$\begin{cases} \dot{\mu}(t) = -\mathrm{P}_{u_1^\perp}\big(\nabla\mathcal{R}_{\text{lin},L}(\mu(t))\big), \\ \mu(0) = \mu_0, \end{cases} \tag{14}$$

where $\mathrm{P}_{u_1^\perp} = I_d - u_1 u_1^\top$ denotes the orthogonal projection onto $u_1^\perp$. This projection removes the component along u_1, allowing the dynamics to generically recover u_2.

B Proofs

In the following section, we present the proofs of the propositions and lemmas stated in the main text.

B.1 Proof of Lemma 1

Proof. Along the gradient flow (GF_∞), the map $t \mapsto \mathcal{R}_{\text{soft},\infty}(\mu_\infty(t))$ is non-increasing. Hence, for all $t \geq 0$,

$$\mathcal{R}_{\text{soft},\infty}(\mu_\infty(t)) \leq \mathcal{R}_{\text{soft},\infty}(\mu_0).$$

By coercivity of $\mathcal{R}_{\text{soft},\infty}$, the sublevel set $[\mathcal{R}_{\text{soft},\infty} \leq R_{\text{soft},\infty}(\mu_0)]$ is bounded. Therefore, there exists $\rho \geq \|\mu_0\|$ such that $\mu_\infty(t) \in B(0,\rho)$ for all $t \geq 0$, which proves the claim. □

B.2 Proof of Proposition 1

Proof. Compute gradients using $\nabla_\mu a = 2\Sigma\mu$ and $\nabla_\mu b = 2\Sigma^2\mu$:

$$\nabla_\mu\mathcal{R}_{\text{soft},\infty}(\mu) = -4\lambda\Sigma^2\mu + 2\lambda^2\big(b\Sigma\mu + a\Sigma^2\mu\big).$$

At a nonzero critical point, divide by 2 and rearrange to get

$$(\lambda a - 2)\Sigma^2\mu + \lambda b\Sigma\mu = 0.$$

Multiplying by Σ^{-1} yields

$$(\lambda a - 2)\Sigma\mu + \lambda b\mu = 0$$

Assuming $\lambda a - 2 \neq 0$, we can rearrange the previous equation into the following one

$$\Sigma\mu = c(\mu)\mu,$$

for $c(\mu) = -\dfrac{\lambda b}{\lambda a - 2}$. If $\lambda a - 2 = 0$, we would have

$$\lambda b\mu = 0 \iff \frac{b}{a}\mu = 0,$$

which has no solution for $\mu \neq 0$ (since Σ is positive definite). Thus any nonzero critical μ is an eigenvector of Σ. Writing $\Sigma v = \sigma v$ and $\mu = \alpha v$ we obtain

$$a = \alpha^2 \sigma, \qquad b = \alpha^2 \sigma^2,$$

and stationarity reduces to

$$\lambda \sigma \alpha^2 = 1 \quad \Longrightarrow \quad \alpha = \pm \frac{1}{\sqrt{\lambda \sigma}}.$$

Hence the nonzero critical points are precisely $\pm \frac{1}{\sqrt{\lambda \sigma}} v$ for eigenpairs (σ, v), together with the trivial critical point $\mu = 0$.

Now let us compute the Hessian,

$$\nabla^2_\mu \mathcal{R}_{\text{soft},\infty}(\mu) = -4\lambda \Sigma^2 + 4\lambda^2 \big(\Sigma^2 \mu \mu^\top \Sigma + \Sigma \mu \mu^\top \Sigma^2\big) + 2\lambda^2 \big(b\Sigma + a\Sigma^2\big).$$

Evaluate $\nabla^2_\mu \mathcal{R}_{\text{soft},\infty}$ at a critical $\mu^\pm_\star = \pm \alpha v$ with $\Sigma v = \sigma v$ and $\alpha^2 = \frac{1}{\lambda \sigma}$. For any eigenvector w of Σ with $\Sigma w = \tau w$ one obtains

$$\nabla^2_\mu \mathcal{R}_{\text{soft},\infty}(\mu^\pm_\star) w = \begin{cases} 2\lambda \tau(\sigma - \tau) w, & w \perp v, \\ 8\lambda \sigma^2 w, & w = v, \end{cases}$$

which follows from substituting $a = 1/\lambda$, $b = \sigma/\lambda$ and simplifying. From this:

- At $\mu = 0$ we have $a = b = 0$ and $\nabla^2_\mu \mathcal{R}_{\text{soft},\infty}(0) = -4\lambda \Sigma^2$, which is negative definite; hence $\mu = 0$ is a strict local maximum.
- If there exists $\tau > \sigma$ (in particular if $\sigma < \sigma_1$) then for that τ the curvature $2\lambda\tau(\sigma - \tau)$ is negative, so the critical point is a strict saddle.
- If $\sigma = \sigma_1$, then for every eigenvalue $\tau < \sigma_1$ we have $\tau(\sigma_1 - \tau) > 0$, so the Hessian at $\mu^\pm_\star = \pm \frac{1}{\sqrt{\lambda \sigma_1}} u_1$ is positive definite. Moreover, evaluating the objective yields

  $$\mathcal{R}_{\text{soft},\infty}(\mu^\pm_\star) = \operatorname{tr}(\Sigma) - \sigma_1,$$

  and by Lemma 2, these local minimizers are, in fact, global minimizers.

□

B.3 Proof of Proposition 2

Proof. The claim follows from the fact that this function is algebraic, and for such a function, gradient flow (and gradient descent with a proper stepsize) avoids strict saddle points for almost every initialization. By the Stable Manifold Theorem (Shub, 1987, Theorem III.7), the set of initial conditions whose trajectories converge to such a point is contained in its stable manifold, which has Lebesgue measure zero. Hence almost every initialization converges to a local minimum (see, e.g., (Lee et al., 2016; Panageas and Piliouras, 2017)). □

B.4 Proof of Proposition 3

Proof. Since $\mathcal{R}_{\text{soft},\infty}$ is a polynomial, it is real analytic. Moreover, by Proposition 1 its critical points are finite and nondegenerate. Therefore, by Corollary 2 it satisfies the Łojasiewicz inequality with exponent $\theta = \frac{1}{2}$ in a neighborhood of each critical point. Let

$$E(t) = \mathcal{R}_{\text{soft},\infty}(\mu_\infty(t)) - \mathcal{R}_{\text{soft},\infty}(\mu^\star).$$

Along the flow,

$$\dot{E}(t) = -\|\nabla \mathcal{R}_{\text{soft},\infty}(\mu_\infty(t))\|^2.$$

By the Łojasiewicz inequality, there exists $t_0 > 0$ and $c > 0$ such that for every $t \geq t_0$

$$\|\nabla \mathcal{R}_{\text{soft},\infty}(\mu_\infty(t))\| \geq c\, E(t)^{1/2}.$$

Thus

$$\dot{E}(t) \leq -c^2 E(t),$$

and Grönwall's inequality yields

$$E(t) \leq E(0)e^{-c^2 t}.$$

Next, since $\mu^\star$ is a nondegenerate minimizer, $\nabla^2 \mathcal{R}_{\text{soft},\infty}(\mu^\star)$ is positive definite. By continuity of the Hessian, for any $\varepsilon > 0$ there exists a neighborhood of $\mu^\star$ such that for all μ in this neighborhood,

$$\|\nabla^2 \mathcal{R}_{\text{soft},\infty}(\mu) - \nabla^2 \mathcal{R}_{\text{soft},\infty}(\mu^\star)\| \leq \varepsilon.$$

Using the integral form of Taylor's theorem,

$$\nabla \mathcal{R}_{\text{soft},\infty}(\mu) = \nabla^2 \mathcal{R}_{\text{soft},\infty}(\mu^\star)(\mu - \mu^\star) + r(\mu),$$

where $\|r(\mu)\| \leq \varepsilon \|\mu - \mu^\star\|$. Therefore,

$$\langle \nabla \mathcal{R}_{\text{soft},\infty}(\mu), \mu - \mu^\star \rangle \geq \langle \nabla^2 \mathcal{R}_{\text{soft},\infty}(\mu^\star)(\mu - \mu^\star), \mu - \mu^\star \rangle - \varepsilon \|\mu - \mu^\star\|^2.$$

Since the smallest eigenvalue of $\nabla^2 \mathcal{R}_{\text{soft},\infty}(\mu^\star)$ is

$$\tilde{s} = 2\lambda \min\{\sigma_2(\sigma_1 - \sigma_2), \sigma_d(\sigma_1 - \sigma_d)\},$$

it follows that

$$\langle \nabla \mathcal{R}_{\text{soft},\infty}(\mu), \mu - \mu^\star \rangle \geq (\tilde{s} - \varepsilon)\|\mu - \mu^\star\|^2.$$

By Proposition 2, we have $\mu_\infty(t) \to \mu^\star$, then there exists $t_0 > 0$ such that this inequality holds for all $t \geq t_0$. Then

$$\frac{d}{dt}\|\mu_\infty(t) - \mu^\star\|^2 = -2\langle \nabla \mathcal{R}_{\text{soft},\infty}(\mu_\infty(t)), \mu_\infty(t) - \mu^\star \rangle \leq -2(\tilde{s} - \varepsilon)\|\mu_\infty(t) - \mu^\star\|^2.$$

Applying Grönwall's inequality,

$$\|\mu_\infty(t) - \mu^\star\|^2 \leq \|\mu_\infty(t_0) - \mu^\star\|^2 e^{-2(\tilde{s}-\varepsilon)(t-t_0)},$$

we conclude that

$$\|\mu_\infty(t) - \mu^\star\| \leq Ce^{-(\tilde{s}-\varepsilon)(t-t_0)}.$$

□

B.5 Proof of Proposition 4

Proof. For simplicity, we will use the following shortcut notation

$$T_L := T_L^{\text{soft},\mu}(\mathbb{X})_1, \qquad T_\infty := T_\infty^{\text{soft},\mu}(X_1).$$

We first note that Assertion 1 is a direct consequence of Lemma 7. Besides, Assertion 2 directly follows from

$$\sup_{\mu \in B(0,\rho)} \|D_\mu^k T_\infty\|_F \leq C_k' \|X_1\|,$$

for $C_0' = \lambda\|\Sigma\|\rho^2, C_1' = 2\lambda\|\Sigma\|\rho, C_2' = 2\lambda\|\Sigma\|$. Defining

$$C_k = (C_k')^2 \text{tr}(\Sigma), \quad k \in \{0, 1, 2\},$$

we get the required.

Now we prove Assertion 3, from Assertion 1 with $k = 0$ and Assertion 2,

$$\mathbb{E}[\|T_L\|^2] \leq 2\mathbb{E}[\|T_L - T_\infty\|^2] + 2\mathbb{E}[\|T_\infty\|^2] \leq 2\psi_0(L) + 2C_0.$$

Hence $\sup_{L,\mu} \mathbb{E}[\|T_L\|^2] < \infty$. The same argument using $k = 1, 2$ gives uniform L^2-bounds for $D_\mu T_L$ and $D_\mu^2 T_L$.

Case $k = 0$. Using

$$\|a\|^2 - \|b\|^2 = 2\langle b, a - b\rangle + \|a - b\|^2,$$

we obtain

$$\mathcal{R}_{\mathrm{soft},L} - \mathcal{R}_{\mathrm{soft},\infty} = 2\mathbb{E}[\langle X_1 - T_\infty, T_\infty - T_L\rangle] + \mathbb{E}[\|T_L - T_\infty\|^2].$$

By Cauchy–Schwarz and the uniform L^2-bounds,

$$|\mathcal{R}_{\mathrm{soft},L} - \mathcal{R}_{\mathrm{soft},\infty}| \le C\mathbb{E}[\|T_L - T_\infty\|^2]^{1/2}.$$

Taking the supremum over μ and squaring gives

$$\sup_{\mu \in B(0,\rho)} |\mathcal{R}_{\mathrm{soft},L} - \mathcal{R}_{\mathrm{soft},\infty}|^2 \le C\psi_0(L).$$

Case $k = 1$. Differentiating under the expectation gives

$$\nabla\mathcal{R}_{\mathrm{soft},L} = -2\mathbb{E}[(X_1 - T_L)^\top D_\mu T_L].$$

Hence

$$\nabla\mathcal{R}_{\mathrm{soft},L} - \nabla\mathcal{R}_{\mathrm{soft},\infty} = -2\mathbb{E}[(X_1 - T_L)^\top (D_\mu T_L - D_\mu T_\infty)] - 2\mathbb{E}[(T_\infty - T_L)^\top D_\mu T_\infty].$$

Each term is bounded using Cauchy–Schwarz and the uniform L^2 bounds:

$$\|\nabla\mathcal{R}_{\mathrm{soft},L} - \nabla\mathcal{R}_{\mathrm{soft},\infty}\|_F^2 \le C\big(\mathbb{E}[\|D_\mu T_L - D_\mu T_\infty\|_F^2] + \mathbb{E}[\|T_L - T_\infty\|^2]\big).$$

Taking the supremum over μ yields

$$\sup_{\mu \in B(0,\rho)} \|\nabla\mathcal{R}_{\mathrm{soft},L} - \nabla\mathcal{R}_{\mathrm{soft},\infty}\|_F^2 \le C\big(\psi_1(L) + \psi_0(L)\big).$$

Case $k = 2$. We compute

$$\nabla^2\mathcal{R}_{\mathrm{soft},L} = 2\mathbb{E}[(D_\mu T_L)^\top (D_\mu T_L)] - 2\mathbb{E}[(X_1 - T_L)^\top D_\mu^2 T_L].$$

and the analogous formula for $\mathcal{R}_{\mathrm{soft},\infty}$. Set

$$\Delta T := T_L - T_\infty, \qquad \Delta DT := D_\mu T_L - D_\mu T_\infty, \qquad \Delta D^2 T := D_\mu^2 T_L - D_\mu^2 T_\infty.$$

Regarding the first term of the Hessian, we rewrite it as

$$(D_\mu T_L)^\top (D_\mu T_L) - (D_\mu T_\infty)^\top (D_\mu T_\infty) = (\Delta DT)^\top (\Delta DT) + (D_\mu T_\infty)^\top (\Delta DT) + (\Delta DT)^\top (D_\mu T_\infty).$$

Therefore,

$$\begin{aligned}&\mathbb{E}\big[(D_\mu T_L)^\top (D_\mu T_L)\big] - \mathbb{E}\big[(D_\mu T_\infty)^\top (D_\mu T_\infty)\big] \\ &= \mathbb{E}\big[(\Delta DT)^\top (\Delta DT)\big] + \mathbb{E}\big[(D_\mu T_\infty)^\top (\Delta DT)\big] + \mathbb{E}\big[(\Delta DT)^\top (D_\mu T_\infty)\big].\end{aligned}$$

With respect to the second term, we have

$$(X_1 - T_L)^\top D_\mu^2 T_L - (X_1 - T_\infty)^\top D_\mu^2 T_\infty = A_1 + A_2,$$

where

$$A_1 := (X_1 - T_L)^\top (D_\mu^2 T_L - D_\mu^2 T_\infty),$$
$$A_2 := \big[(X_1 - T_L) - (X_1 - T_\infty)\big]^\top D_\mu^2 T_\infty.$$

Since $(X_1 - T_L) - (X_1 - T_\infty) = T_\infty - T_L = -\Delta T$, we obtain

$$A_2 = -(\Delta T)^\top D_\mu^2 T_\infty.$$

Hence,

$$\begin{aligned}&\mathbb{E}\big[(X_1 - T_L)^\top D_\mu^2 T_L\big] - \mathbb{E}\big[(X_1 - T_\infty)^\top D_\mu^2 T_\infty\big] \\ &= \mathbb{E}\big[(X_1 - T_L)^\top \Delta D^2 T\big] - \mathbb{E}\big[(\Delta T)^\top D_\mu^2 T_\infty\big].\end{aligned}$$

Combining the previous expansions,

$$\begin{aligned}\nabla^2 \mathcal{R}_{\text{soft},L} - \nabla^2 \mathcal{R}_{\text{soft},\infty} = \; & 2\mathbb{E}[(\Delta DT)^\top (\Delta DT)] \\ & + 2\mathbb{E}[(D_\mu T_\infty)^\top (\Delta DT)] + 2\mathbb{E}[(\Delta DT)^\top (D_\mu T_\infty)] \\ & - 2\mathbb{E}[(X_1 - T_L)^\top \Delta D^2 T] \\ & + 2\mathbb{E}[(\Delta T)^\top D_\mu^2 T_\infty].\end{aligned}$$

