Attention-based clustering

Rodrigo Maulen-Soto, Claire Boyer, Pierre Marion

September 2, 2025

Abstract

Transformers have emerged as a powerful neural network architecture capable of tackling a wide range of learning tasks. In this work, we provide a theoretical analysis of their ability to automatically extract structure from data in an unsupervised setting. In particular, we demonstrate their suitability for clustering when the input data is generated from a Gaussian mixture model. To this end, we study a simplified two-head attention layer and define a population risk whose minimization with unlabeled data drives the head parameters to align with the true mixture centroids.

This phenomenon highlights the ability of attention-based layers to capture underlying distributional structure. We further examine an attention layer with key, query, and value matrices fixed to the identity, and show that, even without any trainable parameters, it can perform in-context quantization, revealing the surprising capacity of transformer-based methods to adapt dynamically to input-specific distributions.

1 Introduction

Attention-based models (Bahdanau et al., 2015), in particular Transformers (Vaswani et al., 2017), have achieved state-of-the-art performance across a wide range of learning tasks. These include applications 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 success of the attention mechanism has been linked to its ability to capture long-range relationships in input sequences (Bahdanau et al., 2015; Vaswani et al., 2017). They do this by computing pairwise dependencies between tokens based on learned projections, without regard to the tokens' positions in the sequence.

On the theoretical side, a full understanding of attention-based mechanisms has not yet been developed. This is due to the complexity of the architectures and the diversity of relevant tasks they manage to achieve. A promising research direction to bridge this gap involves identifying essential features from real-world problems and constructing minimal yet representative tasks that retain the essential difficulty—paired with provable models that solve them using attention-based mechanisms. Notable recent efforts in this vein include Ahn et al. (2023); von Oswald et al. (2023); Yang et al. (2025); Zhang et al. (2024); Li et al. (2024, 2023). However, the existing literature mainly focuses on supervised learning aspects, and in particular in-context learning (von Oswald et al., 2023; Zhang et al., 2024; Garg et al., 2023; Li et al., 2023; Furuya et al., 2024). The goal of in-context learning is to predict the output corresponding to a new query, given a prompt consisting of input/output pairs.

Beyond the standard supervised setting, Transformer models are often (pre)trained in practice with semi-supervised objectives such as masked language modeling (Phuong and Hutter, 2022). This raises important questions about their statistical behavior and training dynamics in unsupervised regimes. In this work, we examine Transformers through the lens of clustering, thereby revealing the inherent capacity of attention mechanisms to perform unsupervised representation learning. To the best of our knowledge, the only prior theoretical work that explicitly explores clustering with Transformers is He et al. (2025), who demonstrate that attention layers can mimic the EM algorithm (Lloyd, 1982), albeit assuming known cluster labels during training. In contrast, our analysis focuses on the fully unsupervised setting and further provides insight into the functional roles of individual attention heads in the context of model-based clustering.

Contributions. In this paper, we investigate the behavior of attention layers in an unsupervised learning setting, where input data is drawn from mixture distributions. We focus on a two-component mixture model, beginning with a simplified setup based on Dirac masses and extending to a more realistic scenario involving Gaussian components. Within this classical clustering framework, we introduce a two-headed linear attention layer designed to capture cluster membership through attention scores, while remaining analytically and computationally tractable. To assess the quality of the embeddings produced by the attention mechanism, we define a theoretical risk analogous to the classical quantization error in unsupervised learning. We analyze the training dynamics of the proposed predictor under projected gradient descent and prove that, with appropriate initialization, the algorithm can learn the true latent centroids of the mixture components, despite the non-convexity of the loss landscape and without access to cluster labels. To relax the initialization requirements in practice, we further propose a regularization scheme that promotes disentanglement between attention heads. Our theoretical findings are supported by numerical experiments under varying conditions, including different initialization regimes, mixture separability levels, and problem dimensionalities. Overall, we show that attention-based predictors can successfully adapt to mixture models by learning the underlying centroids through training. We also study their quantization properties in the oracle regime, where parameters have converged to the true centroids. Finally, we focus on a particular attention layer in which the key, query, and value parameters are fixed to the identity matrix. Surprisingly, we show that this type of layer, despite having no trainable parameters, can still perform in-context quantization, meaning it adapts to the case where the distribution of each input sequence comes with its own mixing parameters. This further demonstrates the remarkable ability of transformer-based methods to adapt on the fly to the underlying data distribution, even when no attention parameters are trained.

Organization. Section 2 introduces the problem and outlines the proposed approach. Section A presents an oversimplified version of the problem, using linear attention to solve a two-component Dirac mixture model. In Section 3, we address the general problem with linear attention applied to a two-component Gaussian mixture model. In Section 4, we discuss the quantization properties of attention-based predictor with oracle parameters. In Section 5, we explore an in-context clustering framework and examine the quantization capabilities of a simple attention-based layer with no learned parameters. The proofs of all the theoretical results can be found in the corresponding appendices.

2 A starter on attention-based layers and clustering

Data distribution in model-based clustering. In attention-based learning, the key idea is to map a set of input tokens to a transformed set of output tokens. With this in mind, we consider an input sequence $\mathbb{X} \in \mathbb{R}^{L\times d}$ composed of L tokens $(X_1, \dots, X_L)$, each token being a vector of $\mathbb{R}^d$, i.e.,

$$\mathbb{X} = \begin{pmatrix} X_1^\top \\ \vdots \\ X_L^\top \end{pmatrix} \in \mathbb{R}^{L\times d}.$$

We assume that the tokens are i.i.d. drawn from a simple mixture model: for $1 \leq \ell \leq L$,

$$X_\ell \sim \frac{1}{2}\mathcal{N}(\mu_0^\star, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\mu_1^\star, \sigma^2 I_d), \tag{P$_\sigma$}$$

with balanced components and where the centroids $(\mu_0^\star, \mu_1^\star) \in (\mathbb{S}^{d-1})^2$ (i.e., $\|\mu_0^\star\|_2 = \|\mu_1^\star\|_2 = 1$) are assumed to be orthogonal, i.e., such that $\langle \mu_0^\star, \mu_1^\star\rangle = 0$. Therefore, for each token, there exists an associated latent variable, denoted by Z_ℓ, corresponding to a Bernoulli random variable of parameter $1/2$ and encoding its corresponding cluster.

To initiate the mathematical analysis of Transformer-based layers in a clustering setting, we have considered the degenerate case where $\sigma^2 = 0$, i.e., where samples are drawn from a mixture of Dirac masses. We carried out a detailed analysis of the training dynamics for such a degenerate model. This preliminary study proved insightful and helped guide our analysis of the non-degenerate mixture model. For the reader's interest, it is provided in Appendix A.

Attention-based predictors. An attention head made of a self-attention layer can be written as

$$H^{\mathrm{soft}_\lambda}(\mathbb{X}) = \mathrm{softmax}_\lambda\left(\mathbb{X}QK^\top\mathbb{X}^\top\right)\mathbb{X}V$$

where the softmax of temperature $\lambda > 0$ is applied row-wise, no skip connection is considered and the matrices $K, Q, V \in \mathbb{R}^{d\times p}$ are usually referred to as keys, queries and values. We adopt the convention that the values are identity matrices; thus, the attention head simply outputs combinations of the initial tokens weighted by the attention scores. While this simplification is certainly convenient for facilitating the mathematical analysis that follows, it is also supported by experimental studies showing comparable performance when the value matrices are removed (He and Hofmann, 2024). Furthermore, assume that the key and query matrices are equal to the same column matrix $\mu \in \mathbb{R}^{d\times 1}$, we obtain

$$H^{\mathrm{soft}_\lambda,\mu}(\mathbb{X}) = \mathrm{softmax}_\lambda\left(\mathbb{X}\mu\mu^\top\mathbb{X}^\top\right)\mathbb{X}. \tag{1}$$

With such an architecture, the ℓ-th output vector is therefore given by

$$H^{\mathrm{soft}_\lambda,\mu}(\mathbb{X})_\ell = \sum_{k=1}^{L}\mathrm{softmax}_\lambda\left(X_\ell^\top\mu\mu^\top\mathbb{X}^\top\right)_k X_k, \tag{2}$$

which corresponds to aggregating the X_k's when X_k and X_ℓ are simultaneously aligned with μ. This suggests that attention heads may act as effective learners in a clustering framework.

The softmax nonlinearity used in the attention head (1) introduces a coupling between tokens, which undoubtedly complicates the mathematical analysis. To address this difficulty, we propose to look at a simplified linear attention head, still parameterized by $\mu \in \mathbb{R}^d$, and defined for $1 \le \ell \le L$, as

$$H^{\mathrm{lin},\mu}(\mathbb{X})_\ell = \frac{2}{L}\sum_{k=1}^{L}(\lambda X_\ell^\top\mu\mu^\top X_k)X_k. \tag{3}$$

This head uses a linear activation function instead of the traditional softmax found in practical architectures, and has already received interest in several mathematical studies (see Zhang et al., 2024; von Oswald et al., 2023; Han et al., 2023; Katharopoulos et al., 2020).

Note that when μ is chosen to be $\mu_0^\star$, then for tokens X_ℓ and X_k whose corresponding latent variables Z_ℓ and Z_k are both equal to 0 (i.e., the samples belong to the same cluster centered at $\mu_0^\star$), the vectors X_ℓ and X_k are likely to be aligned with $\mu_0^\star$. In this case, we have $(X_\ell^\top\mu\mu^\top X_k)X_k \simeq X_k$. Conversely, if X_ℓ and X_k are associated with different latent variables (e.g., $Z_\ell = 0$ and $Z_k = 1$), then $(X_\ell^\top\mu\mu^\top X_k)X_k \simeq 0$. This behavior suggests that when setting $\mu = \mu_0^\star$, and if X_ℓ belongs to the cluster centered at $\mu_0^\star$, the sum $\sum_{k=1}^{L}(\lambda X_\ell^\top\mu\mu^\top X_k)X_k$ effectively aggregates the X_k's from the same cluster, whose expected number is $L/2$, motivating the renormalizing factor of $2/L$. Overall, $H^{\mathrm{lin},\mu_0^\star}(\mathbb{X})_\ell$ can be seen as producing an empirical mean of the tokens belonging to the same cluster, serving as an estimator of the corresponding centroid.

Therefore, assuming that the number of clusters in the data is known, it is natural to consider an attention-based predictor composed of two attention heads, parameterized by μ_0 and $\mu_1 \in \mathbb{R}^d$,

$$T^{\mathrm{lin},\mu_0,\mu_1}(\mathbb{X}) = H^{\mathrm{lin},\mu_0}(\mathbb{X}) + H^{\mathrm{lin},\mu_1}(\mathbb{X}). \tag{4}$$

Metric loss. As no label is available, we focus on minimizing the following theoretical loss:

$$\mathcal{L}(T) \overset{\mathrm{def}}{=} \frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}\left[\|X_\ell - T(\mathbb{X})_\ell\|_2^2\right], \tag{5}$$

where T is an arbitrary attention-based predictor. The distinctive feature of this risk lies in the fact that, if the predictor were able to return, for each token X_ℓ, its associated centroid $\mu_{Z_\ell}^\star$, the risk would exactly correspond to a quantization error, characteristic of a standard clustering task. Note that, due to the independence of the tokens, we have $\mathcal{L}(T) = \mathbb{E}\|X_1 - T(\mathbb{X})_1\|_2^2$, so we can confine the following theoretical analysis on the minimization of the predictive error for the first token only.

PGD iterates. In this paper, we focus on the training dynamics of Transformer-based predictors when minimizing the theoretical risk $\mathcal{L}$. While we acknowledge that, in practice, an empirical version of this risk is typically used, analyzing the optimization of the theoretical risk is already a non-trivial task, which offers valuable insights into the behavior observed in practice.