Using $\|\mathbb{E}[Z]\|_F^2 \le \mathbb{E}[\|Z\|_F^2]$ and Cauchy–Schwarz:

$$\|\mathbb{E}[(\Delta DT)^\top (\Delta DT)]\|_F \le \mathbb{E}[\|\Delta DT\|_F^2],$$

$$\|\mathbb{E}[(D_\mu T_\infty)^\top (\Delta DT)]\|_F \le \mathbb{E}[\|D_\mu T_\infty\|_F^2]^{1/2} \mathbb{E}[\|\Delta DT\|_F^2]^{1/2},$$

$$\|\mathbb{E}[(X_1 - T_L)^\top \Delta D^2 T]\|_F \le \mathbb{E}[\|X_1 - T_L\|^2]^{1/2} \mathbb{E}[\|\Delta D^2 T\|_F^2]^{1/2},$$

$$\|\mathbb{E}[(\Delta T)^\top D_\mu^2 T_\infty]\|_F \le \mathbb{E}[\|\Delta T\|^2]^{1/2} \mathbb{E}[\|D_\mu^2 T_\infty\|_F^2]^{1/2}.$$

By Assertion 2, for $\mu \in B(0, \rho)$ there exists $C > 0$ such that

$$\|\nabla^2 \mathcal{R}_{\text{soft},L} - \nabla^2 \mathcal{R}_{\text{soft},\infty}\|_F^2 \le C\big(\mathbb{E}[\|T_L - T_\infty\|^2] + \mathbb{E}[\|D_\mu T_L - D_\mu T_\infty\|_F^2] + \mathbb{E}[\|D_\mu^2 T_L - D_\mu^2 T_\infty\|_F^2]\big).$$

Taking the supremum over μ concludes the proof. □

B.6 Proof of Proposition 5

Proof. By Lemma 1, if $\|\mu\| \ge \rho$, then $\mathcal{R}_{\text{soft},\infty}(\mu) > \mathcal{R}_{\text{soft},\infty}(\mu_0)$. Thus, for every $\rho' > \rho$ we have that if $\|\mu\| = \rho' > \rho$ then

$$\mathcal{R}_{\text{soft},\infty}(\mu) > \mathcal{R}_{\text{soft},\infty}(\mu_0),$$

by compactness of $\{\mu \in \mathbb{R}^d : \|\mu\| = \rho'\}$ and continuity of $\mathcal{R}_{\text{soft},\infty}$, thus

$$\min_{\|\mu\| = \rho'} \mathcal{R}_{\text{soft},\infty}(\mu) > \mathcal{R}_{\text{soft},\infty}(\mu_0),$$

so there exists $\varepsilon > 0$ such that

$$\min_{\|\mu\| = \rho'} \mathcal{R}_{\text{soft},\infty}(\mu) > \mathcal{R}_{\text{soft},\infty}(\mu_0) + 2\varepsilon,$$

and then

$$\|\mu\| = \rho' \quad \Longrightarrow \quad \mathcal{R}_{\text{soft},\infty}(\mu) > \mathcal{R}_{\text{soft},\infty}(\mu_0) + 2\varepsilon. \tag{15}$$

Fix any $\rho' > \rho$, then there exists $\varepsilon > 0$ that satisfies (15). By Proposition 4, we have uniform convergence of $\mathcal{R}_{\text{soft},L}$ to $\mathcal{R}_{\text{soft},\infty}$ on $B(0, \rho')$, there exists $L' \in \mathbb{N}$ such that for all $L \ge L'$,

$$\sup_{\|\mu\| \le \rho'} \big|\mathcal{R}_{\text{soft},L}(\mu) - \mathcal{R}_{\text{soft},\infty}(\mu)\big| \le \varepsilon.$$

We argue by contradiction. Assume that for some $L \geq L'$, the trajectory μ_L does not remain in $B(0, \rho')$ for every $t \geq 0$. Then, there exists a first exit time $t^* > 0$ such that

$$\|\mu_L(t^*)\| = \rho', \qquad \|\mu_L(t)\| < \rho' \text{ for all } t < t^*.$$

Since $\mathcal{R}_{\text{soft},L}$ decreases along the flow,

$$\mathcal{R}_{\text{soft},L}(\mu_L(t^*)) \leq \mathcal{R}_{\text{soft},L}(\mu_0).$$

Using uniform convergence of the risk on $B(0, \rho')$ at $\mu_L(t^*)$ and at μ_0 gives

$$\mathcal{R}_{\text{soft},\infty}(\mu_L(t^*)) \leq \mathcal{R}_{\text{soft},L}(\mu_L(t^*)) + \varepsilon \leq \mathcal{R}_{\text{soft},L}(\mu_0) + \varepsilon \leq \mathcal{R}_{\text{soft},\infty}(\mu_0) + 2\varepsilon.$$

This is a contradiction with (15). Hence, for all $L \geq L'$, the trajectory μ_L remains in $B(0, \rho')$ for all $t \geq 0$, proving the lemma. □

B.7 Proof of Proposition 6

Proof. Fix $T > 0$, by Proposition Lemma 5, there exists $\rho > 0$ and $L' \in \mathbb{N}$, such that for every $L \geq L'$ the trajectories $\mu_L(t)$ and $\mu_\infty(t)$ of (GF_L) and (GF_∞) satisfy $\mu_L(t), \mu_\infty(t) \in B(0, \rho)$ for all $t \in [0, T]$. Besides, in Proposition 4 we have shown that for $k \in \{0, 1, 2\}$

$$\nabla^k \mathcal{R}_{\text{soft},L} \underset{L\to\infty}{\to} \nabla^k \mathcal{R}_{\text{soft},\infty}$$

uniformly on $B(0, \rho)$, and $\nabla^2 \mathcal{R}_{\text{soft},L}$ are uniformly bounded on $B(0, \rho)$ (since it converges uniformly). Hence, there exists a constant $K > 0$ such that for all $L \geq L'$ and all $x, y \in B(0, \rho)$

$$\|\nabla \mathcal{R}_{\text{soft},L}(x) - \nabla \mathcal{R}_{\text{soft},L}(y)\| \leq K\|x - y\|, \qquad \|\nabla \mathcal{R}_{\text{soft},\infty}(x) - \nabla \mathcal{R}_{\text{soft},\infty}(y)\| \leq K\|x - y\|.$$

Set

$$\zeta(L) := \sup_{x \in B(0,\rho)} \|\nabla \mathcal{R}_{\text{soft},L}(x) - \nabla \mathcal{R}_{\text{soft},\infty}(x)\|.$$

By uniform convergence $\zeta(L) \to 0$ as $L \to \infty$. Consider the difference $e_L(t) := \mu_L(t) - \mu_\infty(t)$. Differentiating and using the definitions (GF_L), (GF_∞) we obtain

$$e_L'(t) = -\nabla \mathcal{R}_{\text{soft},L}(\mu_L(t)) + \nabla \mathcal{R}_{\text{soft},\infty}(\mu_\infty(t)).$$

Using the triangle inequality and the Lipschitz bound,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\|e_L(t)\| &\leq \|e_L'(t)\| \\
&\leq \|\nabla \mathcal{R}_{\text{soft},L}(\mu_L(t)) - \nabla \mathcal{R}_{\text{soft},L}(\mu_\infty(t))\| + \|\nabla \mathcal{R}_{\text{soft},L}(\mu_\infty(t)) - \nabla \mathcal{R}_{\text{soft},\infty}(\mu_\infty(t))\| \\
&\leq K\|e_L(t)\| + \zeta(L).
\end{aligned}$$

This differential inequality together with $e_L(0) = 0$, let us integrate to obtain

$$\|e_L(t)\| \leq \zeta(L) \int_0^t e^{K(t-s)}\, \mathrm{d}s = \zeta(L)\frac{e^{Kt} - 1}{K}.$$

Therefore, for every $t \in [0, T]$,

$$\sup_{0 \leq s \leq T} \|\mu_L(s) - \mu_\infty(s)\| \leq \zeta(L)\frac{e^{KT} - 1}{K} \xrightarrow[L\to\infty]{} 0,$$

which proves that $\mu_L \to \mu_\infty$ uniformly on $[0, T]$. □

B.8 Proof of Proposition 7

Proof. We use the following bound

$$\|\mathcal{R}_{\text{soft},L}(\mu_L) - \mathcal{R}_{\text{soft},\infty}(\mu_\infty)\| \leq \|\mathcal{R}_{\text{soft},L}(\mu_L) - \mathcal{R}_{\text{soft},\infty}(\mu_L)\| + \|\mathcal{R}_{\text{soft},\infty}(\mu_L) - \mathcal{R}_{\text{soft},\infty}(\mu_\infty)\|.$$

The first term goes to 0 uniformly on $[0,T]$ as μ_L is bounded by Proposition Lemma 5 and Proposition 4 for $k=0$. The second term goes to 0 uniformly on $[0,T]$ as $\mathcal{R}_{\text{soft},\infty}$ is C^2, combined with Proposition 6. □

B.9 Proof of Proposition 8

Proof. We use the characterization of critical points for $\mathcal{R}_{\text{soft},\infty}$ given in Proposition 1, this gives us $\text{crit}(\mathcal{R}_{\text{soft},\infty}) = \{\mu^{(1)}, \dots, \mu^{(2d+1)}\}$. Let $f_L = \mathcal{R}_{\text{soft},L}$ and $f = \mathcal{R}_{\text{soft},\infty}$, Proposition 4 shows convergence in C^2_{loc} of f_L towards f as $L \to \infty$. Then we use Lemma 5 with f_L and f, this gives us $r_k > 0$ such that $B(\mu^{(k)}, r_k)$ are pairwise disjoint and that for every $\rho > 0$ big enough such that $\cup_{k=1}^{2d+1} B(\mu^{(k)}, r_k) \subset B(0,\rho)$. Then for big enough L,

$$\text{crit}(f_L) \cap B(0,\rho) = \{\mu_L^{(1)}, \dots, \mu_L^{(\Lambda)}\},$$

with $\mu_L^{(k)} \to \mu^{(k)}$ as $L \to \infty$ for every $k \in \{1, \dots, 2d+1\}$, and if $\mu^{(k)}$ is a strict saddle point (resp. local minimum) for f, then $\mu_L^{(k)}$ is a strict saddle point (resp. local minimum) for f_L. □

B.10 Proof of Proposition 9

Proof. For L large enough, Proposition 8 ensures that the critical points of $\mathcal{R}_{\text{soft},L}$ are finite and nondegenerate, with two local minimizers $\pm\mu^\star_{L,\sigma_1}$. Hence, for a generic initialization, the gradient flow converges to one of them.

The exponential decay of the objective follows from the Łojasiewicz inequality, while the convergence rate of the iterates is given by linearization around $\mu_L^\star$, yielding $\tilde{s}_L = \sigma_{\min}(\nabla^2 \mathcal{R}_{\text{soft},L}(\mu_L^\star)) > 0$. The limit $\tilde{s}_L \to \tilde{s}$ follows from C^2_{loc} convergence of the risks. □

B.11 Proof of Proposition 10

Proof. The result follows from classical properties of linear transformations of Gaussian vectors, yielding

$$T_\infty^{\text{soft},\mu}(X_1) \sim \mathcal{N}(0, \Gamma(\mu)).$$

The rank-one structure is immediate. Evaluating at $\mu^\star$ and using $\Sigma u_1 = \sigma_1 u_1$ gives $\Gamma(\mu^\star) = \sigma_1 u_1 u_1^\top$, which corresponds to the law of $\langle X, u_1 \rangle u_1$. □

B.12 Proof of Proposition 11

Proof. The convergence rate is analogous to the proof of Proposition 3. From algebraic manipulation we can check that for a_i, b_i defined in (18), we have

$$\frac{(a_1+a_2)(a_3+b_3)}{a_1+2a_2+b_2} - a_3 < 2(a_3+b_3),$$

thus

$$\begin{aligned}
\hat{s} &= 2\,\frac{(a_1+a_2)b_3 - a_3(a_2+b_2)}{a_1+2a_2+b_2} \\
&= \frac{2\lambda n\theta\xi^2\left(dn\xi^2 - d\xi^2 + n^2\theta + n^2\xi^2 + 5n\theta + 4n\xi^2 + 2\theta + 3\xi^2\right)}{d\xi^2 + n\theta + n\xi^2 + 4\theta + 3\xi^2} \\
&= \frac{2\lambda n\theta\xi^2\left[\theta(n^2+5n+2) + \xi^2(n^2+dn+4n-d+3)\right]}{(n+4)\theta + (d+n+3)\xi^2}.
\end{aligned} \tag{16}$$

For ξ^2 small enough, this reduces to

$$\hat{s} \sim 2\frac{\lambda n(n^2+5n+2)\theta\,\xi^2}{n+4}.$$

Hence the gradient flow converges locally with an exponential rate arbitrarily close to $\hat{s}$. □

B.13 Proof of Proposition 12

Proof. Since we have proven in Proposition 4 that $\nabla^k \mathcal{R}^{(\Sigma)}_{\text{soft},L}$ converges uniformly (as $L \to \infty$) on compact sets $\nabla^k \mathcal{R}^{(\Sigma)}_{\text{soft},\infty}$ for $k \in \{0,1,2\}$, i.e., there exists $C(\Sigma) > 0$,

$$\psi(L,\Sigma) = L^{-\frac{1}{c^2\|\Sigma\|_2^2+1}}(1+\ln L)^{\frac{c^2\|\Sigma\|_2^2}{c^2\|\Sigma\|_2^2+1}}$$

with $c = 20\lambda\rho^2 > 0$, and $\lim_{L\to\infty}\psi(L,\Sigma) = 0$, such that

$$\sup_{\mu\in B(0,\rho)} \|\nabla^k \mathcal{R}_{\text{soft},L}(\mu) - \nabla^k \mathcal{R}_{\text{soft},\infty}(\mu)\|_F^2 \le C(\Sigma)\psi(L,\Sigma).$$

Thus for $k \in \{0,1,2\}$, since $\mathcal{R}^{(\Sigma)}_{\text{soft},L}$ and $\mathcal{R}^{(\Sigma)}_{\text{soft},\infty}$ are $C^2(\mathbb{R}^d)$ functions and

$$\mathbb{E}_\Sigma\left[\sup_{\mu\in B(0,\rho)} \nabla^k \mathcal{R}^{(\Sigma)}_{\text{soft},L}(\mu)\right] < \infty, \quad \mathbb{E}_\Sigma\left[\sup_{\mu\in B(0,\rho)} \nabla^k \mathcal{R}^{(\Sigma)}_{\text{soft},\infty}(\mu)\right] < \infty.$$

By the dominated convergence theorem we have that

$$\nabla^k \mathcal{R}^{\text{ICL}}_L(\mu) = \mathbb{E}_\Sigma[\nabla^k \mathcal{R}^{(\Sigma)}_{\text{soft},L}(\mu)], \quad \nabla^k \mathcal{R}^{\text{ICL}}_\infty(\mu) = \mathbb{E}_\Sigma[\nabla^k \mathcal{R}^{(\Sigma)}_{\text{soft},\infty}(\mu)],$$

and then get

$$\begin{aligned}
\sup_{\mu\in B(0,\rho)} \|\nabla^k \mathcal{R}^{\text{ICL}}_L(\mu) - \nabla^k \mathcal{R}^{\text{ICL}}_\infty(\mu)\|_F^2 &= \sup_{\mu\in B(0,\rho)} \|\nabla^k \mathbb{E}_\Sigma[\mathcal{R}^{(\Sigma)}_{\text{soft},L}(\mu)] - \nabla^k \mathbb{E}_\Sigma[\mathcal{R}^{(\Sigma)}_{\text{soft},\infty}(\mu)]\|_F^2 \\
&\le \sup_{\mu\in B(0,\rho)} \|\mathbb{E}_\Sigma[\nabla^k \mathcal{R}^{(\Sigma)}_{\text{soft},L}(\mu) - \nabla^k \mathcal{R}^{(\Sigma)}_{\text{soft},\infty}(\mu)]\|_F^2 \\
&\le \mathbb{E}_\Sigma\left[\sup_{\mu\in B(0,\rho)} \|\nabla^k \mathcal{R}^{(\Sigma)}_{\text{soft},L}(\mu) - \nabla^k \mathcal{R}^{(\Sigma)}_{\text{soft},\infty}(\mu)\|_F^2\right] \\
&\le \mathbb{E}_{\Sigma\sim W_d(V,n)}[C(\Sigma)\psi(L,\Sigma)].
\end{aligned}$$

Now, we define $\zeta(\Sigma) = C(\Sigma)\sup_{L>1}\psi(L,\Sigma)$, it is direct to show that

$$\sup_{L>1}\psi(L,\Sigma) \le 1 + c^2\|\Sigma\|_2^2,$$

and then

$$\mathbb{E}_{\Sigma\sim W_d(V,n)}[\zeta(\Sigma)] \le \mathbb{E}_{\Sigma\sim W_d(V,n)}[C(\Sigma)(1+c^2\|\Sigma\|_2^2)].$$

Since $C(\Sigma)$ depends polynomially on $\|\Sigma\|_2$ and that for the Wishart distribution every finite moment is bounded, we get that $\mathbb{E}_{\Sigma\sim W_d(V,n)}[\zeta(\Sigma)] < \infty$, by dominated convergence theorem we get that $\lim_{L\to\infty}\mathbb{E}_{\Sigma\sim W_d(V,n)}[C(\Sigma)\psi(L,\Sigma)]$ $\mathbb{E}_{\Sigma\sim W_d(V,n)}[C(\Sigma)\lim_{L\to\infty}\psi(L,\Sigma)] = 0$, and then for $k \in \{0,1,2\}$,

$$\lim_{L\to\infty}\sup_{\mu\in B(0,\rho)} \|\nabla^k \mathcal{R}^{\text{ICL}}_L(\mu) - \nabla^k \mathcal{R}^{\text{ICL}}_\infty(\mu)\|_F^2 = 0$$

□

B.14 Proof of Proposition 13

Proof. Without loss of generality, by the arguments of Proposition 5, we can assume that the trajectory remains bounded and contained in $B(0,\rho)$. In particular, all critical points in $B(0,\rho)$ are those identified in Proposition 20, and the only stable ones are the two local minimizers $\pm\mu^\star_{L,\|}$. Applying the Stable Manifold Theorem (Shub, 1987, Theorem III.7), we conclude that, for almost every initialization in $B(0,\rho)$, the trajectory converges to one of these two minimizers. The convergence rates come from the same reasoning as for Proposition 9. □

C Technical results

In this section, we present the technical results required for our analysis.

C.1 Gaussian computations

We compute specific higher-order moments required for the construction of the objective function.

Proposition 17. *Let $X_1, X_2 \overset{\text{iid}}{\sim} \mathcal{N}(0,\Sigma)$ in $\mathbb{R}^d$ and let $\mu \in \mathbb{R}^d$. We have*

1. $\mathbb{E}[\|X_1\|^2] = tr(\Sigma)$.
2. $\mathbb{E}\left[(X_1^\top \mu)^2\right] = \mu^T\Sigma\mu$.
3. $\mathbb{E}\left[\langle X_1, X_2\rangle\langle X_1,\mu\rangle\langle X_2,\mu\rangle\right] = \mu^T\Sigma^2\mu$.
4. $\mathbb{E}\left[\langle X_1, X_2\rangle\langle X_1,\mu\rangle^3\langle X_2,\mu\rangle\right] = 3(\mu^T\Sigma\mu)(\mu^T\Sigma^2\mu)$.
5. $\mathbb{E}\left[\langle X_1,\mu\rangle^2\|X_1\|^2\right] = tr(\Sigma)(\mu^T\Sigma\mu) + 2\mu^T\Sigma^2\mu$.
6. $\mathbb{E}\left[\langle X_1,\mu\rangle^4\|X_1\|^2\right] = 3tr(\Sigma)(\mu^T\Sigma\mu)^2 + 12(\mu^T\Sigma\mu)(\mu^T\Sigma^2\mu)$.

Proof. All identities follow from linearity of expectation, independence of X_1 and X_2, and Isserlis' theorem (Isserlis, 1918) for centered Gaussian vectors.

1. Since $\|X_1\|^2 = \sum_{i=1}^d X_{1,i}^2$ and $\mathbb{E}[X_{1,i}^2] = \Sigma_{ii}$,
$$\mathbb{E}[\|X_1\|^2] = \sum_{i=1}^d \Sigma_{ii} = \mathrm{tr}(\Sigma).$$
2. Expanding $(X_1^\top\mu)^2$ and using $\mathbb{E}[X_{1,i}X_{1,j}] = \Sigma_{ij}$ gives
$$\mathbb{E}[(X_1^\top\mu)^2] = \sum_{i,j}\mu_i\mu_j\Sigma_{ij} = \mu^\top\Sigma\mu.$$
3. Conditioning on X_1 and using independence,
$$\mathbb{E}[\langle X_1,X_2\rangle\langle X_2,\mu\rangle \mid X_1] = X_1^\top\mathbb{E}[X_2X_2^\top]\mu = X_1^\top\Sigma\mu.$$
Therefore,
$$\mathbb{E}[\langle X_1,X_2\rangle\langle X_1,\mu\rangle\langle X_2,\mu\rangle] = \mathbb{E}[(X_1^\top\mu)(X_1^\top\Sigma\mu)] = \mu^\top\Sigma^2\mu.$$
4. Conditioning again on X_1,
$$\mathbb{E}[\langle X_1,X_2\rangle\langle X_1,\mu\rangle^3\langle X_2,\mu\rangle] = \mathbb{E}[(X_1^\top\mu)^3(\mu^\top\Sigma X_1)].$$
For a centered Gaussian vector X and vectors μ_0,μ_1, Isserlis' theorem yields
$$\mathbb{E}[(\mu_0^\top X)^3(\mu_1^\top X)] = 3(\mu_0^\top\Sigma\mu_0)(\mu_0^\top\Sigma\mu_1).$$
Applying this with $a=\mu_0$ and $\mu_1 = \Sigma\mu$ gives the claim.

5. Expanding
$$\langle X_1, \mu\rangle^2 \|X_1\|^2 = \sum_{i,j,k} \mu_i \mu_j X_{1,i} X_{1,j} X_{1,k}^2$$
and applying Isserlis' theorem to the fourth-order moment yields two types of pairings, leading to
$$\mathbb{E}[\langle X_1, \mu\rangle^2 \|X_1\|^2] = \mathrm{tr}(\Sigma)(\mu^\top \Sigma \mu) + 2\mu^\top \Sigma^2 \mu.$$

6. Similarly,
$$\langle X_1, \mu\rangle^4 \|X_1\|^2 = \sum_{i,j,\ell,m,k} \mu_i \mu_j \mu_\ell \mu_m X_{1,i} X_{1,j} X_{1,\ell} X_{1,m} X_{1,k}^2.$$
Applying Isserlis' theorem to the sixth-order moment, pairings where the two copies of k are paired together contribute $3\mathrm{tr}(\Sigma)(\mu^\top \Sigma \mu)^2$, while the remaining 12 pairings contribute $12(\mu^\top \Sigma \mu)(\mu^\top \Sigma^2 \mu)$. Summing both terms gives the result.

□

C.2 Optimization preliminaries

In this subsection, we gather several optimization results that will be used throughout this work.

Lemma 2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a $C^1(\mathbb{R}^d)$ coercive function. Assume that the set of critical points of f is completely characterized and every critical point is either a strict saddle or a local minimum. Assume moreover that all local minima attain the same value m. Then every local minimum of f is a global minimum, and*
$$\min_{x \in \mathbb{R}^d} f(x) = m.$$

Proof. Since f is coercive and continuous, it attains its global minimum at some point $x^\star \in \mathbb{R}^d$. Since f is C^1, any global minimizer satisfies $\nabla f(x^\star) = 0$, hence $x^\star$ is a critical point. By assumption, every critical point is either a strict saddle or a local minimum. A strict saddle cannot be a minimizer, therefore $x^\star$ must be a local minimum. Since all local minima have value m, we obtain
$$f(x^\star) = m.$$
Thus $m = \min f$, and every local minimum is a global minimum. □

Lemma 3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be C^2 and let $x^\star$ be a nondegenerate critical point, i.e.,*
$$\nabla f(x^\star) = 0 \quad \textit{and} \quad \nabla^2 f(x^\star) \textit{ is invertible.}$$
Then there exist constants $C > 0$ and a neighborhood U of $x^\star$ such that
$$\|\nabla f(x)\| \geq C\, |f(x) - f(x^\star)|^{1/2} \quad \textit{for all } x \in U.$$

Proof. Since $\nabla^2 f(x^\star)$ is invertible, there exists $c_1 > 0$ such that
$$\|\nabla^2 f(x^\star) v\| \geq c_1 \|v\| \quad \text{for all } v \in \mathbb{R}^d.$$

By Taylor's theorem with integral remainder,
$$\nabla f(x) = \nabla^2 f(x^\star)(x - x^\star) + r(x),$$
where
$$r(x) = \int_0^1 \big(\nabla^2 f(x^\star + t(x - x^\star)) - \nabla^2 f(x^\star)\big)(x - x^\star)\, dt.$$

Since $\nabla^2 f$ is continuous, for any $\varepsilon > 0$ there exists $\delta > 0$ such that for all $\|x - x^\star\| \leq \delta$,
$$\|\nabla^2 f(x) - \nabla^2 f(x^\star)\| \leq \varepsilon.$$

Hence

$$\|r(x)\| \le \varepsilon \|x - x^\star\|.$$

Thus

$$\|\nabla f(x)\| \ge \|\nabla^2 f(x^\star)(x - x^\star)\| - \|r(x)\| \ge (c_1 - \varepsilon)\|x - x^\star\|.$$

Choosing $\varepsilon = c_1/2$, we obtain

$$\|\nabla f(x)\| \ge c_2 \|x - x^\star\|$$

for $c_2 = \frac{c_1}{2} > 0$ and all $\|x - x^\star\| \le \delta$.

Next, by Taylor expansion of f,

$$f(x) - f(x^\star) = \frac{1}{2}(x - x^\star)^T \nabla^2 f(x^\star)(x - x^\star) + \tilde{r}(x),$$

where $|\tilde{r}(x)| \le \varepsilon \|x - x^\star\|^2$ for $\|x - x^\star\|$ small enough.

Since $\nabla^2 f(x^\star)$ is bounded, there exists $c_3 > 0$ such that

$$|(x - x^\star)^T \nabla^2 f(x^\star)(x - x^\star)| \le c_3 \|x - x^\star\|^2.$$

Thus

$$|f(x) - f(x^\star)| \le c_4 \|x - x^\star\|^2$$

for some $c_4 > 0$.

Combining both estimates yields

$$\|\nabla f(x)\| \ge c_2 \|x - x^\star\| \ge \frac{c_2}{\sqrt{c_4}} |f(x) - f(x^\star)|^{1/2}.$$

□

Corollary 2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be real analytic and assume that all its critical points are nondegenerate and finite. Then for each critical point $x^\star$, there exist $C > 0$ and a neighborhood U of $x^\star$ such that*

$$\|\nabla f(x)\| \ge C |f(x) - f(x^\star)|^{1/2} \quad \text{for all } x \in B(x^\star, \varepsilon).$$

Lemma 4. *Let $\alpha, \beta > 0$ and $L \ge 1$. Define, for $\Psi \ge 0$,*

$$\varphi_L(\Psi) \;:=\; \frac{e^{\alpha \Psi^2}}{L} \;+\; (1 + \ln L) e^{-\beta \Psi^2}.$$

Then φ_L admits a unique minimizer $\Psi(L) \ge 0$, given by

$$\Psi^2(L) = \frac{1}{\alpha + \beta} \Big(\ln L + \ln(1 + \ln L) + \ln(\beta/\alpha) \Big).$$

Moreover, the optimal value satisfies

$$\inf_{\Psi \ge 0} \varphi_L(\Psi) = \Theta(\psi_{\alpha,\beta}(L)),$$

where $\psi_{\alpha,\beta}(L) = L^{-\frac{\beta}{\alpha+\beta}} (1 + \ln L)^{\frac{\alpha}{\alpha+\beta}}$.

Proof. Set $x = \Psi^2 \ge 0$ and define

$$f_L(x) = \frac{e^{\alpha x}}{L} + (1 + \ln L) e^{-\beta x}.$$

The function f_L is strictly convex on $\mathbb{R}_+$ Therefore, any critical point is the unique global minimizer. Differentiating,

$$f_L'(x) = \frac{\alpha e^{\alpha x}}{L} - \beta (1 + \ln L) e^{-\beta x}.$$

Solving $f_L'(x) = 0$ yields

$$e^{(\alpha+\beta)x} = \frac{\beta}{\alpha} L(1 + \ln L),$$

which gives the stated expression for $\Psi^2(L)$. Substituting this value into either term of f_L,

$$\frac{e^{\alpha\Psi^2(L)}}{L} = L^{-\frac{\beta}{\alpha+\beta}} (1 + \ln L)^{\frac{\alpha}{\alpha+\beta}} \left(\frac{\beta}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}},$$

and the term $(1 + \ln L)e^{-\beta\Psi^2(L)}$ has the same order. The claim follows. □

Lemma 5 (Persistence of nondegenerate critical points)**.** *Let $f \in C^2(\mathbb{R}^d)$ and let $x^\star$ be a nondegenerate critical point of f, i.e.*

$$\nabla f(x^\star) = 0, \qquad \nabla^2 f(x^\star) \text{ is invertible.}$$

Assume $f_n \to f$ in $C^2_{\text{loc}}(\mathbb{R}^d)$. Then there exist $r > 0$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$:

1. *f_n has a unique critical point $x_n^\star$ in $B(x^\star, r)$,*
2. *$x_n^\star$ is nondegenerate and moreover $x_n^\star \to x^\star$ as $n \to \infty$,*
3. *If the set of critical points of f is finite, i.e.,*

$$\text{crit}(f) = \{x^{(1)}, \ldots, x^{(\Lambda)}\},$$

and each $x^{(k)}$, $k \in \{1, \ldots, \Lambda\}$, is nondegenerate, let $r_k > 0$ be such that $B(x^{(k)}, r_k)$ are pairwise disjoint. Let $\rho > 0$ such that

$$\bigcup_{k=1}^{\Lambda} B(x^{(k)}, r_k) \subset B(0, \rho).$$

Then for n large enough,

$$\text{crit}(f_n) \cap B(0, \rho) = \{x_n^{(1)}, \ldots, x_n^{(\Lambda)}\}.$$

Moreover, if for some $k \in \{1, \ldots, \Lambda\}$, $x^{(k)}$ is a strict saddle (resp. a strict local minimum), then $x_n^{(k)}$ is also a strict saddle (resp. a strict local minimum) for n large enough.

Proof. 1. **Uniform invertibility of the Hessian.** Since $\nabla^2 f(x^\star)$ is invertible and $\nabla^2 f$ is continuous, there exist $r > 0$ and $m > 0$ such that

$$\|\nabla^2 f(x)^{-1}\| \leq m \quad \text{for all } x \in B(x^\star, r).$$

Because $f_n \to f$ in C^2_{loc}, for n sufficiently large,

$$\sup_{x \in B(x^\star, r)} \|\nabla^2 f_n(x) - \nabla^2 f(x)\| \leq \frac{1}{2m}.$$

Hence for $x \in B(x^\star, r)$, we define $E_n(x) = \nabla^2 f_n(x) - \nabla^2 f(x)$

$$\nabla^2 f_n(x) = \nabla^2 f(x) + E_n(x), \qquad \|E_n(x)\| \leq \frac{1}{2m}.$$

Then, for n sufficiently large $\nabla^2 f_n(x)$ is invertible and let $A(x) = \nabla^2 f(x)$, then

$$\nabla^2 f_n(x) = A(x) + E_n(x)$$

And we have the identity

$$\nabla^2 f_n(x)^{-1} = (I_d + A(x)^{-1} E_n(x))^{-1} A(x)^{-1},$$

where

$$\|A(x)^{-1}E_n(x)\| \le \|A(x)^{-1}\|\|E_n(x)\| \le m\frac{1}{2m} = \frac{1}{2}.$$

By the Neumann Series Theorem we get that

$$\|(I_d + A(x)^{-1}E_n(x))^{-1}\| \le \frac{1}{1-\frac{1}{2}} = 2.$$

Thus, for n sufficiently large, we get

$$\|\nabla^2 f_n(x)^{-1}\| = \|(I_d + A(x)^{-1}E_n(x))^{-1}\|\|A(x)^{-1}\| \le 2m.$$

Definition of the homotopy. Define

$$H_n(t,x) = (1-t)\nabla f(x) + t\nabla f_n(x), \qquad t \in [0,1].$$

Then

$$D_x H_n(t,x) = (1-t)\nabla^2 f(x) + t\nabla^2 f_n(x),$$
$$D_t H_n(t,x) = \nabla f_n(x) - \nabla f(x).$$

By the uniform invertibility of the Hessian, $D_x H_n(t,x)$ is invertible on $[0,1] \times B(x^\star, r)$ for n sufficiently large.