For a given predictor $T^{\mathrm{lin},\mu_0,\mu_1}$ made of two linear attention heads parameterized respectively by μ_0 and μ_1, one can reinterpret the objective $\mathcal{L}$ as a function $\mathcal{R}$ of the parameters (μ_0, μ_1), defined by

$$\mathcal{R}(\mu_0, \mu_1) = \mathcal{L}(T^{\mathrm{lin},\mu_0,\mu_1}). \tag{6}$$

Note that the computation of the risk also depends on the choice of the underlying data distribution. However, as is common practice in the literature, we do not explicitly indicate the dependence of the risk on the data distribution. This should not hinder understanding, as the distributional assumptions and context will always be made clear. As we rely on the theoretical analysis on an expression of this risk restricted to the sphere, we consider as a gradient strategy, the Projected (Riemaniann) Gradient Descent (PGD) algorithm (Boumal, 2023). Given an initialization $(\mu_0^0, \mu_1^0) \in (\mathbb{S}^{d-1})^2$ and a step size $\gamma > 0$, the PGD iterates $(\mu_0^k, \mu_1^k) \in (\mathbb{S}^{d-1})^2$ are recursively defined by:

$$\begin{aligned} \mu_0^{k+1} &= \frac{\mu_0^k - \gamma(I_d - \mu_0^k(\mu_0^k)^\top)\nabla_{\mu_0}\mathcal{R}(\mu_0^k, \mu_1^k)}{\|\mu_0^k - \gamma(I_d - \mu_0^k(\mu_0^k)^\top)\nabla_{\mu_0}\mathcal{R}(\mu_0^k, \mu_1^k)\|_2}, \\ \mu_1^{k+1} &= \frac{\mu_1^k - \gamma(I_d - \mu_1^k(\mu_1^k)^\top)\nabla_{\mu_1}\mathcal{R}(\mu_0^k, \mu_1^k)}{\|\mu_1^k - \gamma(I_d - \mu_1^k(\mu_1^k)^\top)\nabla_{\mu_1}\mathcal{R}(\mu_0^k, \mu_1^k)\|_2}. \end{aligned} \tag{PGD}$$

In what follows, we analyze the convergence of these iterates to the oracle centroids, both theoretically, progressing from simplified mixtures to more complex ones, and numerically.

3 Training dynamics: The centroids are learned as attention parameters

We now turn our attention to a more standard and practically relevant setting, that of Gaussian mixtures, as defined in (P_σ). Specifically, we consider the case of i.i.d. tokens drawn as

$$X_\ell \sim \frac{1}{2}\mathcal{N}(\mu_0^\star, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\mu_1^\star, \sigma^2 I_d),$$

for $1 \leq \ell \leq L$, where $(\mu_0^\star, \mu_1^\star) \in (\mathbb{S}^{d-1})^2$ are orthogonal unit vectors, i.e., $\langle \mu_0^\star, \mu_1^\star \rangle = 0$. Our analysis continues to rely on the risk $\mathcal{R}$, now evaluated under this Gaussian mixture model.

3.1 Theoretical analysis

Preliminary computations. We start by introducing the following quantities:

$$\kappa_0 \overset{\text{def}}{=} \langle \mu_0^\star, \mu_0 \rangle, \quad \kappa_1 \overset{\text{def}}{=} \langle \mu_1^\star, \mu_1 \rangle, \quad \eta_0 \overset{\text{def}}{=} \langle \mu_1, \mu_0^\star \rangle, \quad \eta_1 \overset{\text{def}}{=} \langle \mu_0, \mu_1^\star \rangle, \quad \xi \overset{\text{def}}{=} \langle \mu_0, \mu_1 \rangle, \tag{7}$$

and derive a closed-form expression for the risk w.r.t. this reparameterization. Although the full expression is somewhat complex (see Appendix C.1), the following proposition highlights its key structural properties, as being a polynomial in these five variables.

Proposition 3.1. *Under the Gaussian mixture model* (P_σ)*, consider the attention-based predictor* $T^{\mathrm{lin},\mu_0,\mu_1}$ *composed of two linear heads parameterized by* $(\mu_0, \mu_1) \in (\mathbb{S}^{d-1})^2$*. Then, there exists a function* $\mathcal{R}^< : [-1,1]^5 \mapsto \mathbb{R}$ *such that* $\mathcal{R}(\mu_0, \mu_1) = \mathcal{R}^<(\kappa_0, \kappa_1, \eta_0, \eta_1, \xi)$ *and*

$$\begin{aligned} \mathcal{R}^<(\kappa_0, \kappa_1, \eta_0, \eta_1, \xi) \in \operatorname{span}\big(\{&\kappa_0^4, \eta_0^4, \kappa_1^4, \eta_1^4, \kappa_0^2\eta_0^2, \kappa_1^2\eta_1^2, \kappa_0^2\eta_1^2, \kappa_1^2\eta_0^2, \kappa_0\eta_0\kappa_1\eta_1, \\ &\kappa_0^2, \eta_0^2, \kappa_1^2, \eta_1^2, \kappa_0\eta_0\xi, \kappa_1\eta_1\xi, \xi^2, 1\}\big). \end{aligned} \tag{8}$$

We remark that when η_0, η_1, and ξ are fixed to 0, most of the monomials vanish, yielding a fully explicit formula for the risk (see Lemma C.1 in the appendices). Far from being a mere rewrite, this step provides the algebraic foundation for all the exact calculations and insights that follow.

Optimality conditions. Given the complexity of the theoretical analysis, we focus on a simplified setting by restricting our study to parameters (μ_0, μ_1) lying on a specific manifold[1]:

$$\mathcal{M} = \{(\mu_0, \mu_1) \in (\mathbb{S}^{d-1})^2 : \langle \mu_1^\star, \mu_0 \rangle = 0, \langle \mu_0^\star, \mu_1 \rangle = 0, \langle \mu_0, \mu_1 \rangle = 0\}. \tag{9}$$

In terms of notation, it is equivalent to assume that $\eta_0 = 0, \eta_1 = 0, \xi = 0$. Therefore on this manifold, we adopt the shorthand notation $\mathcal{R}^<(\kappa_0, \kappa_1) \overset{\text{def}}{=} \mathcal{R}^<(\kappa_0, \kappa_1, 0, 0, 0)$.

[1] up to relabeling the head parameters, since a priori, μ_0 (resp. μ_1) does not have to be automatically related to $\mu_0^\star$ (resp. $\mu_1^\star$).

Lemma 3.2. *Under the Gaussian mixture model* (P_σ)*, the risk* $\mathcal{R}^<$ *restricted to* $\mathcal{M}$ *has the form*

$$\mathcal{R}^<(\kappa_0, \kappa_1) = A(\kappa_0^4 + \kappa_1^4) + B(\kappa_0^2 + \kappa_1^2) + C\kappa_0^2\kappa_1^2 + D, \tag{10}$$

for A, B, C, D *non-negatives constants, dependent on* σ *and* L*, made explicit in Lemma C.1.*

Proposition 3.3. *Consider* $\mathcal{R}^<(\kappa_0, \kappa_1)$*, there exists* $\lambda^\star(\sigma, L) > 0$ *such that the points* $(\pm 1, \pm 1)$ *are global minima of* $\mathcal{R}^<(\kappa_0, \kappa_1)$*.*

This result demonstrates that, under a suitable condition on the temperature parameter—specifically, when $\lambda = \lambda^\star(\sigma, L)$—the points $\pm\mu_0^\star$ and $\pm\mu_1^\star$ are global minimizers of the risk. The explicit form of $\lambda^\star(\sigma, L)$ is provided in Proposition C.2 in the appendices. Moreover, it is worth noting that as the variance σ^2 of the Gaussian components tends to zero, $\lambda^\star(\sigma, L)$ approaches the value $\lambda_0^\star = \frac{L+1}{L+3}$,which coincides with the value previously identified in the degenerate case (P_0). On the other hand, when σ is fixed and L grows large, $\lambda^\star(\sigma, L)$ tends toward $\lambda_\infty^\star = \frac{1+4\sigma^2}{1+5\sigma^2+6\sigma^4}$, which will guide us to properly choose λ in our numerical experiments.

Convergence analysis. In what follows, we show that the (PGD) iterates can indeed converge to global minimizers, provided they are suitably initialized on the manifold $\mathcal{M}$.

Theorem 3.4. *Under the Gaussian mixture model* (P_σ)*, consider the attention-based predictor* $T^{\mathrm{lin},\mu_0,\mu_1}$ *composed of two linear heads. Take* $\lambda \in]0, \lambda^\star(\sigma, L)]$*, with* $\lambda^\star(\sigma, L)$ *defined as in Proposition 3.3. Then there exists* $\bar{\gamma} > 0$ *such that for any stepsize* $0 < \gamma < \bar{\gamma}$*, and for a generic initialization* $(\mu_0^0, \mu_1^0) \in \mathcal{M}$*, the iterates* (μ_0^k, μ_1^k) *generated by* (PGD) *converge to the centroids (up to a sign), i.e.*

$$(\mu_0^k, \mu_1^k) \xrightarrow[k\to\infty]{} (\pm\mu_0^\star, \pm\mu_1^\star).$$

Theorem 3.4 underlines the capabilities of linear attention-based predictors in a clustering framework. With appropriate initialization, the attention heads align with the true underlying centroids even when trained without access to labels. This result shows that attention layers can uncover and encode the latent structure of the input distribution in a fully unsupervised setting through their parameters.

3.2 Experimental verification of the theoretical results

Setting. To better reflect practical algorithmic behavior, we implement Projected Stochastic Gradient Descent (PSGD; see Appendix F.1), which serves as an empirical counterpart to the (PGD) iterates by relying on sample-based estimations.

In what follows, we use the metric referred to as *distance to the centroids (up to a sign)*, given by

$$\sqrt{\min\{\mathrm{dist}_1, \mathrm{dist}_2\}}, \tag{11}$$

where

$$\mathrm{dist}_1 = \min\{\|\hat{\mu}_0 - \mu_0^\star\|^2, \|\hat{\mu}_0 + \mu_0^\star\|^2\} + \min\{\|\hat{\mu}_1 - \mu_1^\star\|^2, \|\hat{\mu}_1 + \mu_1^\star\|^2\},$$
$$\mathrm{dist}_2 = \min\{\|\hat{\mu}_0 - \mu_1^\star\|^2, \|\hat{\mu}_0 + \mu_1^\star\|^2\} + \min\{\|\hat{\mu}_1 - \mu_0^\star\|^2, \|\hat{\mu}_1 + \mu_0^\star\|^2\},$$

and $\mu_0^\star, \mu_1^\star$ denote the true centroids, respectively, while $\hat{\mu}_0, \hat{\mu}_1$ are the parameters returned by (PSGD). Note that this distance is invariant under relabeling and sign flips of the head parameters. More implementation details related to the following experiments can be found in Appendix F.2.

It is worth noting that the assumption of orthogonality of the centroids on the unit sphere always results in a constant distance between the centroids, namely $\|\mu_0^\star - \mu_1^\star\|_2 = \sqrt{2}$. In this setting, to characterize the separation between the two modes of the mixture, one can introduce a notion of *interference* that depends solely on the variance level of each mode and which is defined as $\mathrm{I}(\sigma) = \mathbb{P}(X_\sigma > \frac{\sqrt{2}}{2})$, where $X_\sigma \sim \mathcal{N}(0, \sigma^2)$. Remark that this function is increasing with supremum 0.5. This motivates the choice of two contrasting scenarios for the numerical experiments: a low-interference regime with $\sigma = 0.3$, where $\mathrm{I}(0.3) \approx 0.01$, and a high-interference regime with $\sigma = 1$, where $\mathrm{I}(1) \approx 0.24$. More implementation details of the following numerical experiments can be found in Appendix F.2.

Results. When initialization is done on the manifold, the training analysis depicted in Figure 1 demonstrates linear convergence of the head parameters toward the centroids (up to a sign) during the first 10^3 iterations, which is in line with the obtained theoretical results. The error then plateaus at around 10^{-2} in the low-interference setting ($\sigma = 0.3$), and around 10^{-1} in the high-interference setting ($\sigma = 1$). This saturation phenomenon is attributed to the stochasticity introduced by using (PSGD) in place of (PGD) in the simulations.

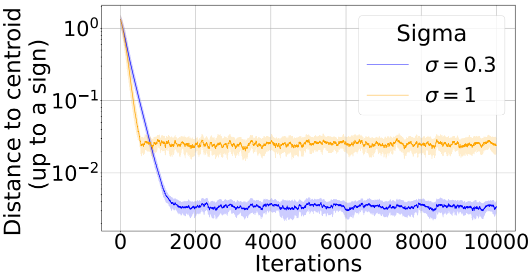


Figure 1: Distance to centroids vs (PSGD) iterations for the minimization of $\mathcal{R}$, with an initialization on the manifold $\mathcal{M}$. 10 runs, 95% percentile intervals are plotted.

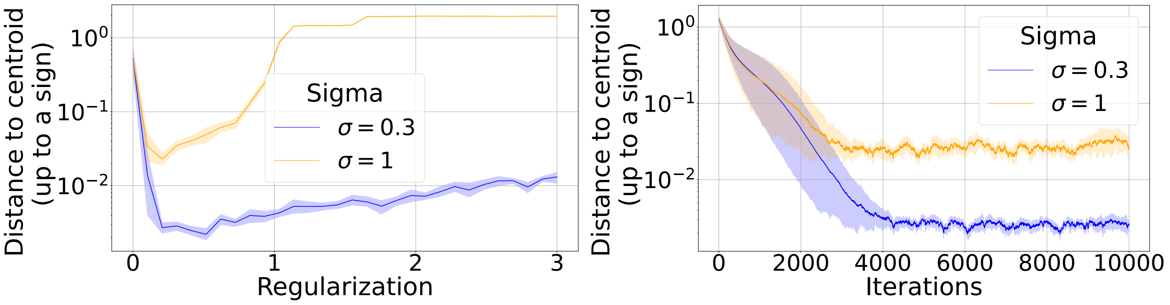
3.3 Generalizations

We consider several generalizations of our theoretical setting, to give insight on the role of our assumptions.