Technical condition. Let

$$M := \sup_{(t,x)\in[0,1]\times B(x^\star,r)} \|(D_x H_n(t,x))^{-1}\|.$$

From the uniform invertibility of $\nabla^2 f$ and $\nabla^2 f_n$, $M < \infty$ uniformly for n large.

Since $f_n \to f$ in C^1_{loc},

$$\sup_{x\in B(x^\star,r)} \|\nabla f_n(x) - \nabla f(x)\| \longrightarrow 0.$$

Therefore, for n sufficiently large,

$$\sup_{(t,x)\in[0,1]\times B(x^\star,r)} \|(D_x H_n(t,x))^{-1} D_t H_n(t,x)\| \le M \sup_{x\in B(x^\star,r)} \|\nabla f_n(x) - \nabla f(x)\| < r.$$

Conclusion. Since

$$H_n(0,x^\star) = \nabla f(x^\star) = 0,$$

Then by (Krantz and Parks, 2013, Theorem 4.2.1) yields a continuous curve $x_n(t)$ with

$$H_n(t, x_n(t)) = 0, \qquad x_n(0) = x^\star, \qquad x_n(t) \in B(x^\star, r).$$

Define $x_n^\star := x_n(1)$. Then

$$\nabla f_n(x_n^\star) = 0, \qquad x_n^\star \in B(x^\star, r).$$

Uniqueness follows from the local invertibility of $D_x H_n$.

2. Since $\nabla^2 f_n(x_n^\star) \to \nabla^2 f(x^\star)$ and the limit is invertible, $x_n^\star$ is nondegenerate for n large. Let $\Gamma := \frac{1}{\|[\nabla^2 f(x^\star)]^{-1}\|}$, by continuity of $\nabla^2 f$, shrinking r if necessary, we have for every $x \in B(x^\star, r)$,

$$\|\nabla^2 f(x) - \nabla^2 f(x^\star)\| \le \frac{\Gamma}{2},$$

and since $\|\nabla^2 f(x^\star)v\| \geq \Gamma\|v\|$, using Taylor we get

$$\nabla f(x) = \nabla^2 f(x^\star)(x - x^\star) + \int_0^1 (\nabla^2 f(x^\star + t(x - x^\star)) - \nabla^2 f(x^\star))(x - x^\star)dt.$$

We take norms and after bounding we obtain

$$\|x - x^\star\| \leq \frac{2}{\Gamma}\|\nabla f(x)\|,$$

choosing $x = x_n^\star$, in order to conclude that $x_n^\star \to x^\star$ as $n \to \infty$ we need to obtain that $\lim_{n\to\infty} \|\nabla f(x_n^\star)\| = 0$, but since

$$\nabla f(x_n^\star) = \nabla f(x_n^\star) - \nabla f_n(x_n^\star),$$

then we conclude since $f_n \to f$ in C^1_{loc}.

3. Assume now that

$$\text{crit}(f) = \{x^{(1)}, \ldots, x^{(\Lambda)}\}, \qquad \nabla^2 f(x^{(k)}) \text{ is invertible for all } k.$$

Since the set is finite, define

$$\delta := \frac{1}{2}\min_{k\neq\ell} \|x^{(k)} - x^{(\ell)}\| > 0.$$

Applying items (1)–(2) to each $x^{(k)}$, there exist $r_k \in (0, \delta)$ and $n_k \in \mathbb{N}$ such that for all $n \geq n_k$:

(a) f_n has a unique critical point $x_n^{(k)}$ in $B(x^{(k)}, r_k)$,

(b) $x_n^{(k)} \to x^{(k)}$,

(c) $x_n^{(k)}$ is nondegenerate.

Let

$$n_0 := \max_{k\in\{1,\ldots,\Lambda\}} n_k.$$

For $n \geq n_0$ the balls $B(x^{(k)}, r_k)$ are disjoint and

$$\{x_n^{(1)}, \ldots, x_n^{(\Lambda)}\} \subset \text{crit}(f_n).$$

It remains to show that there are no other critical points in $B(0, \rho)$.

Let

$$K := \bigcup_{k=1}^{\Lambda} B(x^{(k)}, r_k) \subset B(0, \rho),$$

then

$$\{x_n^{(1)}, \ldots, x_n^{(\Lambda)}\} \subseteq \text{crit}(f_n) \cap B(0, \rho). \tag{17}$$

Since f has no critical point in $K^c \cap B(0, \rho)$ and ∇f is continuous, the compactness of $K^c \cap B(0, \rho)$ implies

$$\eta_\rho := \inf_{x\in K^c\cap B(0,\rho)} \|\nabla f(x)\| > 0.$$

Because $f_n \to f$ in C^1_{loc}, we have

$$\sup_{x\in B(0,\rho)} \|\nabla f_n(x) - \nabla f(x)\| \to 0.$$

Hence for n large enough and every $x \in K^c \cap B(0,\rho)$,

$$\|\nabla f_n(x)\| \geq \|\nabla f(x)\| - \|\nabla f_n(x) - \nabla f(x)\| \geq \frac{\eta_\rho}{2} > 0.$$

Therefore f_n has no critical point in $K^c \cap B(0,\rho)$. Let $x \in \mathrm{crit}(f_n) \cap B(0,\rho)$, if $x \in K^c$ we get a contradiction with the previous founding. Then $x \in K$, and there exists $k \in \{1, ..., \Lambda\}$ such that $x \in B(x^{(k)}, r_k)$, since the only critical point of f_n inside this ball is $x_n^{(k)}$, we conclude that $x = x_n^{(k)}$ and

$$\mathrm{crit}(f_n) \cap B(0,\rho) \subseteq \{x_n^{(1)}, \dots, x_n^{(\Lambda)}\},$$

then by (17) we conclude the equality of sets.

Finally, since $x_n^{(k)} \to x^{(k)}$ and $f_n \to f$ in C^2_{loc},

$$\nabla^2 f_n(x_n^{(k)}) \to \nabla^2 f(x^{(k)}).$$

Because $\nabla^2 f(x^{(k)})$ is invertible, its eigenvalues are bounded away from zero. By continuity of the spectrum, the inertia of $\nabla^2 f_n(x_n^{(k)})$ coincides with that of $\nabla^2 f(x^{(k)})$ for n large enough. Hence if $x^{(k)}$ is a strict local minimum (resp. a strict saddle), then $x_n^{(k)}$ is also a strict local minimum (resp. a strict saddle) for n large enough.

□

C.3 Almost sure and L^2- convergence of encodings

In the following subsection, we present the lemmas needed to establish almost sure and L^2 convergence of $T^\mu_{\mathrm{soft},L}$ to $T^\mu_{\mathrm{soft},\infty}$.

Lemma 6. *Consider $X_1, \dots, X_L$ i.i.d $\mathcal{N}(0,\Sigma)$, and*

$$T_L(X_1) = \frac{\sum_{k=1}^L X_k \exp(\lambda \langle X_1, \mu\rangle \langle X_k, \mu\rangle)}{\sum_{k=1}^L \exp(\lambda \langle X_1, \mu\rangle \langle X_k, \mu\rangle)}.$$

And $T_\infty(X_1) = \lambda \Sigma \mu \mu^\top X_1$, we have that $T_L \to T_\infty$ a.s., and similarly for its Jacobian $D_\mu T_L(X_1) \to D_\mu T_\infty(X_1)$ a.s., and its Hessian $D^2_\mu T_L(X_1) \to D^2_\mu T_\infty(X_1)$ a.s..

Proof. Consider $X \sim \mathcal{N}(0,\Sigma)$ and let us fix $X_1 = z$, and define $\eta_k(z) = \eta_k(\mu, z) = \exp(\lambda \langle z, \mu\rangle \langle X_k, \mu\rangle)$, then

$$T_L(z) = \frac{\sum_{k=1}^L \eta_k(z) X_k}{\sum_{k=1}^L \eta_k(z)} = \frac{\frac{1}{L}\eta_1(z) X_1 + \frac{L-1}{L}\frac{1}{L-1}\sum_{k=2}^L \eta_k(z) X_k}{\frac{1}{L}\eta_1(z) + \frac{L-1}{L}\frac{1}{L-1}\sum_{k=2}^L \eta_k(z)}.$$

By the strong law of large numbers, we have that

$$\lim_{L\to\infty} T_L(z) = \frac{\mathbb{E}[X \exp(\lambda \langle X, \mu\rangle \langle z, \mu\rangle)]}{\mathbb{E}[\exp(\lambda \langle X, \mu\rangle \langle z, \mu\rangle)]} = \lambda \Sigma \mu \mu^\top z, \text{ a.s..}$$

Therefore

$$\mathbb{P}\left(\lim_{L\to\infty} T_L(X_1) = \lambda \Sigma \mu \mu^\top X_1 \middle| X_1\right) = 1.$$

Taking expectation w.r.t. X_1, we get that $\mathbb{P}\left(\lim_{L\to\infty} T_L(X_1) = \lambda \Sigma \mu \mu^\top X_1\right) = 1$ or $\lim_{L\to\infty} T_L(X_1) = \lambda \Sigma \mu \mu^\top X_1$ a.s..

First order. Let

$$N_L(\mu, z) = \frac{1}{L-1}\sum_{k=2}^{L} \eta_k(\mu, z) X_k, \quad S_L(\mu, z) = \frac{1}{L-1}\sum_{k=2}^{L} \eta_k(\mu, z),$$

then $T_L(z) = \frac{N_L(\mu,z)}{S_L(\mu,z)} + R_{1,L}(z)$ and

$$D_\mu T_L(z) = \frac{1}{S_L(\mu, z)} D_\mu(N_L(\mu, z)) - \frac{1}{S_L(\mu, z)^2} N_L(\mu, z)(D_\mu S_L(\mu, z))^\top + D_\mu R_{1,L}(z).$$

Using the strong law of large numbers, we get (since $\lim_{L\to\infty} D_\mu R_{1,L}(z) = 0$ a.s.)

$$\begin{aligned}\lim_{L\to\infty} D_\mu T_L(z) &= \frac{\mathbb{E}[X(\nabla_\mu \exp(\lambda\langle X, \mu\rangle\langle z, \mu\rangle))^\top]}{\mathbb{E}[\exp(\lambda\langle X, \mu\rangle\langle z, \mu\rangle)]} - \frac{\mathbb{E}[X\exp(\lambda\langle X, \mu\rangle\langle z, \mu\rangle)](\mathbb{E}[\nabla_\mu \exp(\lambda\langle X, \mu\rangle\langle z, \mu\rangle)])^\top}{\mathbb{E}[\exp(\lambda\langle X, \mu\rangle\langle z, \mu\rangle)]^2} \\ &= \lambda(\Sigma\mu z^\top + (\mu^\top z)\Sigma) = D_\mu T_\infty(z).\end{aligned}$$

And we conclude as before that $\lim_{L\to\infty} D_\mu T_L(X_1) = D_\mu T_\infty(X_1)$ a.s..

Second-order. Condition on $X_1 = z$ and recall that

$$N_L(\mu, z) = \frac{1}{L-1}\sum_{k=2}^{L} \eta_k(\mu, z) X_k, \qquad S_L(\mu, z) = \frac{1}{L-1}\sum_{k=2}^{L} \eta_k(\mu, z),$$

so that

$$T_L(z) = \frac{N_L(\mu, z)}{S_L(\mu, z)} + R_{1,L}(z).$$

Differentiating twice w.r.t. μ and using the quotient rule,

$$D_\mu^2 T_L(z) = D_\mu^2\left(\frac{N_L(z)}{S_L(z)}\right) + D_\mu^2 R_{1,L}(z),$$

where

$$\begin{aligned}D_\mu^2\left(\frac{N_L}{S_L}\right)[h] = \frac{1}{S_L} D_\mu^2 N_L[h] - \frac{1}{S_L^2}\Big((D_\mu N_L)\langle D_\mu S_L, h\rangle + (D_\mu S_L)\langle D_\mu N_L, h\rangle\Big) \\ - \frac{1}{S_L^2} N_L D_\mu^2 S_L[h] + \frac{2}{S_L^3} N_L \langle D_\mu S_L, h\rangle D_\mu S_L.\end{aligned}$$

By the strong law of large numbers applied componentwise to

$$\big\{X_k\eta_k(\mu, z),\ D_\mu[X_k\eta_k(\mu, z)],\ D_\mu^2[X_k\eta_k(\mu, z)]\big\}_{k\geq 2},$$

and to

$$\big\{\eta_k(\mu, z),\ D_\mu\eta_k(\mu, z),\ D_\mu^2\eta_k(\mu, z)\big\}_{k\geq 2},$$

we obtain almost surely the convergence of N_L, $D_\mu N_L$, $D_\mu^2 N_L$ and of S_L, $D_\mu S_L$, $D_\mu^2 S_L$ toward their respective expectations. We verify

$$\lim_{L\to\infty} D_\mu^2 T_L(z) = D_\mu^2 T_\infty(z),$$

where, using the closed form

$$T_\infty(z) = \lambda\Sigma\mu\mu^\top z,$$

the Hessian satisfies,

$$D_\mu^2 T_\infty(z)[h] = \lambda\big(\Sigma h z^\top + \langle h, z\rangle\Sigma\big).$$

Finally, as before we conclude

$$\lim_{L\to\infty} D_\mu^2 T_L(X_1) = D_\mu^2 T_\infty(X_1) \quad \text{a.s.}$$

□

Lemma 7. *Consider $X, X_1, \dots, X_L$ i.i.d $\mathcal{N}(0, \Sigma)$, and*

$$T_L^\mu(X_1) = \frac{\sum_{k=1}^L X_k \exp(\lambda \langle X_1, \mu\rangle \langle X_k, \mu\rangle)}{\sum_{k=1}^L \exp(\lambda \langle X_1, \mu\rangle \langle X_k, \mu\rangle)}, \quad T_\infty^\mu(X_1) = \lambda \Sigma \mu \mu^\top X_1,$$

and we define $g : \mathbb{N} \times \mathbb{R}_+ \times \mathbb{R}^d \times \mathbb{R}^{d\times d} \times \mathbb{N} \to \mathbb{R}$ by

$$g(L, \lambda, \mu, \Sigma, P) = L^{-\frac{1}{P^2\lambda^2(\mu^\top \Sigma \mu)^2+1}} (1 + \ln L)^{\frac{P^2\lambda^2(\mu^\top \Sigma \mu)^2}{P^2\lambda^2(\mu^\top \Sigma \mu)^2+1}},$$

then we have that

$$\mathbb{E}[\|T_L^\mu(X_1) - T_\infty^\mu(X_1)\|^2] = \mathcal{O}(g(L, \lambda, \mu, \Sigma, 12)).$$

Also,

$$\mathbb{E}[\|D_\mu T_L^\mu(X_1) - D_\mu T_\infty^\mu(X_1)\|_F^2] = \mathcal{O}(g(L, \lambda, \mu, \Sigma, 16)),$$

And

$$\mathbb{E}[\|D_\mu^2 T_L^\mu(X_1) - D_\mu^2 T_\infty^\mu(X_1)\|_F^2] = \mathcal{O}(g(L, \lambda, \mu, \Sigma, 20)).$$

We note $\lim_{L\to\infty} \sup_{\mu\in B(0,\rho)} g(L, \lambda, \mu, \Sigma, P) = 0$, since the dependency of μ inside the $\mathcal{O}-$term is polynomial, $T_L^\mu(X_1)$ and $DT_L^\mu(X_1)$ converge in L^2 uniformly over compact sets of μ to $T_\infty^\mu(X_1)$ and $DT_\infty^\mu(X_1)$, respectively, as $L \to \infty$.

Proof. Throughout the proof, $C_{\text{params}} > 0$ denotes a constant depending only on the indicated parameters. We condition on $X_1 = z$.

Notation. Define

$$\eta_k(z) = \exp(\lambda \langle z, \mu\rangle \langle X_k, \mu\rangle), \quad N_L(z) = \frac{1}{L-1} \sum_{k=2}^L \eta_k(z) X_k, \quad S_L(z) = \frac{1}{L-1} \sum_{k=2}^L \eta_k(z),$$

and

$$N = \mathbb{E}[N_L(z)], \quad S = \mathbb{E}[S_L(z)], \quad S_\theta = S + \theta(S_L - S), \quad N_\theta = N + \theta(N_L - N).$$

Let

$$\xi = \lambda^2 \langle z, \mu\rangle^2 \mu^\top \Sigma \mu.$$

General structure. All terms appearing below are finite sums of quantities of the form

$$T_k = \alpha_k \, \mathbb{E}\left[\prod_i |Z_{i,k}|^{p_{i,k}}\right],$$

where $Z_{i,k}$ belongs to

$$N_L - N, \; S_L - S, \; D_\mu N_L - D_\mu N, \; D_\mu S_L - D_\mu S, \; S_\theta^{-1}, \; N_\theta, \; D_\mu N_\theta, \; D_\mu S_\theta.$$

Using Lemmas 9 and 11,

$$T_k \le C_{\lambda,\Sigma} L^{-\frac{1}{2}\sum_i p_{i,k}} (1 + \|z\|^{m_k}) \exp\left(\frac{1}{2}\Big(\sum_i p_{i,k}\Big)^2 \xi\right).$$

We define

$$P := \max_k \sum_i p_{i,k}.$$

0-th order bound. We write

$$T_L(z) = \frac{N_L(z)}{S_L(z)} + R_{1,L}(z),$$

where

$$R_{1,L}(z) = \frac{\eta_1(z)}{L} \frac{zS_L(z) - N_L(z)}{S_L(z)\left(S_L(z) + \frac{\eta_1(z)}{L}\right)}.$$

Using Lemma 9 with $p = 2$,

$$\mathbb{E}[\|R_{1,L}(z)\|^2] \le \frac{C_{\lambda,\Sigma}}{L^2}(1 + \|z\|^2) \exp\left(\frac{2^2}{2}\xi\right).$$

Consider the map

$$F_1 : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d, \qquad F_1(x, y) = \frac{x}{y}.$$

Its differentials are

$$DF_1(x, y)[h, k] = \frac{h}{y} - \frac{xk}{y^2},$$

$$D^2F_1(x, y)[(h, k), (h, k)] = \frac{2x}{y^3}k^2 - \frac{2}{y^2}hk.$$

Applying Taylor's formula at (N, S), there exists $\theta \in (0, 1)$ such that

$$\frac{N_L}{S_L} - \frac{N}{S} = DF_1(N, S)[N_L - N, S_L - S] + R_{2,L},$$

where

$$R_{2,L} = \frac{1}{2}D^2F_1(N_\theta, S_\theta)[(N_L - N, S_L - S), (N_L - N, S_L - S)].$$

Thus

$$R_{2,L} = \frac{(N + \theta(N_L - N))(S_L - S)^2}{(S + \theta(S_L - S))^3} - \frac{(N_L - N)(S_L - S)}{(S + \theta(S_L - S))^2},$$

and

$$\|R_{2,L}\|^2 \le 2\frac{\|N_\theta\|^2(S_L - S)^4}{S_\theta^6} + 2\frac{\|N_L - N\|^2(S_L - S)^2}{S_\theta^4}.$$

Applying Lemmas 9 and 11, there exists $m_0 > 0$ such that

$$\mathbb{E}[\|R_{2,L}(z)\|^2] \le \frac{C_{\lambda,\Sigma}}{L^2}(1 + \|z\|^{m_0}) \exp\left(\frac{12^2}{2}\xi\right).$$

Origin of $P = 12$. The highest order product is

$$\|N_\theta\|^2(S_L - S)^4 S_\theta^{-6}, \quad \Rightarrow \quad P = 2 + 4 + 6 = 12.$$

Moreover,

$$\mathbb{E}\left[\left\|\frac{N_L - N}{S} - \frac{N(S_L - S)}{S^2}\right\|^2\right] = \frac{C_{\lambda,\Sigma}}{L}(1 + \|z\|^{m_0}) \exp\left(\frac{2^2}{2}\xi\right).$$

Thus

$$\mathbb{E}[\|T_L - T_\infty\|^2 \mid X_1 = z] \le \frac{C_{\lambda,\Sigma}}{L}(1 + \|z\|^{m_0}) \exp\left(\frac{12^2}{2}\xi\right).$$

Applying Lemma 4,

$$\mathbb{E}[\|T_L - T_\infty\|^2] = \mathcal{O}\big(g(L, \lambda, \mu, \Sigma, 12)\big).$$

1-st order bound. We write

$$D_\mu T_L - D_\mu T_\infty = \left(\frac{D_\mu N_L}{S_L} - \frac{D_\mu N}{S}\right) - \left(\frac{N_L (D_\mu S_L)^\top}{S_L^2} - \frac{N (D_\mu S)^\top}{S^2}\right) + \widetilde{R}_{1,L}(z),$$

where $\widetilde{R}_{1,L}(z)$ gathers all terms arising from differentiating the contribution of $\eta_1(z)$ in the numerator and denominator of T_L. .

Consider the map

$$F_2 : \mathbb{R}^{d\times d} \times \mathbb{R} \to \mathbb{R}^{d\times d}, \qquad F_2(A, y) = \frac{A}{y}.$$

Applying Taylor's formula at $(D_\mu N, S)$ yields an expansion of

$$\frac{D_\mu N_L}{S_L} - \frac{D_\mu N}{S}.$$

Similarly, define

$$F_3 : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{d\times d}, \qquad F_3(x, y, w) = \frac{xy^\top}{w^2},$$

and apply Taylor's formula at $(N, D_\mu S, S)$.

All resulting terms are finite sums of products handled above. Using Lemmas 9 and 11, there exists $m_1 > 0$ such that

$$\mathbb{E}[\|D_\mu T_L - D_\mu T_\infty\|^2 \mid X_1 = z] \le \frac{C_{\lambda,\Sigma}}{L}(1 + \|z\|^{m_1}) \exp\left(\frac{16^2}{2}\xi\right).$$

Origin of $P = 16$. The highest order term is

$$\|N_\theta\|^2 \|D_\mu S_\theta\|^2 (S_L - S)^4 S_\theta^{-8}, \quad \Rightarrow \quad P = 2 + 2 + 4 + 8 = 16.$$

Applying Lemma 4,

$$\mathbb{E}[\|D_\mu T_L - D_\mu T_\infty\|^2] = \mathcal{O}\big(g(L, \lambda, \mu, \Sigma, 16)\big).$$

2-nd order bound. We use the identity

$$D_\mu^2\left(\frac{N_L}{S_L}\right) = \frac{D_\mu^2 N_L}{S_L} - \frac{2(D_\mu N_L)(D_\mu S_L)^\top}{S_L^2} - \frac{N_L D_\mu^2 S_L}{S_L^2} + \frac{2N_L (D_\mu S_L)(D_\mu S_L)^\top}{S_L^3},$$

and the analogous expression for N/S.

Each difference is expanded using Taylor formulas for maps of the form

$$(A, y) \mapsto \frac{A}{y}, \qquad (x, y, w) \mapsto \frac{xy^\top}{w^2}, \qquad (x, y, z, w) \mapsto \frac{xyz^\top}{w^3}.$$

Using Lemmas 9 and 11, there exists $m_2 > 0$ such that

$$\mathbb{E}[\|D_\mu^2 T_L - D_\mu^2 T_\infty\|^2 \mid X_1 = z] \le \frac{C_{\lambda,\Sigma}}{L}(1 + \|z\|^{m_2}) \exp\left(\frac{20^2}{2}\xi\right).$$

Origin of $P = 20$. The highest order term is

$$\|N_\theta\|^2 \|D_\mu S_\theta\|^2 \|D_\mu S_\theta\|^2 (S_L - S)^4 S_\theta^{-10}, \quad \Rightarrow \quad P = 2 + 2 + 2 + 4 + 10 = 20.$$

Applying Lemma 4,

$$\mathbb{E}[\|D_\mu^2 T_L - D_\mu^2 T_\infty\|_F^2] = \mathcal{O}\big(g(L, \lambda, \mu, \Sigma, 20)\big).$$

□

Lemma 8. *Let $f : \mathbb{R}^d \to \mathbb{R}^{m_1 \times m_2}$ and $X_1, \dots, X_L$ be i.i.d. random variables $\mathcal{N}(0, \Sigma)$, let $A_L = \frac{1}{L}\sum_{k=1}^{L} f(X_k)$, then for $p \geq 2$, we have*

$$\mathbb{E}[\|A_L - \mathbb{E}[A_L]\|_F^p] \leq C_p L^{-\frac{p}{2}} \mathbb{E}[\|f(X_1)\|_F^p],$$

where C_p is a constant that only depends on p, and $\|\cdot\|_F$ is the Frobenius norm on $\mathbb{R}^{m_1 \times m_2}$.

Proof. Let $Y_k = f(X_k) - \mathbb{E}[f(X_1)]$. By construction, $\mathbb{E}[Y_k] = 0$ and Y_k are i.i.d., we have

$$\mathbb{E}[\|A_L - \mathbb{E}[A_L]\|^p] = \frac{1}{L^p}\mathbb{E}\left[\left\|\sum_{k=1}^{L} Y_k\right\|^p\right],$$

Rosenthal's inequality states that there exists a constant R_p depending only on p such that

$$\mathbb{E}\left[\left\|\sum_{k=1}^{L} Y_k\right\|^p\right] \leq R_p L^{\frac{p}{2}} \mathbb{E}[\|Y_1\|^p],$$

Besides by Jensen's inequality,

$$\mathbb{E}[\|Y_1\|^p] \leq 2^p \mathbb{E}[\|f(X_1)\|^p].$$

We conclude by taking $C_p = 2^p R_p$.

□

Lemma 9. *Let $\lambda > 0, z, \mu \in \mathbb{R}^d$, $\theta \in [0,1]$, $X, X_1, \dots, X_L$ i.i.d. $\mathcal{N}(0, \Sigma)$,*

$$\begin{aligned}
\eta_k(z) &= \exp(\lambda \langle z, \mu\rangle \langle X_k, \mu\rangle), \\
N_L(z) &= \frac{1}{L-1}\sum_{k=2}^{L} \eta_k(z) X_k, \\
N &= \mathbb{E}[N_L], \\
S_L(z) &= \frac{1}{L-1}\sum_{k=2}^{L} \eta_k(z), \\
S &= \mathbb{E}[S_L], \\
S_\theta &= S + \theta(S_L - S), \\
N_\theta &= N + \theta(N_L - N),
\end{aligned}$$

Then for $p > 0$, there exists $C_{p,\Sigma}, C_p > 0$ (depending only on the constants in the subscripts) such that letting $\xi = \lambda^2 \langle z, \mu\rangle^2 \mu^\top \Sigma \mu$ we have:

$$\begin{aligned}
\mathbb{E}[S_\theta^{-p}] &\leq 2\exp\left(\frac{p^2}{2}\xi\right), \\
\mathbb{E}[\|N_L - N\|^p] &\leq C_{p,\Sigma} L^{-\frac{p}{2}} (1 + \lambda^p \|\mu\|^{2p} \|z\|^p) \exp\left(\frac{p^2}{2}\xi\right), \\
\mathbb{E}[|S_L - S|^p] &\leq C_p L^{-\frac{p}{2}} \exp\left(\frac{p^2}{2}\xi\right), \\
\mathbb{E}[\|D_\mu N_L - D_\mu N\|^p] &\leq C_{p,\Sigma} L^{-\frac{p}{2}} \lambda^p \|\mu\|^p \|z\|^p (1 + \lambda^{2p} \|\mu\|^{4p} \|z\|^{2p}) \exp\left(\frac{p^2}{2}\xi\right), \\
\mathbb{E}[\|D_\mu S_L - D_\mu S\|^p] &\leq C_{p,\Sigma} L^{-\frac{p}{2}} \lambda^p \|\mu\|^p \|z\|^p (1 + \lambda^p \|\mu\|^{2p} \|z\|^p) \exp\left(\frac{p^2}{2}\xi\right).
\end{aligned}$$

Proof. By Jensen's inequality we have that, for $p > 0$: $S_\theta^{-p} \leq S^{-p} + S_L^{-p}$ and

$$S_L^{-p} \leq \frac{1}{L-1} \sum_{k=2}^{L} \exp(-p\lambda \langle z, \mu \rangle \langle X_k, \mu \rangle),$$

thus

$$\mathbb{E}[S_\theta^{-p}] \leq S^{-p} + \mathbb{E}[\exp(-p\lambda \langle z, \mu \rangle \langle X, \mu \rangle)] \leq 2\exp\left(\frac{1}{2} p^2 \lambda^2 \langle z, \mu \rangle^2 \mu^\top \Sigma \mu\right).$$

Using Lemma 8 with $f_1(x) = x\exp(\lambda \langle z, \mu \rangle \langle x, \mu \rangle)$ and $f_2(x) = \exp(\lambda \langle z, \mu \rangle \langle x, \mu \rangle)$ and Lemma 10, we get

$$\begin{aligned}
\mathbb{E}[\|N_L - N\|^p] &\leq C_p L^{-\frac{p}{2}} \mathbb{E}[\|f_1(X)\|^p] \\
&\leq C_{p,\Sigma} L^{-\frac{p}{2}} (1 + \|\mu\|^p \lambda^p |\langle z, \mu \rangle|^p) \exp\left(\frac{1}{2} p^2 \lambda^2 \langle z, \mu \rangle^2 \mu^\top \Sigma \mu\right) \\
&\leq C_{p,\Sigma} L^{-\frac{p}{2}} (1 + \lambda^p \|\mu\|^{2p} \|z\|^p) \exp\left(\frac{1}{2} p^2 \lambda^2 \langle z, \mu \rangle^2 \mu^\top \Sigma \mu\right),
\end{aligned}$$

and

$$\mathbb{E}[|S_L - S|^p] \leq C_p L^{-\frac{p}{2}} \mathbb{E}[\|f_2(X)\|^p] = C_p L^{-\frac{p}{2}} \exp\left(\frac{1}{2} p^2 \lambda^2 \langle z, \mu \rangle^2 \mu^\top \Sigma \mu\right).$$

Besides, with

$$f_3(x) = \lambda \exp(\lambda \langle z, \mu \rangle \langle x, \mu \rangle) x (\langle z, \mu \rangle x + \langle x, \mu \rangle z)^\top,$$

and

$$f_4(x) = \lambda \exp(\lambda \langle z, \mu \rangle \langle x, \mu \rangle)(\langle z, \mu \rangle x + \langle x, \mu \rangle z),$$

we derive the bounds

$$\begin{aligned}
\mathbb{E}[\|D_\mu N_L - D_\mu N\|^p] &\leq C_p L^{-\frac{p}{2}} \mathbb{E}[\|f_3(X)\|_F^p] \\
&\leq C_p L^{-\frac{p}{2}} C_{p,\Sigma} \lambda^p \|\mu\|^p \|z\|^p (1 + \lambda^{2p} \langle z, \mu \rangle^{2p} \|\mu\|^{2p}) \exp\left(\frac{1}{2} p^2 \lambda^2 \langle z, \mu \rangle^2 \mu^\top \Sigma \mu\right) \\
&\leq C_{p,\Sigma} L^{-\frac{p}{2}} \lambda^p \|\mu\|^p \|z\|^p (1 + \lambda^{2p} \|\mu\|^{4p} \|z\|^{2p}) \exp\left(\frac{1}{2} p^2 \lambda^2 \langle z, \mu \rangle^2 \mu^\top \Sigma \mu\right),
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\|D_\mu S_L - D_\mu S\|^p] &\leq C_p L^{-\frac{p}{2}} \mathbb{E}[\|f_4(X)\|^p] \\
&\leq C_p L^{-\frac{p}{2}} C_{p,\Sigma} \lambda^p \|\mu\|^p \|z\|^p (1 + \lambda^p \|\mu\|^{2p} \|z\|^p) \exp\left(\frac{1}{2} p^2 \lambda^2 \langle z, \mu \rangle^2 \mu^\top \Sigma \mu\right) \\
&\leq C_{p,\Sigma} L^{-\frac{p}{2}} \lambda^p \|\mu\|^p \|z\|^p (1 + \lambda^p \|\mu\|^{2p} \|z\|^p) \exp\left(\frac{1}{2} p^2 \lambda^2 \langle z, \mu \rangle^2 \mu^\top \Sigma \mu\right).
\end{aligned}$$