Random initialization on the unit sphere (outside of the manifold). When the initialization is performed outside the manifold, PGD iterates only partially align with the underlying centroids. A way to handle arbitrary initializations (suggested by our analysis in the degenerate case, see Appendix A), is to introduce a regularized risk minimization problem:

$$\min_{\mu_0,\mu_1\in\mathbb{S}^{d-1}} \mathcal{R}^{\rho}(\mu_0,\mu_1), \quad \text{with} \quad \mathcal{R}^{\rho}(\mu_0,\mu_1) \overset{\text{def}}{=} \mathcal{R}(\mu_0,\mu_1) + \rho r(\mu_0,\mu_1), \tag{$\mathcal{P}_\rho$}$$

for $\rho > 0$ and the regularization term defined by $r(\mu_0,\mu_1) = \mathbb{E}[\langle\mu_0, X_1\rangle^2\langle\mu_1, X_1\rangle^2]$. The role of this term is to encourage the orthogonality conditions on μ_0, μ_1, thereby compensating for initializations that may fall outside the manifold $\mathcal{M}$. Numerical results show that the centroids can be recovered with an appropriate level of regularization (see Figure 6a). Note that, as the strength of the regularization increases, it gradually overrides the original objective and impairs the alignment of the head parameters with the true centroids —an effect that becomes more pronounced at higher noise levels.



(a) Distance to centroids after 5000 iterations vs regularization strength ρ for the minimization of $\mathcal{R}^{\rho}$.

(b) Distance to centroids vs (PSGD) iterations for the minimization of $\mathcal{R}^{\rho}$, with regularization $\rho = 0.2$.

Figure 2: Performance of (PSGD), when initializing on the unit sphere and minimizing the regularized risk ($\mathcal{P}_\rho$). 10 runs, 95% percentile intervals are plotted.

In Appendix G.2, we present additional experiments in higher-dimensional settings, highlighting the impact of dimensionality on the training dynamics.

Orthogonality of the clusters. The framework studied in this paper assumes that the centroids of the clusters are orthogonal. Relaxing this assumption to allow for potential overlap between centroid directions significantly complicates the theoretical analysis by introducing additional terms.

More clusters. A natural extension is to consider the case of K centroids and attention heads. We perform an experiment in the case $K = 3$, which shows recovery of the centroids, see details in Appendix G.4. The main difference with the case $K = 2$ is the regularization term which now writes

$$r(\mu_0,\mu_1,\mu_2) = \sum_{0\le i<j\le 2} \langle\mu_i, X_1\rangle^2\langle\mu_j, X_1\rangle^2$$

to promote pairwise orthogonality between the parameters. This approach should further generalize to the case of K orthonormal centroids with $K < d$.

4 Attention-based layers as approximate quantizers

We have seen that attention-based predictors can adapt to mixture models by learning the underlying centroids through training. In this section, we investigate the quantization properties of an attention-based predictor whose parameters have converged to the true centroids. To guide our analysis, we introduce the optimal quantizer $T^\star$ as a statistical benchmark within a standard clustering framework. This oracle predictor returns the true centroid of each token, that is, for $1 \leq \ell \leq L$, $T^\star(\mathbb{X})_\ell = \mu^\star_{Z_\ell}$ where Z_ℓ is encoding the latent cluster of the token X_ℓ. One can immediately note that the risk of the optimal quantizer is given by

$$\mathcal{L}(T^\star) = \mathbb{E}\left[\left\|X_1 - \mu^\star_{Z_1}\right\|_2^2\right] = d\sigma^2.$$

Returning to the attention-based predictor $T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}$ with oracle parameters, the first key observation is that it closely resembles an optimal quantizer: its ℓ-th output aligns, on average, with the centroid of the cluster to which the ℓ-th token belongs, as shown by the next lemma.

Proposition 4.1. *Under the Gaussian mixture model* (P_σ)*, it holds that*

$$\mathbb{E}[T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}(\mathbb{X})_1|Z_1 = c] = \mu_c^\star \frac{\lambda}{L}[(L+1) + 2(L+3)\sigma^2], \quad \textit{for} \quad c = \{0,1\}.$$

Therefore, choosing $\lambda = \frac{L}{(L+1)+2(L+3)\sigma^2}$ *leads to unbiased approximate quantization, i.e.,*

$$\mathbb{E}[T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}(\mathbb{X})_1|Z_1 = c] = \mu_c^\star.$$

One can next characterize the asymptotic risk and variance of the oracle attention-based predictor.

Proposition 4.2. *Under the Gaussian mixture model* (P_σ)*, fix the temperature* $\lambda = \frac{1+4\sigma^2+4\sigma^4}{1+6\sigma^2+12\sigma^4+8\sigma^6}$. *Then, the risk of the attention-based predictor* $T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}$ *with oracle parameters* $\mu_0^\star$ *and* $\mu_1^\star$ *satisfies*

$$\lim_{L\to\infty} \mathcal{L}(T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}) = \sigma^2(d-2).$$

Moreover, for an arbitrary value of λ*, when* $L \to \infty$*, we get*

$$\lim_{L\to\infty} \mathrm{Var}[T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}(\mathbb{X})_1|Z_1 = c] = 2\lambda^2\sigma^2(1+2\sigma^2)^2.$$

Strikingly, as the input sequence length L increases, we find that

$$\lim_{L\to\infty} \frac{\mathcal{L}(T^{\mathrm{lin},\mu_0^\star,\mu_1^\star})}{\mathcal{L}(T^\star)} = 1 - \frac{2}{d}.$$

This result shows that the attention-based predictor asymptotically achieves a lower risk than the optimal quantizer. This phenomenon can be partly explained by the fact that the comparison is not entirely fair: the predictors $T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}$ and $T^\star$ do not belong to the same class of functions. Indeed, the optimal quantizer $T^\star$ is only allowed to return two fixed vectors and relies on a single input token to predict the associated centroid (albeit with access to the latent label). On the other hand, the image of attention-based encoder $T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}$ is not discrete, and moreover this estimator aggregates a growing sequence of random variables, all drawn from the same mixture. The aggregation of multiple inputs can be seen as a variance reduction mechanism, which lowers the risk (this is also evident in the proof of Proposition 4.2, where the risk is shown to decrease as L increases). Note that the gap vanishes in a high-dimensional setting $d \to \infty$. Another insightful comparison between the predictors $T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}$ and $T^\star$ is through their conditional variances $\mathrm{Var}[T(\mathbb{X})_1|Z_1 = c]$. While the conditional variance of the optimal quantizer is null by definition, it is positive for the linear attention layer, and asymptotically independent of d as shown in Proposition 4.2. This once again highlights the fact that these two quantifiers belong to function classes of different complexity.

These properties are illustrated in Figure 3: we observe that the attention-based embeddings are approximate projections of the inputs on the line between the two centroids. In particular, the variance of the embedded point cloud is lower than the variance of the inputs (which precisely means that $\mathcal{L}(T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}) < \mathcal{L}(T^\star)$), while remaining positive (i.e., $\mathrm{Var}[T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}(\mathbb{X})_1|Z_1 = c] > 0$).

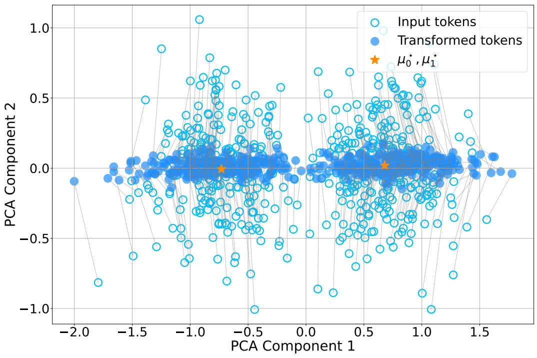


Figure 3: Comparison between inputs and attention-based embeddings ($d = 5, \sigma = 0.3$).

5 In-context clustering

5.1 Setting

So far, we have considered the traditional setting for model-based clustering, with a mixture of Gaussian made of two components: each token was assumed to be distributed as follows

$$X_\ell \sim \frac{1}{2}\mathcal{N}(\mu_0^\star, \sigma^2 I) + \frac{1}{2}\mathcal{N}(\mu_1^\star, \sigma^2 I)$$

with fixed centroids $\mu_0^\star$ and $\mu_1^\star$ of unit-norm and orthogonal.

We have shown that despite the non-convexity of the problem, attention-based predictors including two heads could perform approximate quantization and discover the underlying centroids encoded in their parameters. Note that for an input sequence of tokens $\mathbb{X} = (X_1|\dots|X_L)^\top \in \mathbb{R}^{L\times d}$, the first output of the type of attention-based predictor considered in this paper, parameterized by μ_0 and μ_1, can be rewritten as

$$\begin{aligned} T^{\mathrm{lin},\mu_0,\mu_1}(\mathbb{X})_1 &= \frac{2}{L}\sum_{\ell=1}^{L} \lambda(X_1^\top \mu_0 \mu_0^\top X_\ell)X_\ell + \lambda(X_1^\top \mu_1 \mu_1^\top X_\ell)X_\ell \\ &= \frac{2}{L}\sum_{\ell=1}^{L} \lambda X_1^\top (\mu_0\mu_0^\top + \mu_1\mu_1^\top)X_\ell X_\ell. \end{aligned}$$

Therefore, when we consider two linear heads parameterized by row vectors for the queries and keys, and constrain them to be orthogonal, the setup can be interpreted as a single attention head with a query/key matrix of rank 2. Interestingly, we are able to effectively train this rank-2 query/key head by leveraging the non-convex optimization of two simple, row-structured heads.

Now imagine that we challenge the clustering setting, and we assume that each input sequence is still drawn from a 2-component mixture but with its own centroids, i.e., for each input sequence $\mathbb{X}_i = (X_{i1}|\dots|X_{iL}) \in \mathbb{R}^{L\times d}$, $i = 1,\dots,n$, we assume the tokens $(X_{i\ell})_\ell$ to be i.i.d., such that the ℓ-th token is distributed as

$$X_{i\ell}|\mu_{i0}^\star, \mu_{i1}^\star \sim \frac{1}{2}\mathcal{N}(\mu_{i0}^\star, \sigma^2 I) + \frac{1}{2}\mathcal{N}(\mu_{i1}^\star, \sigma^2 I),$$

for some random orthogonal centroids $\mu_{i0}^\star$ and $\mu_{i1}^\star$ of unit-norm.

If the prior distribution over the centroids is concentrated along specific preferred directions, denoted for instance by $\mu_0^{\star\star}$ and $\mu_1^{\star\star}$, then it is highly likely that the predictor studied in this paper, $T^{\mathrm{lin},\mu_0,\mu_1}$, will perform well. A more formal analysis would likely show that, after training, the Transformer's parameters align with these underlying preferred directions. Pushing this idea further, imagine now that the centroids are instead distributed in an isotropic way. In such a case, one can anticipate that $T^{\mathrm{lin},\mu_0,\mu_1}$ will struggle to adapt to the task of in-context clustering due to limited flexibility: only two parameters, $\mu_0, \mu_1 \in \mathbb{S}^{d-1}$, are used to perform the embedding task in which the centroids vary significantly from one input sequence to another. To address this issue, one can think about increasing the degrees of freedom of the predictor. By building upon previous developments, one could initially increase the number of linear attention heads. Specifically, if we consider d attention heads whose parameters are constrained to be row vectors

$(\mu_c)_{c=1,\dots,d} \in (\mathbb{R}^d)^d$, each of unit norm and mutually orthogonal, we obtain the following attention layer: for an input sequence $\mathbb{X} = (X_1, \dots, X_L) \in \mathbb{R}^{L\times d}$, and for $1 \le \ell \le L$,

$$T^{\text{ctx}}(\mathbb{X})_\ell = \sum_{c=1}^{d} H^{\text{lin},\mu_c}(\mathbb{X})_\ell = \frac{2\lambda}{L}\sum_{c=1}^{d}\sum_{k=1}^{L}(X_\ell^\top \mu_c\mu_c^\top X_k)X_k = \frac{2\lambda}{L}\sum_{k=1}^{L}\Big(X_\ell^\top \underbrace{\Big(\sum_{c=1}^{d}\mu_c\mu_c^\top\Big)}_{\text{Id}} X_k\Big)X_k$$

so finally,

$$T^{\text{ctx}}(\mathbb{X})_\ell = \frac{2\lambda}{L}\sum_{k=1}^{L} X_\ell^\top X_k X_k. \tag{12}$$

In a way that is both expected and somewhat surprising, using d simplified linear heads in parallel, while enforcing orthogonality among their parameters, ultimately amounts to employing an attention layer with no trainable parameters. In what follows, we discuss the properties of T^{ctx} in an in-context clustering framework.