Furthermore, using Lemma 8 with

$$f_5(x) = x\exp(\lambda \langle z, \mu \rangle \langle x, \mu \rangle)[\lambda^2 (\langle z, h \rangle \langle x, \mu \rangle + \langle z, \mu \rangle \langle x, h \rangle)^2 + 2\lambda \langle z, h \rangle \langle x, h \rangle]^\top,$$

and

$$f_6(x) = \exp(\lambda \langle z, \mu \rangle \langle x, \mu \rangle)[\lambda^2 (\langle z, h \rangle \langle x, \mu \rangle + \langle z, \mu \rangle \langle x, h \rangle)^2 + 2\lambda \langle z, h \rangle \langle x, h \rangle],$$

and Lemma 10, we get

$$\begin{aligned}
\mathbb{E}[\|D_\mu^2 N_L[h] - D_\mu^2 N[h]\|_F^p] &\leq C_p L^{-\frac{p}{2}} \mathbb{E}[\|f_5(X_1)\|_F^p] \\
&\leq C_{p,\Sigma} \lambda^{2p} \|z\|^p \|h\|^p (1 + \lambda^{2p} \langle z, \mu \rangle^{2p} \|\mu\|^{2p}) \exp\left(\frac{p^2}{2} \lambda^2 \langle z, \mu \rangle^2 \mu^\top \Sigma \mu\right).
\end{aligned}$$

and

$$\mathbb{E}[\|D_\mu^2 S_L[h] - D_\mu^2 S[h]\|^p] \leq C_p L^{-\frac{p}{2}} \mathbb{E}[\|f_6(X_1)\|^p]$$
$$\leq C_{p,\Sigma} \lambda^p \|z\|^p \|h\|^p (1 + \lambda^p \langle z, \mu \rangle^p \|\mu\|^p) \exp\left(\frac{p^2}{2} \lambda^2 \langle z, \mu \rangle^2 \mu^\top \Sigma \mu\right).$$

□

Lemma 10. *Consider $\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}$ symmetric and positive definite, $p, q > 0$ and $X \sim \mathcal{N}(0, \Sigma)$, then*

$$\mathbb{E}[\|X\|^p \exp(q\langle X, \mu \rangle)] \leq C_{p,\Sigma}(1 + q^p \|\mu\|^p) \exp\left(\frac{q^2}{2} \mu^\top \Sigma \mu\right).$$

Proof. We have that

$$\mathbb{E}[\|X\|^p \exp(q\langle X, \mu \rangle)] = \exp\left(\frac{q^2}{2} \mu^\top \Sigma \mu\right) \mathbb{E}[\|X + q\Sigma\mu\|^p].$$

And we bound

$$\|X + q\Sigma\mu\|^p \leq 2^{p-1}(\|X\|^p + \|\Sigma\|_{op}^p q^p \|\mu\|^p).$$

So there exists $C_{p,\Sigma} \stackrel{\text{def}}{=} 2^{p-1} \max\{\mathbb{E}[\|X\|^p], \|\Sigma\|_{op}^p\} > 0$ such that

$$\mathbb{E}[\|X + q\Sigma\mu\|^p] \leq C_{p,\Sigma}(1 + q^p \|\mu\|^p).$$

And we conclude.

□

Lemma 11. *Let $Z_1, \ldots, Z_n$ be non-negative random variables such that for each $i \in \{1, \ldots, n\}$ there exists $C_{i,p}$ that grows at most exponentially in p such that,*

$$\mathbb{E}[|Z_i|^p] \leq C_{i,p} \exp\left(\frac{p^2}{2} \xi\right),$$

for some $\xi > 0$. Let $p_1, \ldots, p_n > 0$. Then there exists C that grows at most exponentially in $(p_1, \ldots, p_n)$ such that

$$\mathbb{E}\left[\prod_{i=1}^n |Z_i|^{p_i}\right] \leq C \exp\left(\frac{1}{2}\left(\sum_{i=1}^n p_i\right)^2 \xi\right)$$

Proof. We apply the generalized Hölder's inequality to get that for every $q_1, \ldots, q_n > 0$ such that $\sum_{i=1}^n \frac{1}{q_i} = 1$,

$$\mathbb{E}\left[\prod_{i=1}^n |Z_i|^{p_i}\right] \leq \prod_{i=1}^n \mathbb{E}[|Z_i|^{p_i q_i}]^{\frac{1}{q_i}}$$
$$\leq \prod_{i=1}^n \left[C_{i,p_i q_i} \exp\left(\frac{p_i^2 q_i^2}{2} \xi\right)\right]^{\frac{1}{q_i}}$$
$$= \left(\prod_{i=1}^n (C_{i,p_i q_i})^{\frac{1}{q_i}}\right) \exp\left(\frac{1}{2} \sum_{i=1}^n \left(p_i^2 q_i\right) \xi\right).$$

Choosing $q_i = \frac{\sum_{j=1}^n p_j}{p_i}$, which minimizes $\sum_{i=1}^n p_i^2 q_i$ given $\sum_{i=1}^n \frac{1}{q_i} = 1$, we conclude since $C_{i, \sum_{j=1}^n p_j}^{\frac{p_i}{\sum_{j=1}^n p_j}}$ grows exponentially in p_i and $C = \prod_{i=1}^n C_{i, \sum_{j=1}^n p_j}^{\frac{p_i}{\sum_{j=1}^n p_j}}$ grows exponentially in $(p_1, \ldots, p_n)$. □

C.4 ICL technical propositions

Lemma 12. *Let $\Sigma \sim W_d(V, n)$, and let $\mu \in \mathbb{R}^d$ be deterministic. Then:*

1. $\mathbb{E}[\operatorname{tr} \Sigma] = n \operatorname{tr} V$.
2. $\mathbb{E}[\mu^\top \Sigma^2 \mu] = n(n+1)\, \mu^\top V^2 \mu + n\, \operatorname{tr}(V)\, \mu^\top V \mu$.
3. $\mathbb{E}[(\mu^\top \Sigma^2 \mu)(\mu^\top \Sigma^2 \mu)] = n(n+2)[(n+3)(\mu^\top V \mu)(\mu^\top V^2 \mu) + (\mu^\top V \mu)^2 \operatorname{tr}(V)]$.

Proof. We use the standard Gaussian representation of the Wishart distribution:

$$\Sigma = \sum_{r=1}^{n} x_r x_r^\top, \qquad x_r \overset{iid}{\sim} \mathcal{N}(0, V).$$

1. Since $\operatorname{tr}(x_r x_r^\top) = \|x_r\|^2$ and $\mathbb{E}[\|x_r\|^2] = \operatorname{tr} V$, linearity of expectation gives

$$\mathbb{E}[\operatorname{tr} \Sigma] = \sum_{r=1}^{n} \mathbb{E}[\|x_r\|^2] = n \operatorname{tr} V.$$

2. Write

$$\mu^\top \Sigma^2 \mu = \sum_{i,j=1}^{n} (\mu^\top \alpha_i)(\alpha_i^\top x_j)(\mu^\top x_j).$$

Splitting the sum into the cases $i = j$ and $i \neq j$:

(i) Diagonal terms. For $i = j$,

$$\mathbb{E}\big[(\mu^\top x)^2 (x^\top x)\big] = \mu^\top V \mu \operatorname{tr} V + 2\, \mu^\top V^2 \mu,$$

by Isserlis' formula.

(ii) Off-diagonal terms. For $i \neq j$, independence yields

$$\mathbb{E}[(\mu^\top \alpha_i)(\alpha_i^\top x_j)(\mu^\top x_j)] = \mu^\top V^2 \mu.$$

Counting terms,

$$\mathbb{E}[\mu^\top \Sigma^2 \mu] = n\big(\mu^\top V \mu \operatorname{tr} V + 2 \mu^\top V^2 \mu\big) + n(n-1) \mu^\top V^2 \mu,$$

which simplifies to

$$\mathbb{E}[\mu^\top \Sigma^2 \mu] = n(n+1) \mu^\top V^2 \mu + n \operatorname{tr} V\, \mu^\top V \mu.$$

3. Let $s_r = \mu^\top x_r$. Then

$$\mu^\top \Sigma \mu = \sum_k s_k^2, \qquad \mu^\top \Sigma^2 \mu = \sum_{i,j} s_i (\alpha_i^\top x_j) s_j.$$

Hence

$$\mathbb{E}[(\mu^\top \Sigma \mu)(\mu^\top \Sigma^2 \mu)] = \sum_{i,j,k} \mathbb{E}\big[s_k^2 s_i (\alpha_i^\top x_j) s_j\big].$$

The expectation depends on coincidences among the indices (i, j, k). Using Isserlis' theorem and independence, one obtains:

- i, j, k all distinct: contribution $(\mu^\top V \mu)(\mu^\top V^2 \mu)$.
- $i = j \neq k$: contribution $(\mu^\top V \mu)\big(\mu^\top V \mu \operatorname{tr} V + 2 \mu^\top V^2 \mu\big)$.
- $i = k \neq j$ or $a = j \neq i$: contribution $3(\mu^\top V \mu)(\mu^\top V^2 \mu)$.
- $i = j = k$: contribution $4(\mu^\top V \mu)(\mu^\top V^2 \mu) + (\mu^\top V \mu)^2 \operatorname{tr} V$.

Summing all contributions with their combinatorial counts yields

$$\mathbb{E}[(\mu^\top \Sigma\mu)(\mu^\top \Sigma^2\mu)] = n(n+2)\big[(n+3)(\mu^\top V\mu)(\mu^\top V^2\mu) + (\mu^\top V\mu)^2 \operatorname{tr} V\big] .$$

This concludes the proof.

□

Lemma 13. *Let $\Sigma \sim W_d(V,n)$, then $\mathcal{R}_\infty^{\mathrm{ICL}}(\mu) = \mathbb{E}_{\Sigma\sim W_d(V,n)}[\mathcal{R}_{\mathrm{soft},\infty}^{(\Sigma)}(\mu)]$, and*

$$\begin{aligned}\mathcal{R}_\infty^{\mathrm{ICL}}(\mu) = n\operatorname{tr}(V) - 2\lambda n[(n+1)\,\mu^\top V^2\mu + \operatorname{tr}(V)\,\mu^\top V\mu] \\ + \lambda^2 n(n+2)[(n+3)(\mu^\top V\mu)(\mu^\top V^2\mu) + (\mu^\top V\mu)^2 \operatorname{tr}(V)].\end{aligned}$$

In particular, if $V = \xi^2 I_d + \theta vv^t$ for $\|v\| = 1, \xi > 0, \theta > 0$, and we let $\alpha = \langle \mu, v\rangle$ and $r = \|\mu\|$, we have that there exists $\tilde{\mathcal{R}}^{\mathrm{ICL}} : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ such that

$$\mathcal{R}_\infty^{\mathrm{ICL}}(\mu) = \tilde{\mathcal{R}}^{\mathrm{ICL}}(r^2, \alpha^2),$$

where

$$\begin{aligned}\tilde{\mathcal{R}}^{\mathrm{ICL}}(r^2,\alpha^2) = n\big(d\xi^2+\theta\big) - 2\lambda n\Big[(n+1)\Big(\xi^4 r^2 + (2\xi^2\theta+\theta^2)\alpha^2\Big) + \big(d\xi^2+\theta\big)\Big(\xi^2 r^2 + \theta\alpha^2\Big)\Big] \\ + \lambda^2 n(n+2)\Bigg\{(n+3)\Big(\xi^2 r^2 + \theta\alpha^2\Big)\Big(\xi^4 r^2 + (2\xi^2\theta+\theta^2)\alpha^2\Big) \\ + \big(d\xi^2+\theta\big)\Big(\xi^2 r^2 + \theta\alpha^2\Big)^2\Bigg\}.\end{aligned}$$

Furthermore, for this particular V, the gradient satisfies

$$\nabla\mathcal{R}_\infty^{\mathrm{ICL}}(\mu) = 2\big(A(r,\alpha)\mu + B(r,\alpha)\alpha v\big),$$

where

$$A(r,\alpha) = a_1 r^2 + a_2\alpha^2 - a_3, \qquad B(r,\alpha) = b_1 r^2 + b_2\alpha^2 - b_3,$$

for constants $a_i, b_i > 0$ defined as

$$\begin{cases} a_1 &= 2\lambda^2 n(n+2)\xi^4[(n+d+3)\xi^2 + \theta], \\ a_2 &= \lambda^2 n(n+2)\xi^2\theta[(3n+2d+9)\xi^2 + (n+5)\theta], \\ a_3 &= 2\lambda n\xi^2[(n+d+1)\xi^2 + \theta], \\ b_1 &= a_2, \\ b_2 &= 2\lambda^2 n(n+2)\theta^2[(2n+d+6)\xi^2 + (n+4)\theta], \\ b_3 &= 2\lambda n\theta[(2n+d+2)\xi^2 + (n+2)\theta]. \end{cases} \tag{18}$$

Proof. The expression of $\mathcal{R}_\infty^{\mathrm{ICL}}(\mu)$ follows directly from (4) together with Lemma 12.

In the case where $V = \xi^2 I_d + \theta vv^\top$, expanding the quadratic forms $\mu^\top V\mu$ and $\mu^\top V^2\mu$ in terms of $r^2 = \|\mu\|^2$ and $\alpha = \langle\mu, v\rangle$ yields the representation

$$\mathcal{R}_\infty^{\mathrm{ICL}}(\mu) = \tilde{\mathcal{R}}^{\mathrm{ICL}}(r^2, \alpha^2).$$

Differentiating this expression with respect to μ, using $\nabla r^2 = 2\mu$ and $\nabla\alpha^2 = 2\alpha v$, gives the stated gradient form

$$\nabla\mathcal{R}_\infty^{\mathrm{ICL}}(\mu) = 2\big(A(r,\alpha)\mu + B(r,\alpha)\alpha v\big),$$

where $A = \nabla_{r^2}\tilde{\mathcal{R}}^{\mathrm{ICL}}$ and $B = \nabla_{\alpha^2}\tilde{\mathcal{R}}^{\mathrm{ICL}}$ are polynomials in (r^2, α^2) whose coefficients are obtained by explicit identification. The expressions of a_i, b_i follow from direct computation. □

Lemma 14. *Let $\xi^2, \theta, \lambda > 0$ and $n \geq d \geq 1$, and a_i, b_i defined as in (18), then $a_1 b_3 > a_2 a_3$ and $(a_1 + a_2) b_3 > (a_2 + b_2) a_3$.*

Proof. Developing the terms, we have that

$$a_1 b_3 - a_2 a_3 = 2\lambda^3 n^2 [c_1 \theta \xi^8 + c_2 \theta^2 \xi^6 + c_3 \theta^3 \xi^4],$$
$$(a_1 + a_2) b_3 - (a_2 + b_2) a_3 = 2\lambda^3 n^2 (n+2) [c_4 \theta \xi^8 + c_5 \theta^2 \xi^6 + c_6 \theta^3 \xi^4 + c_7 \theta^4 \xi^2],$$

with

$$\begin{aligned}
c_1 &= (n+2)(d(n-1) + n^2 + 4n + 3), \\
c_2 &= (n+2)(d(n-1) + n^2 + 5n + 2), \\
c_3 &= (n-1)(n+2), \\
c_4 &= d(n-1) + n^2 + 4n + 3, \\
c_5 &= (2d(n-1) + 3n^2 + 13 + 8), \\
c_6 &= d(n-1) + 3n^2 + 14n + 7, \\
c_7 &= n^2 + 5n + 2.
\end{aligned}$$

Since $n \geq d \geq 1$, we have that every constant $c_1, \ldots, c_7$ is positive, concluding the lemma. □

Proposition 18 (Families of stationary points)**.** *Under the notation of Lemma 13 and the parametrization $\mu = \alpha v + w$ with $w \perp v$ and $r = \|\mu\|$, all stationary points of $\mathcal{R}_\infty^{\mathrm{ICL}}$ belong to one of the following families:*

1. ***Trivial:*** *$\mu = 0$.*
2. ***Orthogonal:*** *$\alpha = 0$, $w \neq 0$, with $A(r, 0) = 0$.*
3. ***Aligned:*** *$w = 0$, $\alpha \neq 0$, so that $\mu = \alpha v$, with*
$$A(r, \alpha) + B(r, \alpha) = 0.$$
4. ***Off-axis:*** *$\alpha \neq 0$, $w \neq 0$, with*
$$A(r, \alpha) = 0 \quad \text{and} \quad B(r, \alpha) = 0.$$

Proof. A stationary point satisfies

$$A(r, \alpha)\mu + B(r, \alpha)\alpha v = 0.$$

Writing $\mu = \alpha v + w$ with $w \perp v$ and projecting onto $\mathrm{span}(v)$ and its orthogonal complement yields

$$A(r, \alpha) w = 0, \qquad \alpha\big(A(r, \alpha) + B(r, \alpha)\big) = 0.$$

The conclusion follows by considering whether $\alpha = 0$ or not and whether $w = 0$ or not. □

Proposition 19 (Characterization of families)**.** *The function $\mathcal{R}_\infty^{\mathrm{ICL}} : \mathbb{R}^d \to \mathbb{R}$ satisfies:*

1. *$\mu = 0$ is a local maximum: the Hessian has only negative eigenvalues along all nonzero directions.*
2. *Any admissible orthogonal point is a strict saddle: the Hessian has exactly one negative eigenvalue along v, one positive eigenvalue along the vector's own direction and $d - 2$ zero eigenvalues.*
3. *Any admissible aligned point is a strict local minimum. Moreover, the only admissible aligned points are $\mu^\star = \pm\alpha^\star v$, for $\alpha^\star > 0$ defined in (22).*
4. *There are no admissible off-axis solutions.*

Since $\mathcal{R}_\infty^{\mathrm{ICL}}$ is coercive, these two points are the global minimizers of the function. Finally, for almost every initialization $\mu_0 \in \mathbb{R}^d$, the gradient flow of $\mathcal{R}_\infty^{\mathrm{ICL}}$ converges to one of the two global minimizers $\pm\alpha^\star v$.

Proof. We recall by Lemma 13 that $\nabla \mathcal{R}_\infty^{\mathrm{ICL}}(\mu) = 2(A(r,\alpha)\mu + B(r,\alpha)\alpha v)$, where

$$A(r,\alpha) = a_1 r^2 + a_2 \alpha^2 - a_3, \quad B(r,\alpha) = b_1 r^2 + b_2 \alpha^2 - b_3,$$

for $a_i, b_i, i = \{1,2,3\}$ defined as

$$\begin{cases} a_1 &= 2\lambda^2 n(n+2)\xi^4[(n+d+3)\xi^2 + \theta], \\ a_2 &= \lambda^2 n(n+2)\xi^2\theta[(3n+2d+9)\xi^2 + (n+5)\theta], \\ a_3 &= 2\lambda n\xi^2[(n+d+1)\xi^2 + \theta], \\ b_1 &= a_2, \\ b_2 &= 2\lambda^2 n(n+2)\theta^2[(2n+d+6)\xi^2 + (n+4)\theta], \\ b_3 &= 2\lambda n\theta[(2n+d+2)\xi^2 + (n+2)\theta]. \end{cases} \tag{19}$$

By Lemma 14, we have that

$$a_1 b_3 > a_2 a_3, \tag{20}$$
$$(a_1 + a_2)b_3 > (a_2 + b_2)a_3, \tag{21}$$

holds for any $\lambda, \xi^2, \theta > 0$ and $n \geq d \geq 1$.

Differentiating the gradient we obtain the following Hessian:

$$\begin{aligned} \nabla^2 \mathcal{R}_\infty^{\mathrm{ICL}}(\mu) &= 2AI_d + 2Bvv^\top + 4(\partial_{r^2} A)\mu\mu^\top + 4\alpha^2(\partial_{\alpha^2} B)vv^\top + 4\alpha(\partial_{\alpha^2} A)(\mu v^\top + v\mu^\top). \\ &= 2AI_d + 2Bvv^\top + 4a_1\mu\mu^\top + 4\alpha^2 b_2 vv^\top + 4\alpha a_2(\mu v^\top + v\mu^\top). \end{aligned}$$

We check that:

- Origin: We have that $\nabla^2 \mathcal{R}_\infty^{\mathrm{ICL}}(0) = 2A(0,0)I_d + 2B(0,0)vv^\top = -2a_3 - 2b_3 vv^\top$, since $a_3, b_3 > 0$, all eigenvalues are strictly negative.
- Orthogonal: Here $r^2 = \frac{a_3}{a_1}$ and the Hessian simplifies to $\nabla^2 \mathcal{R}_\infty^{\mathrm{ICL}}(\mu) = 4a_1\mu\mu^\top + 2B(r,0)vv^\top$. In direction μ the associated eigenvalue is $4a_1 r^2 > 0$ and in direction v, is $2B(r,0) = 2(b_1\frac{a_3}{a_1} - b_3) = 2(a_2\frac{a_3}{a_1} - b_3)$, which is negative due to (20). In any other direction $w \in \{\mu, v\}^\perp$, we will have null eigenvalues, in particular $\dim(\{\mu, v\}^\perp) = d - 2$.
- Aligned: Here $\mu = \alpha v$, so $r^2 = \alpha^2 = \frac{a_3+b_3}{a_1+a_2+b_1+b_2} = \frac{a_3+b_3}{a_1+2a_2+b_2}$, if

  $$\alpha^\star = \sqrt{\frac{a_3 + b_3}{a_1 + 2a_2 + b_2}}, \tag{22}$$

  then $\mu = \pm\alpha^\star v$ and the Hessian becomes $\nabla^2 \mathcal{R}_\infty^{\mathrm{ICL}}(\mu) = 2A(I - vv^\top) + 4(a_3 + b_3)vv^\top$, in direction v the associated eigenvalue is $4(a_3 + b_3) > 0$. In perpendicular directions to v, the associated eigenvalue is

  $$2A(r,\alpha) = 2[(a_1 + a_2)\frac{a_3 + b_3}{a_1 + 2a_2 + b_2} - a_3],$$

  which is positive due to (21).
- Off-axis: This solution exists only if $a_1 r^2 + a_2\alpha^2 - a_3 = 0$ and $a_2 r^2 + b_2\alpha^2 - b_3 = 0$, and $r^2 > \alpha^2 > 0$, we solve the system of equations

  $$\begin{pmatrix} a_1 & a_2 \\ a_2 & b_2 \end{pmatrix} \begin{pmatrix} r^2 \\ \alpha^2 \end{pmatrix} = \begin{pmatrix} a_3 \\ b_3 \end{pmatrix}.$$

  Let $\Delta = a_1 b_2 - a_2^2$, if:

  - $\Delta = 0$, the matrix is singular and due to (20), we have that the system has no solution.

- $\Delta < 0$, the solution $\alpha^2 = \frac{a_1 b_3 - a_2 a_3}{\Delta}$ is negative due to (20), and we have a contradiction with the fact that $\alpha^2 \geq 0$.
- $\Delta > 0$, solving the system yields $r^2 - \alpha^2 = \frac{(a_2+b_2)a_3 - (a_1+a_2)b_3}{\Delta}$, and by (21), then $r^2 - \alpha^2 < 0$, which is a contradiction.

Thus, under our two conditions (20)-(21), no off-axis solution exists.

□

Proposition 20. *Let $\Sigma \sim W_d(V, n)$ with $V = \xi^2 I_d + \theta v v^t$ and $\|v\|=1$. Then, for L large enough, we can characterize the landscape of critical points of $\mathcal{R}_L^{\mathrm{ICL}}$ within a compact region. More precisely, for sufficiently large $\rho > 0$, there exists $L_0 \in \mathbb{N}$, such that for $L \geq L_0$, the set of critical points of $\mathcal{R}_L^{\mathrm{ICL}}$ that lies in $B(0, \rho)$ is contained in the union of the following sets:*

$$\mathrm{crit}(\mathcal{R}_L^{\mathrm{ICL}}) \cap B(0, \rho) \subseteq \{\mu_{L,0}^\star\} \cup \{\pm \mu_{L,\|}^\star\} \cup \{U_j : j = 1, \dots, 2(d-1)\},$$

where:

1. *The point $\mu_{L,0}^\star$ satisfies $\mu_{L,0}^\star \to 0$ as $L \to \infty$. This point is a strict local maximum.*
2. *The points $\pm\mu_{L,\|}^\star$ satisfy $\pm\mu_{L,\|}^\star \to \pm\alpha^\star v$ as $L \to \infty$, with $\alpha^\star$ according to (22). These points are local minimizers.*
3. *For each orthogonal critical point $\mu_\perp^{(j)}$, $j = 1, \dots, 2(d-1)$ of $\mathcal{R}_\infty^{\mathrm{ICL}}$, there exists a neighborhood U_j such that every critical point of $\mathcal{R}_L^{\mathrm{ICL}}$ in U_j is a strict saddle.*

Proof. The first two items 1 and 2 follow directly from Proposition 5. It remains to prove 3. Let

$$\mathrm{crit}(\mathcal{R}_\infty^{\mathrm{ICL}}) = \{0, \pm\alpha^\star v, \mu_\perp^{(1)}, \dots, \mu_\perp^{(2(d-1))}\},$$

where $\mu_\perp^{(j)}$ denote the orthogonal critical points of the limiting objective. The points 0 and $\pm\alpha^\star v$ are nondegenerate, while each $\mu_\perp^{(j)}$ is degenerate as it has $(d-2)$ null eigenvalues. Choose $r_0, r_\| > 0$ such that the balls

$$B(0, r_0), \qquad B(\pm\alpha^\star v, r_\|)$$

contain no other critical point of $\mathcal{R}_\infty^{\mathrm{ICL}}$. By Proposition 5, for L sufficiently large there exist unique critical points

$$\mu_{L,0}^\star \in B(0, r_0), \qquad \mu_{L,\|}^\star \in B(\alpha^\star v, r_\|), \qquad -\mu_{L,\|}^\star \in B(-\alpha^\star v, r_\|).$$

Next, for each orthogonal critical point $\mu_\perp^{(j)}$, choose $r_j > 0$ such that $B(\mu_\perp^{(j)}, r_j)$ contains no other critical point of $\mathcal{R}_\infty^{\mathrm{ICL}}$. Let $\rho > 0$ big enough such that

$$K := B(0, r_0) \cup B(\alpha^\star v, r_\|) \cup B(-\alpha^\star v, r_\|) \cup \bigcup_{j=1}^{2(d-1)} B(\mu_\perp^{(j)}, r_j) \subset B(0, \rho).$$

Since $\nabla\mathcal{R}_\infty^{\mathrm{ICL}}$ is continuous and has no zero on K^c, for every $\rho > 0$ there exists

$$\eta_\rho := \inf_{x \in K^c \cap B(0,\rho)} \|\nabla\mathcal{R}_\infty^{\mathrm{ICL}}(x)\| > 0.$$

Because $\mathcal{R}_L^{\mathrm{ICL}} \to \mathcal{R}_\infty^{\mathrm{ICL}}$ in C^1_{loc}, we have

$$\sup_{x \in B(0,\rho)} \|\nabla\mathcal{R}_L^{\mathrm{ICL}}(x) - \nabla\mathcal{R}_\infty^{\mathrm{ICL}}(x)\| \to 0.$$

Therefore $\mathcal{R}_L^{\mathrm{ICL}}$ has no critical point in $K^c \cap B(0, \rho)$, and then $\mathrm{crit}(\mathcal{R}_L^{\mathrm{ICL}}) \cap B(0, \rho) \subseteq K$.

Finally, fix one orthogonal critical point $\mu_\perp^{(j)}$ and consider the ball $B(\mu_\perp^{(j)}, r_j)$. By Proposition 19, the Hessian $\nabla^2\mathcal{R}_\infty^{\mathrm{ICL}}(\mu_\perp^{(j)})$ has one positive eigenvalue $\lambda_+ > 0$ and one negative eigenvalue $\lambda_- < 0$. Let

$$\gamma := \min\{\lambda_+, -\lambda_-\} > 0.$$

Since $\nabla^2 \mathcal{R}_\infty^{\mathrm{ICL}}$ is continuous, there exists $r_j > 0$ (possibly smaller than before) such that for every $\mu \in B(\mu_\perp^{(j)}, r_j)$ the matrix $\nabla^2 \mathcal{R}_\infty^{\mathrm{ICL}}(\mu)$ has an eigenvalue at least $\gamma/2$ and another at most $-\gamma/2$.

Because $\mathcal{R}_L^{\mathrm{ICL}} \to \mathcal{R}_\infty^{\mathrm{ICL}}$ in C^2_{loc}, we have

$$\sup_{\mu \in B(\mu_\perp^{(j)}, r_j)} \left\| \nabla^2 \mathcal{R}_L^{\mathrm{ICL}}(\mu) - \nabla^2 \mathcal{R}_\infty^{\mathrm{ICL}}(\mu) \right\| \to 0.$$

Hence for L sufficiently large and every $\mu \in B(\mu_\perp^{(j)}, r_j)$,

$$\left\| \nabla^2 \mathcal{R}_L^{\mathrm{ICL}}(\mu) - \nabla^2 \mathcal{R}_\infty^{\mathrm{ICL}}(\mu) \right\| \le \frac{\gamma}{4}.$$

By continuity of the eigenvalues (Weyl's inequality), the Hessian $\nabla^2 \mathcal{R}_L^{\mathrm{ICL}}(\mu)$ has one eigenvalue at least $\gamma/4$ and another at most $-\gamma/4$ for all $\mu \in B(\mu_\perp^{(j)}, r_j)$.

Let μ_L be a critical point of $\mathcal{R}_L^{\mathrm{ICL}}$ in $B(\mu_\perp^{(j)}, r_j)$. Then $\nabla \mathcal{R}_L^{\mathrm{ICL}}(\mu_L) = 0$, and the Hessian at this point has both a positive and a negative eigenvalue. Therefore μ_L is a strict saddle. □

D Numerical experiments

In this section, we present numerical experiments that illustrate and empirically validate the theoretical results developed throughout the paper. In particular, we study the convergence behavior of the different models toward the principal eigenvector of the underlying covariance structure, as well as the effect of key parameters such as the prompt length and the ambient dimension.

D.1 Experimental Setup

We consider a covariance matrix of the form

$$\Sigma = AA^\top + 0.1 I_d,$$

where $A \in \mathbb{R}^{d \times d}$ has i.i.d. standard Gaussian entries. The target direction is given by the principal eigenvector u_1 associated with the largest eigenvalue of Σ.

All experiments are conducted using stochastic gradient descent (SGD) with constant step size. More precisely, let $\mathcal{R}(\mu)$ denote the objective function of interest (either the empirical risk or the population risk, depending on the setting). The iterates $\{\mu_k\}_{k \ge 0}$ are defined by

$$\mu_{k+1} = \mu_k - \gamma\, g_k,$$

where $\gamma > 0$ is a constant learning rate and g_k is a stochastic gradient estimator of $\nabla \mathcal{R}(\mu_k)$.

In the finite-prompt setting, g_k is computed from a random batch of samples. For instance, in the softmax model, we draw $X_1^{(k)}, \dots, X_L^{(k)} \sim \mathcal{N}(0, \Sigma)$ and define

$$g_k = \nabla_\mu \widehat{\mathcal{R}}_L(\mu_k; X_1^{(k)}, \dots, X_L^{(k)}),$$

where $\widehat{\mathcal{R}}_L$ is the empirical risk associated with the sampled prompt. More precisely, at each iteration we draw a batch of size B, consisting of independent prompts $(X_{b,1}^{(k)}, \dots, X_{b,L}^{(k)})_{b=1}^B$ with $X_{b,j}^{(k)} \sim \mathcal{N}(0, \Sigma)$, and define

$$\widehat{\mathcal{R}}_L(\mu) = \frac{1}{B} \sum_{b=1}^{B} \left\| X_{b,1}^{(k)} - T_L^\mu(X_{b,1}^{(k)}, \dots, X_{b,L}^{(k)}) \right\|^2,$$

so that

$$g_k = \nabla_\mu \widehat{\mathcal{R}}_L(\mu_k).$$

In the infinite-prompt setting, we treat Σ as known (or equivalently $V = \xi^2 I_d + vv^\top$ in the spiked Wishart model), and we perform deterministic gradient descent, i.e., the gradient is computed directly from the population objective:

$$g_k = \nabla \mathcal{R}(\mu_k).$$

The initialization μ_0 is sampled uniformly from $\mathbb{S}^{d-1}$. Unless otherwise specified, we use the following parameters:

- Learning rate: $\gamma = 10^{-4}$,
- Batch size: $B = 256$,
- Number of independent runs: 10.

At each iteration k, we assess performance via the alignment between the normalized iterate and the target direction. In the standard setting, this is given by

$$\left| \left\langle \frac{\mu_k}{\|\mu_k\|}, u_1 \right\rangle \right|,$$

where u_1 denotes the principal eigenvector of Σ.