More formally, we refer to as the in-context clustering as a setting in which the input consists of a generic sequence of L tokens $X_1, \dots, X_L$, sampled from a Gaussian mixture model whose component means (centroids) are randomly drawn on the unit sphere:

$$\begin{cases} \mu_0^\star \sim \mathcal{U}(\mathbb{S}^{d-1}) \qquad \text{and} \qquad \mu_1^\star \,|\, \mu_0^\star \quad \text{arbitrarily distributed on } \mathbb{S}^{d-1} \cap (\mu_0^\star)^\perp \\ X_1, \dots, X_L | \mu_0^\star, \mu_1^\star \sim \frac{1}{2}\mathcal{N}(\mu_0^\star, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\mu_1^\star, \sigma^2 I_d) \end{cases}$$

Associated with each token X_ℓ, we still consider a latent variable Z_ℓ, corresponding to a Bernoulli random variable of parameter $1/2$ and encoding its corresponding cluster, so that

$$X_\ell | \mu_0^\star, \mu_1^\star, Z_\ell \sim \mathcal{N}(\mu_{Z_\ell}^\star, \sigma^2 I).$$

5.2 Linear attention layers can perform in-context approximate quantization

In what follows, we characterize the risk of the attention-based embedding T^{ctx} defined in (12), when the input sequence contains an infinite number of tokens.

Proposition 5.1. *In the asymptotic regime where $L \to \infty$, one has that*

$$\lim_{L\to\infty} \mathcal{L}(T^{\text{ctx}}) = (1 + \sigma^2 d) - 2\lambda(1 + 4\sigma^2 + 2d\sigma^4) + 4\lambda^2\left(2\left(\sigma^2 + \frac{1}{2}\right)^3 + (d-2)\sigma^6\right).$$

Choosing the temperature $\lambda = \frac{1+4\sigma^2+2d\sigma^4}{4\left(2\left(\sigma^2+\frac{1}{2}\right)^3+(d-2)\sigma^6\right)}$ gives

$$\begin{aligned} \lim_{L\to\infty} \mathcal{L}(T^{\text{ctx}}) &= \sigma^2(d-2)\frac{1+2\sigma^2}{1+6\sigma^2+12\sigma^4+4d\sigma^6} \\ &\le \sigma^2(d-2). \end{aligned}$$

As in the fixed centoids setting (Proposition 4.2), we retrieve that for a suitable choice of λ, the loss satisfies

$$\lim_{L\to\infty}\frac{\mathcal{L}(T^{\text{ctx}})}{\mathcal{L}(T^\star)} = \left(1 - \frac{2}{d}\right)\frac{1+2\sigma^2}{1+6\sigma^2+12\sigma^4+4d\sigma^6} \le \left(1-\frac{2}{d}\right).$$

This result is all the more surprising given that it emerges in a more complex setting, yet with a simpler mechanism. Unlike the attention-based predictor $T^{\text{lin},\mu_0,\mu_1}$—which benefits from access to the true centroids and a trainable architecture—the in-context encoder T^{ctx} achieves a smaller asymptotic risk without any learned parameters. That such a non-parametric encoder can outmatch the performance of oracle quantizers or more informed or optimized methods reveals the power of attention layers in extracting meaningful representations purely through the attention mechanism.

In effect, the attention-based encoder T^{ctx} performs an approximate in-context quantization of the input distribution by aggregating sequences sampled from a Gaussian mixture. As formalized below, this simple architecture effectively captures the underlying statistical structure of the data.

Proposition 5.2. *For $c \in \{0,1\}$, one has*

$$\mathbb{E}\left[T^{\mathrm{ctx}}(\mathbb{X})_1|\mu_1^\star,\mu_0^\star,Z_1=c\right]=\frac{2\lambda}{L}\left[(1+(d+2)\sigma^2)+(L-1)\left(\frac{1}{2}+\sigma^2\right)\right]\mu_c^\star.$$

Choosing $\lambda=\frac{L}{2}\frac{1}{1+(d+2)\sigma^2+(L-1)\left(\frac{1}{2}+\sigma^2\right)}$ yields an unbiased encoding, i.e.,

$$\mathbb{E}\left[T^{\mathrm{ctx}}(\mathbb{X})_1|\mu_1^\star,\mu_0^\star,Z_1=c\right]=\mu_c^\star.$$

Moreover, the conditional variance of the encoding satisfies

$$\begin{aligned}\mathrm{Var}\left[T^{\mathrm{ctx}}(\mathbb{X})_1|\mu_1^\star,\mu_0^\star,Z_1=c\right]=&\frac{4\lambda^2}{L^2}(1+3(d+4)\sigma^2+3(d+2)(d+4)\sigma^4+d(d+2)(d+4)\sigma^6)\\&+\frac{12\lambda^2}{L^2}(L-1)\left(\frac{1}{2}+\frac{(d+8)}{2}\sigma^2+3(d+2)\sigma^4+d(d+2)\sigma^6\right)\\&+\frac{8\lambda^2}{L^2}\frac{(L-1)(L-2)}{2}\left(2\left(\sigma^2+\frac{1}{2}\right)^3+(d-2)\sigma^6\right)\\&-\frac{4\lambda^2}{L^2}\left[1+(d+2)\sigma^2+(L-1)\left(\frac{1}{2}+\sigma^2\right)\right]^2.\end{aligned}$$

When $L\to\infty$,

$$\lim_{L\to\infty}\mathrm{Var}\left[T^{\mathrm{ctx}}(\mathbb{X})_1|\mu_1^\star,\mu_0^\star,Z_1=c\right]=2\lambda^2\sigma^2(1+4\sigma^2+2d\sigma^4).$$

In this asymptotic regime, choosing $\lambda=\frac{1}{1+2\sigma^2}$, we obtain an unbiased encoding with a conditional variance of

$$2\sigma^2\frac{1+4\sigma^2+2d\sigma^4}{(1+2\sigma^2)^2}.$$

It is worth noting that for $d\geq 2$,

$$2\sigma^2\frac{1+4\sigma^2+2d\sigma^4}{(1+2\sigma^2)^2}\leq\sigma^2 d.$$

Therefore, in the regime of infinite input sequences ($L\to\infty$) and $\lambda=\frac{1}{1+2\sigma^2}$, the encoding becomes unbiased and exhibits a variance reduction effect. However, unlike in Proposition 4.1, this variance remains dimension-dependent in general.

6 Conclusion

This work offers a mathematically grounded, principled perspective on the unsupervised learning capabilities of attention mechanisms within mixture models. By combining a classical clustering framework with simplified yet non-trivial attention architectures, we present theoretical and empirical evidence showing that, when properly trained, attention layers can effectively recover latent structure in data. Our analysis provides insight into the training dynamics, quantization behavior, and practical design choices, such as attention head regularization.

We further investigate an in-context setting, where attention-based models still perform efficient approximate quantization, by achieving lower error than the optimal quantizer. Future directions include exploring richer attention architectures, closer to that used in practice, which may further challenge the theoretical analysis.

Acknowledgments

The authors would like to thank Lenka Zdeborová and Gérard Biau for insightful discussions. This work was supported by the Projet Émergence(s) of the City of Paris.

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.

Hedy Attouch. Viscosity solutions of minimization problems. *SIAM Journal of Optimization*, 3:769–806, 1996.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Nerual 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.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds.* Cambridge University, 2023.

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.

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

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.

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.

Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention, 2023. URL https://arxiv.org/abs/2308.00442.

Bobby He and Thomas Hofmann. Simplifying transformer blocks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RtDok9eS3s.

Yihan He, Hong-Yu Chen, Yuan Cao, Jianqing Fan, and Han Liu. Transformers versus the em algorithm in multi-class clustering. *arXiv:2502.06007v1*, 2025.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention, 2020. URL https://arxiv.org/abs/2006.16236.

Kenneth Lange. *Optimization*, volume 2 edition. Springer, New York, 2013.

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.

Stuart Lloyd. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

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.

Pierre Marion, Raphaël Berthier, Gérard Biau, and Claire Boyer. Attention layers provably solve single-location regression, 2024. URL `https://arxiv.org/abs/2410.01537`.

Lorenzo Noci, Chuning Li, Mufan Bill Li, Bobby He, Thomas Hofmann, Chris J. Maddison, and Daniel M. Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=PqfPjS9JRX`.

Mary Phuong and Marcus Hutter. Formal algorithms for transformers. *arXiv:2207.09238*, 2022.

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.

A Training dynamics in the degenerate case

In this section, we discuss the training behavior of the Transformer-based predictor $T^{\text{lin},\mu_0,\mu_1}$ in the context of clustering, assuming the data are drawn from the degenerate mixture model, where for $1 \leq \ell \leq L$,

$$X_\ell \sim \frac{1}{2}\delta_{\mu_0^\star} + \frac{1}{2}\delta_{\mu_1^\star}, \tag{P$_0$}$$

(the orthogonal centroids $\mu_0^\star$ and $\mu_1^\star$ still lie on the unit sphere). We emphasize that despite its apparent simplicity, this study framework is already sufficient to reveal some of the complexity inherent in the clustering task carried out by a self-attention layer.

A.1 Theoretical analysis

Since training is performed by minimizing the risk $\mathcal{R}$, the first steps of our analysis focus on studying the critical points and extrema of this risk.

Critical points and minimizers. First, we reparameterize the problem using the quantities

$$\kappa_0 \overset{\text{def}}{=} \langle \mu_0^\star, \mu_0 \rangle, \quad \kappa_1 \overset{\text{def}}{=} \langle \mu_1^\star, \mu_1 \rangle, \quad \eta_0 \overset{\text{def}}{=} \langle \mu_1, \mu_0^\star \rangle, \quad \eta_1 \overset{\text{def}}{=} \langle \mu_0, \mu_1^\star \rangle. \tag{13}$$

The scalar products κ_0 and κ_1 measure the alignment of the parameters μ_0 and μ_1 with the true centroids $\mu_0^\star$ and $\mu_1^\star$, respectively, while the scalar products η_0 and η_1 capture their orthogonality with the inverted centroids $\mu_1^\star$ and $\mu_0^\star$. The theoretical risk w.r.t. $\kappa_0, \kappa_1, \eta_0$ and η_1 reads as follows. The proof of this result and the following are given in Appendix B.

Proposition A.1. *Under the degenerate mixture model* (P$_0$)*, considering the attention-based predictor $T^{\text{lin},\mu_0,\mu_1}$ composed of two linear heads parameterized by μ_0 and μ_1, the theoretical risk $\mathcal{R}$ can be re-expressed as a function $\mathcal{R}^< : \mathbb{R}^4 \to \mathbb{R}$ such that $\mathcal{R}(\mu_0, \mu_1) = \mathcal{R}^<(\kappa_0, \kappa_1, \eta_0, \eta_1)$, where*

$$\begin{aligned} \mathcal{R}^<(\kappa_0, \kappa_1, \eta_0, \eta_1) = 1 - \lambda \frac{L+1}{L}(\kappa_0^2 + \kappa_1^2 + \eta_0^2 + \eta_1^2) + \lambda^2 \frac{L+3}{2L}([\kappa_0^2 + \eta_0^2]^2 + [\kappa_1^2 + \eta_1^2]^2) \\ + \lambda^2 \frac{L-1}{L}(\kappa_0\eta_1 + \kappa_1\eta_0)^2. \end{aligned} \tag{14}$$

In addition, if $(\mu_0, \mu_1) \in (\mathbb{S}^{d-1})^2$ are prescribed to the unit sphere, then $\text{dom}(\mathcal{R}^<) = [-1, 1]^4$.

Remark A.2. *After a direct computation, we note that the critical points of the risk $\mathcal{R}$ correspond to those of its reparameterized version, $\mathcal{R}^<$, i.e.,*

$$(\mu_0, \mu_1) \in \text{crit}(\mathcal{R}) \iff (\kappa_0, \kappa_1, \eta_0, \eta_1) \in \text{crit}(\mathcal{R}^<).$$

Proposition A.3 (Characterization of global minima)**.** *Consider $\mathcal{R}^< : \mathbb{R}^4 \to \mathbb{R}$ defined as in Proposition A.1 with $\lambda = \frac{L+1}{L+3}$, then a point $(\kappa_0, \kappa_1, \eta_0, \eta_1)$ belongs to* $\text{argmin}(\mathcal{R}^<)$ *if and only if*

$$\kappa_0^2 + \eta_0^2 = 1, \quad \kappa_1^2 + \eta_1^2 = 1, \quad \textit{and} \quad \kappa_0\eta_1 + \kappa_1\eta_0 = 0. \tag{15}$$

While the characterization can be made for any value of λ, choosing $\lambda = \frac{L+1}{L+3}$ simplifies the system by setting the first two conditions equal to 1. Moreover, this specific value provides a critical upper bound on the temperature parameter that guarantees recovery of the underlying centroids via risk minimization, as highlighted in the theorem below, and discussed in Remark B.8 in the appendices. The proof of this result in Appendix B also characterizes all critical points, beyond global minima.