In the spiked Wishart setting, where $\Sigma \sim W_d(\xi^2 I_d + vv^\top)$, we instead measure alignment with the spike direction:

$$\left| \left\langle \frac{\mu_k}{\|\mu_k\|}, v \right\rangle \right|.$$

Remark 3. Gradient computations in the numerical experiments were carried out using JAX (Bradbury et al., 2018).

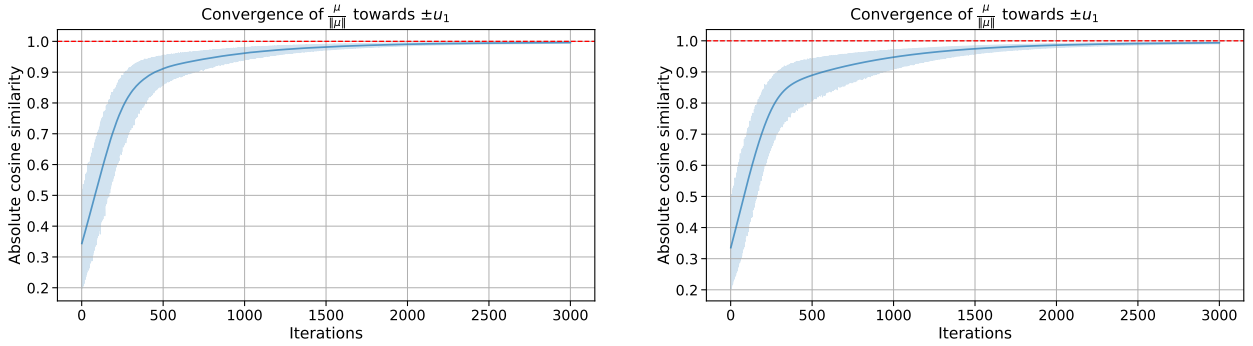
D.2 Softmax Attention: Finite and Infinite Prompt

We first study the softmax attention model in both finite and infinite prompt regimes.

In the finite prompt setting, at each iteration we sample $X_1, \ldots, X_L \sim \mathcal{N}(0, \Sigma)$ and perform stochastic gradient updates using the empirical risk of (2). In the infinite prompt setting, we instead optimize on the closed from of the population risk (4), which corresponds to the limit as $L \to \infty$.

Figure 3a shows the convergence behavior in the finite prompt case, with $L = 100$. Figure 3b shows the corresponding infinite prompt dynamics. Both figures superposed are shown in Figure 1.

We observe that in both regimes the iterates converge toward the principal eigenvector. Moreover, the infinite prompt setting exhibits smoother and more stable convergence, as it removes sampling noise. These observations are consistent with the theoretical analysis and illustrate how the finite prompt model approximates the infinite prompt limit.



(a) SGD on finite prompt risk, $L = 100$.

(b) GD on infinite prompt risk.

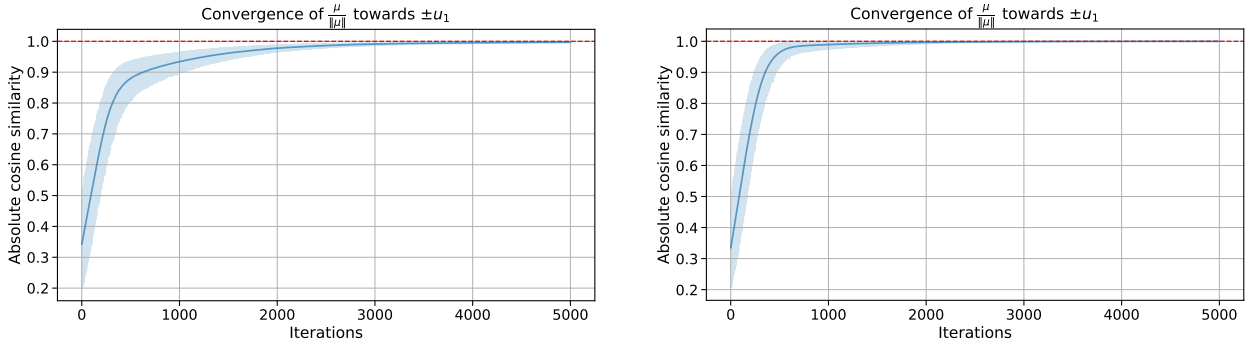
Figure 3: Softmax attention: finite vs. infinite prompt regimes.

D.3 Linear Attention

We next consider the linear attention model introduced in Section A. This model replaces the softmax weighting with a linear aggregation rule, leading to a simpler objective.

As illustrated in Figure 4, we consider the setting $d = 5$, $L = 6$, and $\lambda = 0.01$. In Figure 4a, we run SGD on the empirical risk associated with (11), while in Figure 4b, we perform gradient descent on its analytic counterpart (12).

In both cases, the iterates converge toward the principal eigenvector of Σ. This demonstrates that the recovery of the leading principal component is not specific to the softmax mechanism, but rather reflects a more general phenomenon driven by the structure of the aggregation underlying attention.



(a) SGD on empirical risk of linear attention.

(b) GD on analytic risk of linear attention.

Figure 4: Convergence of linear attention toward the principal eigenvector under empirical and analytic risks.

D.4 Scaling with Prompt Length

We now investigate the role of the prompt length L in both the softmax and linear attention settings. In both cases, we consider prompt lengths L ranging from 3 to 50 using 20 evenly spaced values. We fix the dimension to $d = 5$, and run the optimization for $T = 5000$ iterations. For each value of L, we perform 10 independent runs and measure the final alignment.

The only difference between the two settings lies in the choice of the parameter λ: we use $\lambda = 0.1$ for the softmax model and $\lambda = 0.001$ for the linear attention model.

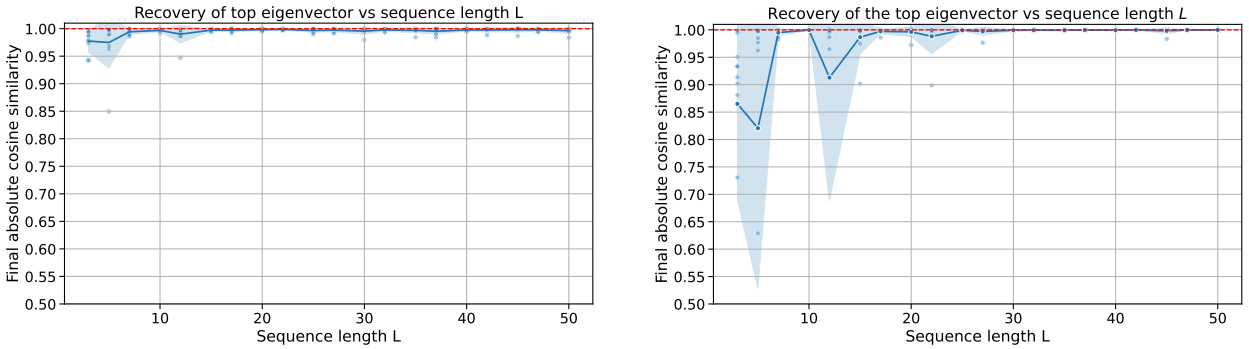
As shown in Figure 5, performance improves with L in both settings, illustrating the transition from the finite-prompt regime to the population regime. In particular, the alignment increases and stabilizes as L grows, providing empirical evidence that the finite-prompt model converges toward its infinite-prompt counterpart. Moreover, the consistency between the softmax and linear cases suggests that this behavior is not specific to the softmax mechanism, but rather reflects a more general phenomenon tied to the structure of the aggregation.

D.5 Scaling with Dimension

We also study the effect of the ambient dimension d across both the softmax and linear attention models. For dimensions d ranging from 3 to 100 in increments of 5, we generate a new covariance matrix Σ for each dimension and evaluate the final alignment after $T = 5000$ iterations, averaging over 10 independent runs in each case.

In all experiments, we scale the hyperparameters with the dimension. In the softmax model, we set the learning rate $\gamma = 0.5/d^2$ and $\lambda = 0.1/d$, while in the linear attention model we use $\gamma = 1/d^2$ and $\lambda = 0.01/d$. We also fix the context length to $L = d$ in the linear case.

For the softmax model, we analyze the infinite-prompt regime, which admits an explicit closed-form expression depending on Σ (see (4)). For the linear attention model, we consider its finite-prompt formulation, which also admits an explicit closed-form expression (for fixed L) depending on Σ (see (12)).

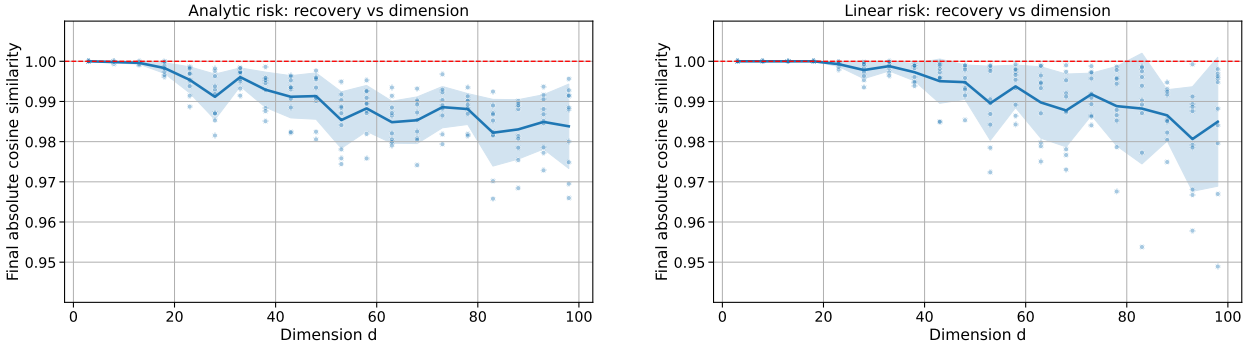


(a) Softmax attention

(b) Linear attention

Figure 5: Final alignment as a function of the prompt length L for both softmax and linear attention models.

Figure 6 reports the resulting performance as a function of d. In both models, we observe that increasing the dimension makes recovery more challenging, reflecting the growing difficulty of estimating the principal component in higher-dimensional spaces. Despite this degradation, both methods consistently retain a strong alignment with the leading eigenvector, highlighting the robustness of the underlying mechanism.



(a) Softmax attention

(b) Linear attention

Figure 6: Final alignment as a function of the dimension d.

D.6 In-Context Learning: Finite and Infinite Prompt

Finally, we consider the in-context learning (ICL) setting, where the covariance matrix is itself random and follows a spiked Wishart distribution:

$$\Sigma \sim W_d\big(\xi^2 I_d + \theta v v^\top, n\big),$$

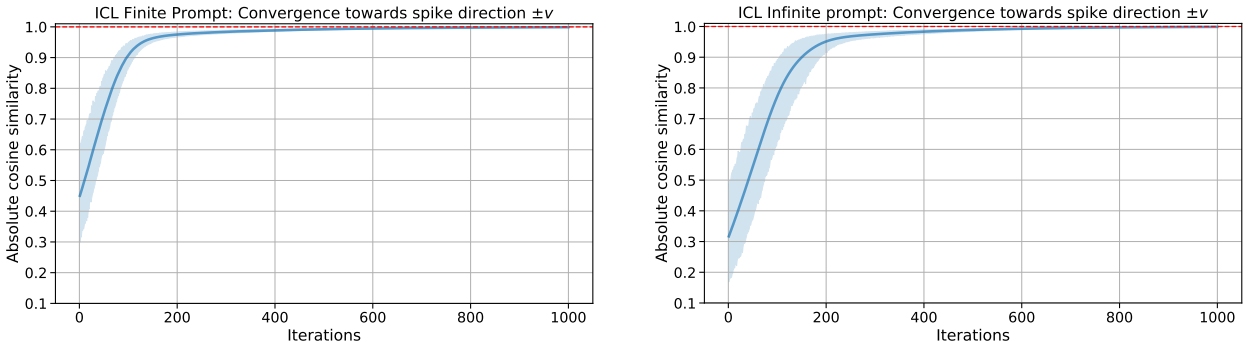
with $d = 5, \xi = 1$, $\theta = 2$, and $n = 10$.

In the finite-prompt regime, we approximate the risk (8) using Monte Carlo sampling with 100 samples of Σ and 100 samples of data per covariance matrix. In the infinite-prompt regime, we optimize the population risk directly using Lemma 13.

Figures 7a and 7b show convergence toward the spike direction v. Both figures superposed are shown in Figure 2.

D.7 In-context learning: Scaling with Prompt Length

We now study the effect of the prompt length L in the in-context learning (ICL) setting. We consider values of L ranging from 3 to 100 in increments of 5. For each value of L, we perform 10 independent runs and report the final alignment after $T = 3000$ iterations.



(a) SGD on finite prompt ICL risk, $L = 100$

(b) GD on infinite prompt ICL risk

Figure 7: In-context learning: finite vs. infinite prompt regimes.

In the finite-prompt regime, for each iteration we sample covariance matrices from the spiked Wishart distribution and generate data accordingly, while in the infinite-prompt regime we directly optimize the corresponding population risk.

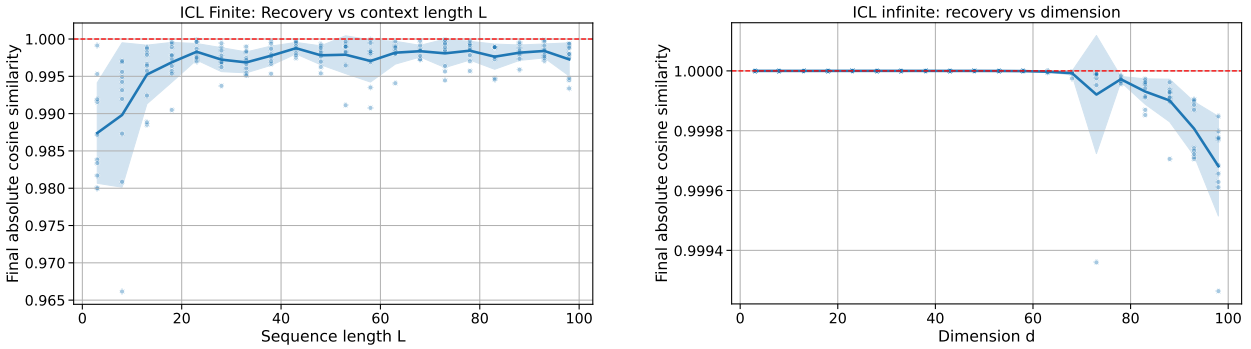
Figure 8a shows the final alignment as a function of L. We observe that, similarly to the standard softmax setting, performance improves as the prompt length increases. This reflects the fact that larger prompts provide a better approximation of the population objective, reducing the variability induced by sampling both the data and the covariance matrices.

These results further support the theoretical prediction that the finite-prompt ICL model converges toward its infinite-prompt counterpart as L grows.

D.8 Scaling with Dimension

In this experiment, we investigate how the ambient dimension d affects the performance of the infinite-prompt in-context learning (ICL) model. We consider dimensions d ranging from 3 to 100 in increments of 5. For each value of d, we fix $n = d$, sample a covariance matrix Σ, and evaluate the alignment of the learned direction with the spike direction v after $T = 2000$ iterations, where the hyperparameters are scaled with the dimension, with learning rate $\gamma = 0.5/d^2$ and $\lambda = 0.1/d$.

Figure 8b reports the resulting alignment as a function of d. As the dimension increases, we observe a gradual degradation in performance, reflecting the increased difficulty of extracting the principal component in higher-dimensional settings. Nevertheless, the model maintains a significant alignment with the leading eigenvector across all dimensions, illustrating the robustness of the ICL mechanism in the infinite-prompt limit.



(a) Final alignment as a function of the prompt length L.

(b) Final alignment as a function of the dimension d.

Figure 8: ICL performance as a function of prompt length and dimension.

D.9 Discussion on the numerical experiments

Overall, the numerical experiments strongly support our theoretical findings. Across all models, we observe consistent recovery of the principal component or spike direction. The experiments highlight the role of finite-prompt effects and illustrate the convergence of finite models toward their corresponding population limits. In this section, we present numerical experiments that illustrate and empirically validate the theoretical results developed throughout the paper. In particular, we study the convergence behavior of the different models toward the principal eigenvector of the underlying covariance structure, as well as the effect of key parameters such as the prompt length and the ambient dimension. The experiments run in a few minutes on a standard laptop, except for Figure 8a, which may take up to an hour due to Monte Carlo sampling of the annealed expectations.