Convergence analysis. From the characterization of the minima of $\mathcal{R}$ given in Proposition A.3, we observe that the points $(\kappa_0, \kappa_1, \eta_0, \eta_1)$ that saturate the first two equations (i.e., satisfy $\kappa_0^2 = \kappa_1^2 = 1$ and $\eta_0^2 = \eta_1^2 = 0$) correspond to global minimizers of the risk that also recover the centroids (up to a sign). However, in general, other global minimizers may exist that do not exhibit this saturation behavior and are therefore disconnected from centroid recovery. In the next result, we show that under appropriate initialization conditions, the (PGD) algorithm converges to the desired global minimum, which aligns with the clustering objective. To this end, we introduce the following manifold

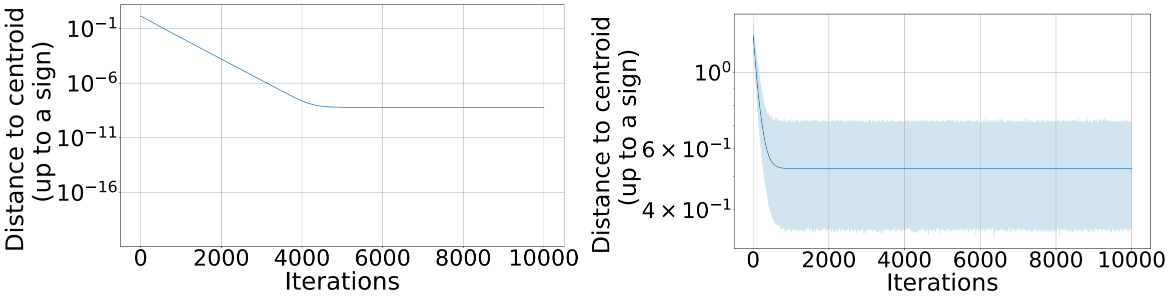
$$\tilde{\mathcal{M}} = \{(\mu_0, \mu_1) \in (\mathbb{S}^{d-1})^2 : \langle \mu_1^\star, \mu_0 \rangle = 0, \langle \mu_0^\star, \mu_1 \rangle = 0\}. \tag{16}$$

Theorem A.4. *Under the Dirac mixture model* (P$_0$)*, consider the attention-based predictor $T^{\text{lin},\mu_0,\mu_1}$ composed of two linear heads. Take $\lambda \in]0, \frac{L+1}{L+3}]$. Then there exists $\bar{\gamma} > 0$ such that for any stepsize $0 < \gamma < \bar{\gamma}$, and for a generic initialization $(\mu_0^0, \mu_1^0) \in \tilde{\mathcal{M}}$, the sequence of iterates (μ_0^k, μ_1^k), generated by* (PGD)*, converges to the centroids (up to a sign), i.e.,*

$$(\mu_0^k, \mu_1^k) \xrightarrow[k \to \infty]{} (\pm\mu_0^\star, \pm\mu_1^\star).$$

This result demonstrates that, despite the non-convexity of the objective function, the key and query row matrices of a linear attention layer trained via (PGD) align with the centroids of the underlying Dirac mixture. Although the setting is simplified, it already highlights the representational role of key and query matrices in attention-based learning, and serves as a foundation for addressing the more general case of Gaussian mixtures.

Note that the convergence is up to a sign, a consequence of the symmetry inherent in $H^{\text{lin},\mu}$. Nonetheless, this sign ambiguity does not affect the output of the attention layer. To resolve this ambiguity and identify the true centroids, one could compare likelihoods or perform a hard assignment



(a) Initialization on the manifold $\tilde{\mathcal{M}}$

(b) Initialization on the sphere $(\mathbb{S}^{d-1})^2$

Figure 4: Distance to centroids vs (PSGD) iterations for the minimization of $\mathcal{R}$, with data drawn from the degenerate case (P_0). 10 runs, 95% percentile intervals are plotted.

step, selecting the centroid that minimizes the total distance to all points within its assigned cluster. Besides, by generic initialization, we mean that the set of initializations $(\mu_0^0, \mu_1^0) \in \tilde{\mathcal{M}}$ for which (PGD) fails to recover the centroids is of Lebesgue measure zero with respect to $\tilde{\mathcal{M}}$.

Initialization on the manifold $\tilde{\mathcal{M}}$ relies on prior knowledge of the centroids, which may be impractical. While a theoretical analysis under generic initialization on the unit sphere would be of interest, it remains analytically intractable due to the complexity of the resulting dynamical system derived from (PGD). In the following, however, we present numerical experiments incorporating a regularization term that proves effective in solving the problem without initialization constraints.

A.2 Numerical experiments

In this section, we study the empirical convergence of the (PGD) iterates when the data follows the degenerate mixture model (P_0).

Results. Figure 4a clearly illustrates that when initialized on the manifold $\tilde{\mathcal{M}}$, the (PSGD) iterates, over the objective function $\mathcal{R}$, converge to the centroids, as established in Theorem A.4.

The situation differs outside the manifold, where numerical evidence shows that the Transformer parameters only partially align with the true centroids as shown in Figure 4b. In fact, we observe empirically that each parameter learns a mixture of both centroids. This indicates that the (PSGD) iterates may converge to optima that do not coincide with the underlying centroids. To mitigate this and better guide the learning process, we propose using a specific form of regularization:

$$r(\mu_0, \mu_1) \overset{\text{def}}{=} \mathbb{E}[\langle \mu_0, X_1 \rangle^2 \langle \mu_1, X_1 \rangle^2]. \tag{17}$$

Therefore, we train the attention-based predictor $H^{\mathrm{lin}, \mu_0, \mu_1}$ now by minimizing the regularized risk

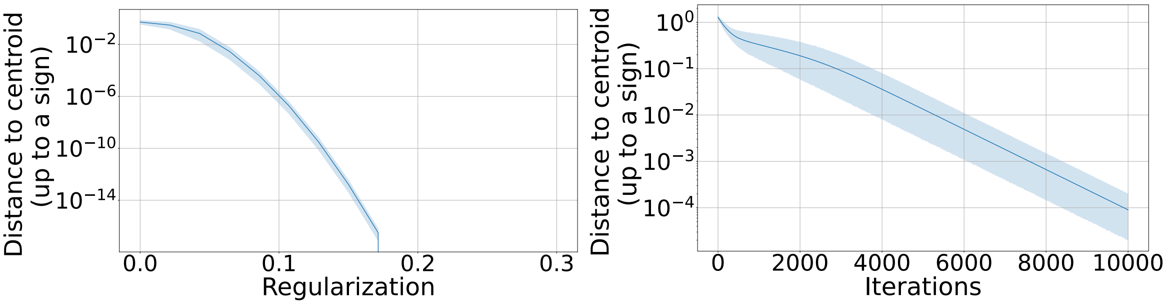
$$\min_{\mu_0, \mu_1 \in \mathbb{S}^{d-1}} \mathcal{R}^\rho(\mu_0, \mu_1) \qquad \text{with} \qquad \mathcal{R}^\rho(\mu_0, \mu_1) \overset{\text{def}}{=} \mathcal{R}(\mu_0, \mu_1) + \rho r(\mu_0, \mu_1), \tag{$\tilde{\mathcal{P}}_\rho$}$$

where $\rho > 0$ denotes the strength of the regularization. It can be rigorously shown that as ρ approaches 0, the minimizers of $\mathcal{R}^\rho$ converge to those of $\mathcal{R}$, exhibiting the saturation phenomenon, desirable to bolster the interpretability of the attention heads. We refer to Appendix B.3 for more details.

In Figure 5a, we observe that a relatively small regularization parameter (of the order of 10^{-1}) is sufficient to achieve centroid alignment, with numerical error below 10^{-14}. In Figure 5b, we fix the regularization parameter and observe that over the course of 10^4 iterations, the attention head parameters exhibit linear convergence towards the true centroids. This numerical experiment highlights the effectiveness of this form of regularization in enhancing the interpretability of attention heads—by promoting their disentanglement—in the context of mixture models.

B Proofs of Section A (degenerate case)

In this section, we present the postponed proofs that support and elaborate on the arguments developed in the main text. We begin by characterizing the critical points of the Dirac mixture risk, then proceed to a discussion on the effects of a regularization term in the Dirac setting. Finally, we outline the proofs of Theorems A.4 and 3.4, which constitute the main theoretical results of our work.



(a) Distance to centroids vs. regularization strength

(b) Distance to centroids vs iterations ($\rho = 0.1$).

Figure 5: Convergence analysis of the (PSGD) iterates for the minimization of the regularized risk $\mathcal{R}^\rho$, under the degenerate mixture case (P_0). 10 runs, 95% percentile intervals are plotted.

B.1 Proof of Proposition A.1 (expression of the risk in the degenerate case)

To facilitate the analysis that follows, we introduce the notation $e_k(\mu) \overset{\text{def}}{=} \lambda X_1^\top \mu\mu^\top X_k$, for $1 \le k \le L$, which allows us to write

$$\mathcal{R}(\mu_0, \mu_1) = \mathbb{E}\left[\left\|X_1 - \frac{2}{L}\sum_{k=1}^{L}(e_k(\mu_0) + e_k(\mu_1))X_k\right\|_2^2\right].$$

In what follows, we give an expression of the risk of an attention-based predictor, in the case where the data is distributed according to the Dirac mixture model (P_0). Then, the risk of an attention-based predictor $T^{\mathrm{lin},\mu_0,\mu_1}$ can be written as

$$\begin{aligned}
\mathcal{R}(\mu_0, \mu_1) &= \mathbb{E}\left[\left\|X_1 - \frac{2}{L}\sum_{k=1}^{L}(e_k(\mu_0) + e_k(\mu_1))X_k\right\|_2^2\right] \\
&= \mathbb{E}\left[\|X_1\|^2\right] - \frac{4}{L}\sum_{k=1}^{L}\mathbb{E}\big[\langle X_1, (e_k(\mu_0) + e_k(\mu_1))X_k\rangle\big] \\
&+ \frac{4}{L^2}\mathbb{E}\left[\left\|\sum_{k=1}^{L}(e_k(\mu_0) + e_k(\mu_1))X_k\right\|^2\right] \\
&= 1 - \frac{4}{L}\sum_{k=1}^{L}\mathbb{E}\big[\langle X_1, (e_k(\mu_0) + e_k(\mu_1))X_k\rangle\big] + \frac{4}{L^2}\sum_{k=1}^{L}\mathbb{E}[\|(e_k(\mu_0) + e_k(\mu_1))X_k\|^2] \\
&+ \frac{8}{L^2}\sum_{1 \le k < j \le L}\mathbb{E}\big[(e_k(\mu_0) + e_k(\mu_1))(e_j(\mu_0) + e_j(\mu_1))\langle X_k, X_j\rangle\big] \\
&= 1 - \frac{4}{L}\mathbb{E}[(e_1(\mu_0) + e_1(\mu_1))\|X_1\|^2] + \frac{4}{L^2}\mathbb{E}\left[\|(e_1(\mu_0) + e_1(\mu_1))X_1\|^2\right] \\
&+ \frac{8}{L^2}\sum_{k=2}^{L}\mathbb{E}\big[(e_1(\mu_0) + e_1(\mu_1))(e_k(\mu_0) + e_k(\mu_1))\langle X_1, X_k\rangle\big] \\
&\underbrace{- \frac{4}{L}\sum_{k=2}^{L}\mathbb{E}\big[\langle X_1, (e_k(\mu_0) + e_k(\mu_1))X_k\rangle\big]}_{\overset{\text{def}}{=}(I)} + \underbrace{\frac{4}{L^2}\sum_{k=2}^{L}\mathbb{E}\left[\|(e_k(\mu_0) + e_k(\mu_1))X_k\|^2\right]}_{\overset{\text{def}}{=}(II)} \\
&\underbrace{+ \frac{8}{L^2}\sum_{1 < k < j \le L}\mathbb{E}\big[(e_k(\mu_0) + e_k(\mu_1))(e_j(\mu_0) + e_j(\mu_1))\langle X_k, X_j\rangle\big]}_{\overset{\text{def}}{=}(III)} \\
&= 1 - \lambda\frac{2}{L}(\kappa_0^2 + \kappa_1^2 + \eta_0^2 + \eta_1^2) + \lambda^2\frac{2}{L^2}([\kappa_0^2 + \eta_0^2]^2 + [\kappa_1^2 + \eta_1^2]^2) \\
&+ \lambda^2\frac{2(L-1)}{L^2}([\kappa_0^2 + \eta_0^2]^2 + [\kappa_1^2 + \eta_1^2]^2) - (I) + (II) + (III).
\end{aligned}$$

(P: *here it is a bit strange to detail the computation for $(I)-(III)$ but not for other terms (where the computation is about as hard). Perhaps say somewhere that other terms are computed in a similar fashion?)* We compute (I) by conditioning on Z_1, Z_k,

$$\begin{aligned}
\mathbb{E}\big[(e_k(\mu_0)+e_k(\mu_1))\langle X_1, X_k\rangle\big] &= \mathbb{E}\big[\mathbb{E}[(e_k(\mu_0)+e_k(\mu_1))\langle X_1, X_k\rangle | Z_1, Z_k]\big] \\
&= \lambda\mathbb{E}[(\langle \mu^\star_{Z_1}, \mu_0\rangle\langle \mu^\star_{Z_k}, \mu_0\rangle + \langle \mu^\star_{Z_1}, \mu_1\rangle\langle \mu^\star_{Z_k}, \mu_1\rangle)\langle \mu^\star_{Z_1}, \mu^\star_{Z_k}\rangle] \\
&= \lambda\frac{\kappa_0^2+\kappa_1^2+\eta_0^2+\eta_1^2}{4}.
\end{aligned}$$

This leads to $(I) = \lambda\frac{L-1}{L}(\kappa_0^2+\kappa_1^2+\eta_0^2+\eta_1^2)$.

Similarly for (II), conditioning on Z_1, Z_k,

$$\begin{aligned}
\mathbb{E}\left[\|(e_k(\mu_0)+e_k(\mu_1))X_k\|^2\right] &= \mathbb{E}\left[\mathbb{E}[\|(e_k(\mu_0)+e_k(\mu_1))X_k\|^2 | Z_1, Z_k]\right] \\
&= \lambda^2\mathbb{E}[\|(\langle \mu^\star_{Z_1}, \mu_0\rangle\langle \mu^\star_{Z_k}, \mu_0\rangle + \langle \mu^\star_{Z_1}, \mu_1\rangle\langle \mu^\star_{Z_k}, \mu_1\rangle)\mu^\star_{Z_k}\|^2] \\
&= \lambda^2\frac{[\kappa_0^2+\eta_0^2]^2+[\kappa_1^2+\eta_1^2]^2}{4} + \lambda^2\frac{(\kappa_0\eta_1+\kappa_1\eta_0)^2}{2}.
\end{aligned}$$

Which gives $(II) = \lambda^2\frac{L-1}{L^2}([\kappa_0^2+\eta_0^2]^2+[\kappa_1^2+\eta_1^2]^2) + \lambda^2\frac{2(L-1)}{L^2}(\kappa_0\eta_1+\kappa_1\eta_0)^2$.

Finally, to compute (III), note that

$$\begin{aligned}
&\mathbb{E}\big[(e_k(\mu_0)+e_k(\mu_1))(e_j(\mu_0)+e_j(\mu_1))\langle X_k, X_j\rangle\big] \\
&\quad= \mathbb{E}\big[\mathbb{E}[(e_k(\mu_0)+e_k(\mu_1))(e_j(\mu_0)+e_j(\mu_1))\langle X_k, X_j\rangle | Z_1, Z_k, Z_j]\big] \\
&\quad= \lambda^2\mathbb{E}[(\langle \mu^\star_{Z_1}, \mu_0\rangle\langle \mu^\star_{Z_k}, \mu_0\rangle + \langle \mu^\star_{Z_1}, \mu_1\rangle\langle \mu^\star_{Z_k}, \mu_1\rangle)(\langle \mu^\star_{Z_1}, \mu_0\rangle\langle \mu^\star_{Z_j}, \mu_0\rangle + \langle \mu^\star_{Z_1}, \mu_1\rangle\langle \mu^\star_{Z_j}, \mu_1\rangle) \\
&\qquad\cdot \langle \mu^\star_{Z_k}, \mu^\star_{Z_j}\rangle] \\
&\quad= \lambda^2\frac{[\kappa_0^2+\eta_0^2]^2+[\kappa_1^2+\eta_1^2]^2}{8} + \lambda^2\frac{(\kappa_0\eta_1+\kappa_1\eta_0)^2}{4},
\end{aligned}$$

leading to $(III) = \lambda^2\frac{(L-1)(L-2)}{2L^2}[[\kappa_0^2+\eta_0^2]^2+[\kappa_1^2+\eta_1^2]^2+2(\kappa_0\eta_1+\kappa_1\eta_0)^2]$.

Putting everything together, we obtain that the risk can be written in terms of $\kappa_0, \kappa_1, \eta_0, \eta_1$, i.e., $\mathcal{R}(\mu_0, \mu_1) = \mathcal{R}^<(\kappa_0, \kappa_1, \eta_0, \eta_1)$, where:

$$\begin{aligned}
\mathcal{R}^< \overset{\text{def}}{=} 1 &- \lambda\left[\frac{2}{L}+\frac{L-1}{L}\right](\kappa_0^2+\kappa_1^2+\eta_0^2+\eta_1^2) \\
&+ \lambda^2\left[\frac{2}{L^2}+\frac{2(L-1)}{L^2}+\frac{L-1}{L^2}+\frac{(L-1)(L-2)}{2L^2}\right]([\kappa_0^2+\eta_0^2]^2+[\kappa_1^2+\eta_1^2]^2) \\
&+ \lambda^2\left[\frac{(L-1)(L-2)}{L^2}+\frac{2(L-1)}{L^2}\right](\kappa_0\eta_1+\kappa_1\eta_0)^2 \\
&= 1-\lambda\frac{L+1}{L}(\kappa_0^2+\kappa_1^2+\eta_0^2+\eta_1^2)+\lambda^2\frac{L+3}{2L}([\kappa_0^2+\eta_0^2]^2+[\kappa_1^2+\eta_1^2]^2) \\
&+ \lambda^2\frac{L-1}{L}(\kappa_0\eta_1+\kappa_1\eta_0)^2.
\end{aligned}$$

B.2 Proof of Proposition A.3 (critical points of the risk in the degenerate case)

Proposition B.1. *Consider $\mathcal{R}^< : \mathbb{R}^4 \to \mathbb{R}$ defined as in Proposition A.1 with $\lambda = \frac{L+1}{L+3}$, then we characterize its critical points by*

1. *The point $(0,0,0,0)$ is a local maximum.*
2. *The points $(\kappa_0, 0, \eta_0, 0)$, where $\kappa_0^2+\eta_0^2 = 1$, and $(0, \kappa_1, 0, \eta_1)$, where $\kappa_1^2+\eta_1^2 = 1$, are strict saddle points.*
3. *The points $(\kappa_0, \kappa_1, \kappa_1, \kappa_0)$ and $(\kappa_0, \kappa_1, -\kappa_1, -\kappa_0)$, where $\kappa_0^2+\kappa_1^2 = \frac{L+3}{2(L+1)}$, are strict saddle points.*

4. $(\kappa_0, \kappa_1, \eta_0, \eta_1)$ *belongs to* $\operatorname{argmin}(\mathcal{R}^<)$ *if and only if:*

$$\begin{cases} \kappa_0^2 + \eta_0^2 & = 1, \\ \kappa_1^2 + \eta_1^2 & = 1, \\ \kappa_0\eta_1 + \kappa_1\eta_0 & = 0. \end{cases} \tag{18}$$

Proof. Let us define $\zeta_0 = (\kappa_0, \eta_0), \zeta_1 = (\eta_1, \kappa_1)$, then there exists a function $\mathcal{R}^{<<} : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$, such that $\mathcal{R}^<(\kappa_0, \kappa_1, \eta_0, \eta_1) = \mathcal{R}^{<<}(\zeta_0, \zeta_1)$, in fact, let us define $A = \frac{(L+1)^2}{L(L+3)}, B = \frac{(L+1)^2}{2L(L+3)}, C = \frac{(L+1)^2(L-1)}{L(L+3)^2}$, then with the value of λ defined in the proposition, we obtain

$$\mathcal{R}^{<<}(\zeta_0, \zeta_1) = 1 - A(\|\zeta_0\|^2 + \|\zeta_1\|^2) + B(\|\zeta_0\|^4 + \|\zeta_1\|^4) + C\langle\zeta_0, \zeta_1\rangle^2.$$

To analyze its critical points, we take the partial derivatives,

$$\begin{aligned} \nabla_{\zeta_0}\mathcal{R}^{<<}(\zeta_0, \zeta_1) &= -2A\zeta_0 + 4B\|\zeta_0\|^2\zeta_0 + 2C\langle\zeta_0, \zeta_1\rangle\zeta_1, \\ \nabla_{\zeta_1}\mathcal{R}^{<<}(\zeta_0, \zeta_1) &= -2A\zeta_1 + 4B\|\zeta_1\|^2\zeta_1 + 2C\langle\zeta_0, \zeta_1\rangle\zeta_0. \end{aligned}$$

And also, we compute its Hessian, we define

$$\begin{aligned} \nabla^2_{\zeta_0,\zeta_0}\mathcal{R}^{<<}(\zeta_0, \zeta_1) &= -2AI_2 + 4B(2\zeta_0\zeta_0^\top + \|\zeta_0\|^2 I_2) + 2C\zeta_1\zeta_1^\top, \\ \nabla^2_{\zeta_0,\zeta_1}\mathcal{R}^{<<}(\zeta_0, \zeta_1) &= 2C(\zeta_0\zeta_1^\top + \zeta_0^\top\zeta_1 I_2), \\ \nabla^2_{\zeta_1,\zeta_0}\mathcal{R}^{<<}(\zeta_0, \zeta_1) &= 2C(\zeta_1\zeta_0^\top + \zeta_0^\top\zeta_1 I_2), \\ \nabla^2_{\zeta_1,\zeta_1}\mathcal{R}^{<<}(\zeta_0, \zeta_1) &= -2AI_2 + 4B(2\zeta_1\zeta_1^\top + \|\zeta_1\|^2 I_2) + 2C\zeta_0\zeta_0^\top \end{aligned}$$

Then the Hessian will be defined by

$$\nabla^2\mathcal{R}^{<<}(\zeta_0, \zeta_1) = \begin{pmatrix} \nabla^2_{\zeta_0,\zeta_0}\mathcal{R}^{<<}(\zeta_0, \zeta_1) & \nabla^2_{\zeta_0,\zeta_1}\mathcal{R}^{<<}(\zeta_0, \zeta_1) \\ \nabla^2_{\zeta_1,\zeta_0}\mathcal{R}^{<<}(\zeta_1, \zeta_0) & \nabla^2_{\zeta_1,\zeta_1}\mathcal{R}^{<<}(\zeta_0, \zeta_1) \end{pmatrix} \tag{19}$$

To find the critical points, we solve the following system of equations:

$$\begin{aligned} -2A\zeta_0 + 4B\|\zeta_0\|^2\zeta_0 + 2C\langle\zeta_0, \zeta_1\rangle\zeta_1 = 0, \\ -2A\zeta_1 + 4B\|\zeta_1\|^2\zeta_1 + 2C\langle\zeta_0, \zeta_1\rangle\zeta_0 = 0. \end{aligned} \tag{20}$$

$(0, 0)$ is a local maximum. We see that a trivial solution to this system is $(\zeta_0, \zeta_1) = (0, 0)$, and replacing into the Hessian matrix, we see directly that this point is a local maximum.

$(0, \zeta_1), (\zeta_0, 0)$ are strict saddle points. We check the case when $\zeta_0 = 0, \zeta_1 \neq 0$, then we need to solve

$$-2A\zeta_1 + 4B\|\zeta_1\|^2\zeta_1 = 0.$$

Since $\zeta_1 \neq 0$, this forces $-2A + 4B\|\zeta_1\|^2 = 0$. Replacing $(0, \zeta_1)$ into the Hessian matrix gives us

$$\nabla^2\mathcal{R}^{<<}(0, \zeta_1) = \begin{pmatrix} -2AI_2 + 2C\zeta_1\zeta_1^\top & 0 \\ 0 & 8B\zeta_1\zeta_1^\top \end{pmatrix}.$$

And $\operatorname{eig}(\nabla^2\mathcal{R}^{<<}(0, \zeta_1)) = \operatorname{eig}(-2AI_2 + 2C\zeta_1\zeta_1^\top) \cup \operatorname{eig}(8B\zeta_1\zeta_1^\top)$, where eig is the set of eigenvalues of a matrix. We have that

$$\begin{aligned} \operatorname{eig}(-2AI_2 + 2C\zeta_1\zeta_1^\top) &= \{-2A, 2(C - A)\}, \\ \operatorname{eig}(8B\zeta_1\zeta_1^\top) &= \{0, 8B\|\zeta_1\|^2\} \end{aligned}$$

where we have used that $8B = 4A$. We also note that $C - A < 0$, then we conclude there are 2 negative eigenvalues and 1 positive eigenvalue, concluding that these points are strict saddle points, due to symmetry we conclude the same for the points of the form $(\zeta_0, 0)$ for $\|\zeta_0\|^2 = 1$.

Non-trivial critical points. We will first show that the critical points that are of the form (ζ_0, ζ_1) for $\zeta_0 \neq 0, \zeta_1 \neq 0$ necessarily satisfy $\|\zeta_0\| = \|\zeta_1\| \neq 0$, multiplying the first equation of (20) by ζ_0 and the second by ζ_1, then subtracting both resulting expressions we obtain $4B(\|\zeta_1\|^4 - \|\zeta_0\|^4) = 2A(\|\zeta_1\|^2 - \|\zeta_0\|^2)$, and then

$$(\|\zeta_0\|^2 - \|\zeta_1\|^2)(-1 + \|\zeta_0\|^2 + \|\zeta_1\|^2) = 0,$$

thus either $\|\zeta_0\| = \|\zeta_1\|$ and we get the first claim, or $\|\zeta_0\|^2 + \|\zeta_1\|^2 = 1$, in this second case we divide in two subcases:

- Let us assume that $\langle \zeta_0, \zeta_1 \rangle = 0$, then multiplying the first equation of (20) by ζ_0 and the second equation by ζ_1, we get that $\|\zeta_0\|^2 = \|\zeta_1\|^2 = 1$, which is a contradiction since we are in the case where $\|\zeta_0\|^2 + \|\zeta_1\|^2 = 1$.
- Therefore $\langle \zeta_0, \zeta_1 \rangle \neq 0$, we multiply the first equation of (20) by ζ_1 and second by ζ_0, after dividing by $2\langle \zeta_0, \zeta_1 \rangle$ we get that

  $$\begin{aligned} -A + A\|\zeta_0\|^2 + C\langle \zeta_0, \zeta_1 \rangle = 0 \\ -A + A\|\zeta_1\|^2 + C\langle \zeta_0, \zeta_1 \rangle = 0. \end{aligned}$$

  Substracting both equations we get that $\|\zeta_0\| = \|\zeta_1\|$.

So we get that necessarily $\|\zeta_0\| = \|\zeta_1\| = r > 0$. Then the equation (20) becomes

$$\begin{aligned} A(r^2 - 1)\zeta_0 + C\langle \zeta_0, \zeta_1 \rangle \zeta_1 = 0, \\ A(r^2 - 1)\zeta_1 + C\langle \zeta_0, \zeta_1 \rangle \zeta_0 = 0. \end{aligned} \tag{21}$$

Adding/substracting both equations we get

$$(\zeta_0 \pm \zeta_1)[A(r^2 - 1) \pm C\langle \zeta_0, \zeta_1 \rangle] = 0.$$

$(\zeta_0, \pm\zeta_0)$ are strict saddle points. In the case where $\zeta_0 = \pm\zeta_1$, by (21) we get that

$$A(r^2 - 1)\zeta_0 + Cr^2\zeta_0 = 0,$$

then $r^2 = \frac{A}{A+C} = \frac{L+3}{2(L+1)}$. Replacing this point on the Hessian matrix $\nabla^2 \mathcal{R}^{<<}(\zeta_0, \pm\zeta_0)$, where $\|\zeta_0\| = r$,

$$\begin{aligned} \nabla^2 \mathcal{R}^{<<}(\zeta_0, \zeta_0) &= \begin{pmatrix} 2A(r^2 - 1)I_2 + 2(2A + C)\zeta_0\zeta_0^\top & 2C(r^2 I_2 + \zeta_0\zeta_0^\top) \\ 2C(r^2 I_2 + \zeta_0\zeta_0^\top) & 2A(r^2 - 1)I_2 + 2(2A + C)\zeta_0\zeta_0^\top \end{pmatrix}, \\ \nabla^2 \mathcal{R}^{<<}(\zeta_0, -\zeta_0) &= \begin{pmatrix} 2A(r^2 - 1)I_2 + 2(2A + C)\zeta_0\zeta_0^\top & -2C(r^2 I_2 + \zeta_0\zeta_0^\top) \\ -2C(r^2 I_2 + \zeta_0\zeta_0^\top) & 2A(r^2 - 1)I_2 + 2(2A + C)\zeta_0\zeta_0^\top \end{pmatrix}. \end{aligned}$$

We can do a similar analysis as before and conclude that this matrix has positive and negative eigenvalues.

Characterization of global minima. If $\zeta_0 \neq \pm\zeta_1$ and $\|\zeta_0\| = \|\zeta_1\| = r > 0$, then both vectors are linearly independent, thus the first equation of (21) is only possible when $r^2 = 1$ and $\langle \zeta_0, \zeta_1 \rangle = 0$, in which case we have to analyze the points (ζ_0, ζ_1) such that $\langle \zeta_0, \zeta_1 \rangle = 0$ and $\|\zeta_0\|^2 = \|\zeta_1\|^2 = 1$, we replace these points on the Hessian matrix and this gives us

$$\nabla^2 \mathcal{R}^{<<}(\zeta_0, \zeta_1) = \begin{pmatrix} 4A\zeta_0\zeta_0^\top + 2C\zeta_1\zeta_1^\top & 2C\zeta_0\zeta_1^\top \\ 2C\zeta_1\zeta_0^\top & 4A\zeta_1\zeta_1^\top + 2C\zeta_0\zeta_0^\top \end{pmatrix}.$$

A direct computation of the eigenvalues with eigenvectors (ζ_1, ζ_0) and $(\zeta_1, -\zeta_0)$ gives us that all the eigenvalues are positive in this case, since $\mathcal{R}^{<<}$ is coercive, these points are in fact global minima. □

B.3 Discussion on regularization

In order to solve the clustering problem in the degenerate case, we train the attention-based predictor $H^{\text{lin},\mu_0,\mu_1}$ now by minimizing the regularized risk

$$\min_{\mu_0,\mu_1 \in \mathbb{S}^{d-1}} \mathcal{R}^\rho(\mu_0, \mu_1) \qquad \text{with} \qquad \mathcal{R}^\rho(\mu_0, \mu_1) \overset{\text{def}}{=} \mathcal{R}(\mu_0, \mu_1) + \rho r(\mu_0, \mu_1), \tag{$\tilde{\mathcal{P}}_\rho$}$$

where $r(\mu_0,\mu_1) = \mathbb{E}[\langle \mu_0, X_1\rangle^2 \langle \mu_1, X_1\rangle^2]$, and $\rho > 0$ denotes the strength of the regularization.

It is direct to check that there exists $r^< : \mathbb{R}^4 \to \mathbb{R}$, such that $r(\mu_0,\mu_1) = r^<(\kappa_0,\kappa_1,\eta_0,\eta_1)$ according to the notation defined in (13), and $r^<(\kappa_0,\kappa_1,\eta_0,\eta_1) = \frac{1}{2}(\kappa_0^2\eta_0^2 + \kappa_1^2\eta_1^2)$. We define the following optimization problem

$$\min_{\kappa_0,\kappa_1,\eta_0,\eta_1 \in [-1,1]} \mathcal{R}^<(\kappa_0,\kappa_1,\eta_0,\eta_1) + \rho r^<(\kappa_0,\kappa_1,\eta_0,\eta_1), \tag{$\tilde{\mathcal{P}}_\rho^<$}$$

where $\mathcal{R}^<$ is defined in Proposition A.1. Since $\mathcal{R}^<$ and $r^<$ are coercive, we apply Attouch (1996, Theorem 2.1) to conclude that if $u_\rho \in [-1,1]^4$ is a solution of $(\tilde{\mathcal{P}}_\rho^<)$, then every limit point $\hat{u}$ of u_ρ, when $\rho \to 0$, satisfies that:

$$\begin{cases} r^<(\hat{u}) \le r^<(v), & \text{for every } v \in \operatorname{argmin} \mathcal{R}^<, \\ \hat{u} \in \operatorname{argmin} \mathcal{R}^<. \end{cases}$$

Due to the geometry of $r^<$ and the characterization of $\operatorname{argmin} \mathcal{R}^<$ we got in Proposition 15, we obtain that if $\hat{u} = (\hat{\kappa_0}, \hat{\kappa_1}, \hat{\eta_0}, \hat{\eta_1})$, then

$$\begin{cases} \hat{\kappa}_0^2 = 1, \hat{\kappa}_1^2 = 1, \hat{\eta}_0^2 = 0, \hat{\eta}_1^2 = 0, & \text{or} \\ \hat{\eta}_0^2 = 1, \hat{\eta}_1^2 = 1, \hat{\kappa}_0^2 = 0, \hat{\kappa}_1^2 = 0. \end{cases} \tag{22}$$

Then the optimal solution for the regularized problem when $\rho \to 0$ achieves a saturation effect, corresponding to global minimizers that recover the centroids. However, due to the non-convex nature of the problem, it is not guaranteed a priori that PGD on $(\tilde{\mathcal{P}}_\rho^<)$ will converge to the desired solution. A possible direction of analysis is to study the dynamics of PGD in the limit where $\rho \to 0$. We know from Proposition B.1 that the only global minimizers of the unregularized problem lie on a manifold, so we expect that PGD converges to this manifold, before evolving on the manifold due to the regularization term, to converge to the minimizers given by (22). Technically, this dynamics could be studied by using two-timescale tools, e.g. similar to Marion and Berthier (2023) and references therein. We leave this analysis for future work.

B.4 Outline of the proof of Theorem A.4

The results in this subsection lead to Theorem A.4. Their proofs are deferred to Subsection C.4, where a generalization of the same lemmas and propositions is established.

Lemma B.2. *At a point $(\kappa_0,\kappa_1,\eta_0,\eta_1)$ such that $\eta_0 = \eta_1 = 0$, we have $\partial_{\eta_0}\mathcal{R}^< = \partial_{\eta_1}\mathcal{R}^< = 0$.*

Proof. We have that

$$\mathcal{R}^<(\kappa_0,\kappa_1,\eta_0,\eta_1) = \mathcal{R}^<(\kappa_0,\kappa_1,-\eta_0,-\eta_1,).$$

Taking the partial derivative in η_0, we get

$$\partial_{\eta_0}\mathcal{R}^<(\kappa_0,\kappa_1,\eta_0,\eta_1) = -\partial_{\eta_0}\mathcal{R}^<(\kappa_0,\kappa_1,-\eta_0,-\eta_1).$$

At a point such that $\eta_0 = \eta_1 = 0$, this gives $\partial_{\eta_0}\mathcal{R}^<(\kappa_0,\kappa_1,0,0) = -\partial_{\eta_0}\mathcal{R}^<(\kappa_0,\kappa_1,0,0)$, therefore $\partial_{\eta_0}\mathcal{R}^<(\kappa_0,\kappa_1,0,0) = 0$, the proof for $\partial_{\eta_1}\mathcal{R}^<$ is analogous. □

Lemma B.3. *The manifold $\tilde{\mathcal{M}}$ is invariant under* (PGD) *dynamics, this is if $(\mu_0^k,\mu_1^k) \in \tilde{\mathcal{M}}$, then $(\mu_0^{k+1},\mu_1^{k+1}) \in \tilde{\mathcal{M}}$.*

Thus, if we initialize (PGD) in $\tilde{\mathcal{M}}$, then we saturate equations (15) in the aforementioned sense, i.e., we retrieve the centroids (up to a sign), the following results will formalize this fact.

Lemma B.4. *When initialized on the manifold $\tilde{\mathcal{M}}$, the iterations generated by* (PGD) *can be reformulated as follows:*

$$(\kappa_0^{k+1},\kappa_1^{k+1}) = \varphi(\kappa_0^k,\kappa_1^k), \tag{23}$$

where $\tilde{\varphi} : [-1,1]^2 \to [-1,1]^2$ is given by

$$\tilde{\varphi}(\kappa_0,\kappa_1) = \left(\frac{\kappa_0 - \gamma(\partial_{\kappa_0}\mathcal{R}^<(\kappa_0,\kappa_1))(1-\kappa_0^2)}{\sqrt{1+\gamma^2(\partial_{\kappa_0}\mathcal{R}^<(\kappa_0,\kappa_1))^2(1-\kappa_0^2)}}, \frac{\kappa_1 - \gamma(\partial_{\kappa_1}\mathcal{R}^<(\kappa_0,\kappa_1))(1-\kappa_1^2)}{\sqrt{1+\gamma^2(\partial_{\kappa_1}\mathcal{R}^<(\kappa_0,\kappa_1))^2(1-\kappa_1^2)}} \right),$$

and $\mathcal{R}^<(\kappa_0,\kappa_1) \overset{\text{def}}{=} \mathcal{R}^<(\kappa_0,\kappa_1,0,0)$.

We explicit that

$$\mathcal{R}^{<}(\kappa_0, \kappa_1) = 1 - \lambda \frac{L+1}{L}(\kappa_0^2 + \kappa_1^2) + \lambda^2 \frac{L+3}{2L}(\kappa_0^4 + \kappa_1^4). \tag{24}$$

In the following propositions, we consider $\mathcal{R}^{<} : [-1,1]^2 \to \mathbb{R}_+$ defined as in (24) with $\lambda \in]0, \frac{L+1}{L+3}]$.

Proposition B.5. *Let $(\mu_0^0, \mu_1^0) \in \tilde{\mathcal{M}}$, consider* (23) *with initial conditions $\kappa_0^0 = \langle \mu_0^0, \mu_0^\star \rangle, \kappa_1^0 = \langle \mu_1^0, \mu_1^\star \rangle$. Then there exists $\bar{\gamma} > 0$ such that for every $0 < \gamma < \bar{\gamma}$, the risk $\mathcal{R}^{<}$ is decreasing along the iterates of* (23)*. Besides, the distance between successive iterates tends to zero, and, if $(\kappa_0^\star, \kappa_1^\star)$ is an accumulation point of the sequence of iterates $(\kappa_0^k, \kappa_1^k)_{k \in \mathbb{N}}$, then*

$$(1 - (\kappa_0^\star)^2)\partial_{\kappa_0} \mathcal{R}^{<}(\kappa_0^\star, \kappa_1^\star) = 0, \quad (1 - (\kappa_1^\star)^2)\partial_{\kappa_1} \mathcal{R}^{<}(\kappa_0^\star, \kappa_1^\star) = 0. \tag{25}$$

Proposition B.6. *The points $(\kappa_0, \kappa_1) \in [-1,1]^2$ satisfying* (25) *belong to the set*

$$\tilde{\mathscr{C}} \overset{\text{def}}{=} \{(\pm 1, \pm 1)^2, (0, \pm 1), (\pm 1, 0), (0,0)\}.$$

Proposition B.7. *The fixed points of the dynamic can be classified as follows:*

1. *The points $(\kappa_0, \kappa_1) = (\pm 1, \pm 1)$ are global minima of $\mathcal{R}^{<}$ on $[-1,1]^2$.*
2. *The points $(\kappa_0, \kappa_1) = (0, \pm 1)$ and $(\pm 1, 0)$ are strict saddle points of $\mathcal{R}^{<}$ on $[-1,1]^2$.*
3. *The point $(\kappa_0, \kappa_1) = (0,0)$ is a global maxima of $\mathcal{R}^{<}$ on $[-1,1]^2$.*

Remark B.8. *Note that the specific characterization of the global minima of $\mathcal{R}^{<}$ was valid only for $\lambda = \frac{L+1}{L+3}$. However, when restricting the analysis to the manifold $\tilde{\mathcal{M}}$ and considering $\lambda \in]0, \frac{L+1}{L+3}[$, the global minima lie outside the domain $[-1,1]^2$. As a result, due to the structure of the update rule in* (PGD)*, the extreme points $(\pm 1, \pm 1)$ of $[-1,1]^2$ become fixed points of the algorithm and serve as global minimizers of $\mathcal{R}^{<}$.*

Proposition B.9. *Consider the context of Proposition B.5, then there exists $\bar{\gamma} > 0$ such that for any stepsize $0 < \gamma < \bar{\gamma}$, the iterates $(\kappa_0^k, \kappa_1^k)_{k \in \mathbb{N}}$ generated by* (23) *converge to an element of $\tilde{\mathscr{C}}$.*

Proposition B.10. *Consider the context of Proposition B.5, then there exists $\bar{\gamma} > 0$ such that for any stepsize $0 < \gamma < \bar{\gamma}$, the set of initializations such that the iterates $(\kappa_0^k, \kappa_1^k)_{k \in \mathbb{N}}$ generated by* (23) *converge to $(0, \pm 1), (\pm 1, 0)$ or $(0,0)$ has Lebesgue measure zero.*

C Proofs of Section 3 (Gaussian mixture model)

C.1 Proof of Proposition 3.1 (expression of the risk in the non-degenerate case).

Recall the notation $e_k(\mu) \overset{\text{def}}{=} \lambda X_1^\top \mu \mu^\top X_k$, for $1 \le k \le L$, which allows us to write

$$\mathcal{R}(\mu_0, \mu_1) = \mathbb{E}\left[\left\| X_1 - \frac{2}{L} \sum_{k=1}^{L} (e_k(\mu_0) + e_k(\mu_1)) X_k \right\|_2^2 \right].$$

Under the Gaussian mixture model, we are going to show that the risk

$$\mathcal{R}(\mu_0, \mu_1) = \mathbb{E}\left[\| X_1 - (H^{\mu_0} + H^{\mu_1})(\mathbb{X})_1 \|_2^2 \right]$$

admits a closed-form representation in terms of elementary functions. It holds that

$$\begin{aligned} \mathcal{R}(\mu_0, \mu_1) &= \mathbb{E}\left[\left\| X_1 - \frac{2}{L} \sum_{k=1}^{L} (e_k(\mu_0) + e_k(\mu_1)) X_k \right\|_2^2 \right] \\ &= \mathbb{E}\left[\|X_1\|^2 \right] - \frac{4}{L} \sum_{k=1}^{L} \mathbb{E}\big[\langle X_1, (e_k(\mu_0) + e_k(\mu_1)) X_k \rangle \big] \end{aligned}$$

[2] This notation means $(\pm 1, \pm 1) \overset{\text{def}}{=} \{(1,1), (1,-1), (-1,1), (-1,-1)\}$

$$
\begin{aligned}
&+ \frac{4}{L^2}\mathbb{E}\left[\Big\|\sum_{k=1}^{L}(e_k(\mu_0)+e_k(\mu_1))X_k\Big\|^2\right].\\
&= (1+d\sigma^2) - \frac{4}{L}\sum_{k=1}^{L}\mathbb{E}\big[\langle X_1,(e_k(\mu_0)+e_k(\mu_1))X_k\rangle\big]\\
&+ \frac{4}{L^2}\sum_{k=1}^{L}\mathbb{E}[\|(e_k(\mu_0)+e_k(\mu_1))X_k\|^2]\\
&+ \frac{8}{L^2}\sum_{1\le k<j\le L}\mathbb{E}\big[(e_k(\mu_0)+e_k(\mu_1))(e_j(\mu_0)+e_j(\mu_1))\langle X_k,X_j\rangle\big]\\
&= (1+d\sigma^2) - \underbrace{\frac{4}{L}\mathbb{E}[(e_1(\mu_0)+e_1(\mu_1))\|X_1\|^2]}_{(I_0)} + \underbrace{\frac{4}{L^2}\mathbb{E}\left[\|(e_1(\mu_0)+e_1(\mu_1))X_1\|^2\right]}_{(II_0)}\\
&\underbrace{+\frac{8}{L^2}\sum_{k=2}^{L}\mathbb{E}\big[(e_1(\mu_0)+e_1(\mu_1))(e_k(\mu_0)+e_k(\mu_1))\langle X_1,X_k\rangle\big]}_{\overset{\text{def}}{=}(III_0)}\\
&\underbrace{-\frac{4}{L}\sum_{k=2}^{L}\mathbb{E}\big[\langle X_1,(e_k(\mu_0)+e_k(\mu_1))X_k\rangle\big]}_{\overset{\text{def}}{=}(I)} + \underbrace{\frac{4}{L^2}\sum_{k=2}^{L}\mathbb{E}\left[\|(e_k(\mu_0)+e_k(\mu_1))X_k\|^2\right]}_{\overset{\text{def}}{=}(II)}\\
&\underbrace{+\frac{8}{L^2}\sum_{1<k<j\le L}\mathbb{E}\big[(e_k(\mu_0)+e_k(\mu_1))(e_j(\mu_0)+e_j(\mu_1))\langle X_k,X_j\rangle\big]}_{\overset{\text{def}}{=}(III)}\\
&= (1+d\sigma^2) - (I_0) + (II_0) + (III_0) - (I) + (II) + (III).
\end{aligned}
$$

We now proceed to compute each of the six terms. To compute (I_0), we can use Lemma E.2, since

$$
\begin{aligned}
&\mathbb{E}[(\langle X_1,\mu_0\rangle^2+\langle X_1,\mu_1\rangle^2)\|X_1\|^2]\\
&= \frac{1}{2}\big(\mathbb{E}[\langle X_1,\mu_0\rangle^2\|X_1\|^2|Z_1=0] + \mathbb{E}[(\langle X_1,\mu_0\rangle^2\|X_1\|^2|Z_1=1]\big)\\
&+ \frac{1}{2}\big(\mathbb{E}[\langle X_1,\mu_1\rangle^2\|X_1\|^2|Z_1=0] + \mathbb{E}[\langle X_1,\mu_1\rangle^2\|X_1\|^2|Z_1=1]\big)\\
&= \frac{1}{2}\left[(\kappa_0^2+\eta_0^2+\kappa_1^2+\eta_1^2)(1+\sigma^2(d+4)) + 2\sigma^2(1+\sigma^2(d+2))(\|\mu_0\|^2+\|\mu_1\|^2)\right]
\end{aligned}
$$

Then, $(I_0) = \frac{2\lambda}{L}\left[(\kappa_0^2+\eta_0^2+\kappa_1^2+\eta_1^2)(1+\sigma^2(d+4)) + 2\sigma^2(1+\sigma^2(d+2))(\|\mu_0\|^2+\|\mu_1\|^2)\right]$.

To compute (II_0), by defining $p_0(\mu_0,\mu_1,\mu^\star)$ as in Lemma E.3, we get

$$
\begin{aligned}
&\mathbb{E}[(\langle X_1,\mu_0\rangle^2+\langle X_1,\mu_1\rangle^2)^2\|X_1\|^2]\\
&= \frac{1}{2}\mathbb{E}[(\langle X_1,\mu_0\rangle^4 + 2\langle X_1,\mu_0\rangle^2\langle X_1,\mu_1\rangle^2 + \langle X_1,\mu_1\rangle^4)\|X_1\|^2|Z_1=0]\\
&+ \frac{1}{2}\mathbb{E}[(\langle X_1,\mu_0\rangle^4 + 2\langle X_1,\mu_0\rangle^2\langle X_1,\mu_1\rangle^2 + \langle X_1,\mu_1\rangle^4)\|X_1\|^2|Z_1=1]\\
&= \frac{1}{2}(p_0(\mu_0,\mu_0,\mu_0^\star) + 2p_0(\mu_0,\mu_1,\mu_0^\star) + p_0(\mu_1,\mu_1,\mu_0^\star))\\
&+ \frac{1}{2}(p_0(\mu_0,\mu_0,\mu_1^\star) + 2p_0(\mu_0,\mu_1,\mu_1^\star) + p_0(\mu_1,\mu_1,\mu_1^\star)).
\end{aligned}
$$

Then,

$$
\begin{aligned}
(II_0) &= \frac{4\lambda^2}{L^2}\mathbb{E}[(\langle X_1,\mu_0\rangle^2+\langle X_1,\mu_1\rangle^2)^2\|X_1\|^2]\\
&= \frac{2\lambda^2}{L^2}\left(p_0(\mu_0,\mu_0,\mu_0^\star + 2p_0(\mu_0,\mu_1,\mu_0^\star) + p_0(\mu_1,\mu_1,\mu_0^\star)\right)
\end{aligned}
$$

$$+\frac{2\lambda^2}{L^2}\left(p_0(\mu_0,\mu_0,\mu_1^\star)+2p_0(\mu_0,\mu_1,\mu_1^\star)+p_0(\mu_1,\mu_1,\mu_1^\star)\right).$$

To compute (III_0), by defining $p_1(\mu_0,\mu_1,\mu_{Z_1}^\star,\mu_{Z_2}^\star)$ as in Lemma E.4, we get

$$\begin{aligned}&\mathbb{E}[(\langle X_1,\mu_0\rangle^2+\langle X_1,\mu_1\rangle^2)(\langle X_1,\mu_0\rangle\langle X_2,\mu_0\rangle+\langle X_1,\mu_1\rangle\langle X_2,\mu_1\rangle)\langle X_1,X_2\rangle]\\&=\frac{1}{4}\sum_{(z_1,z_2)\in\{0,1\}^2}\Upsilon_1(z_1,z_2),\end{aligned}$$

where

$$\begin{aligned}\Upsilon_1(z_1,z_2)=\mathbb{E}[&(\langle X_1,\mu_0\rangle^2+\langle X_1,\mu_1\rangle^2)(\langle X_1,\mu_0\rangle\langle X_2,\mu_0\rangle+\langle X_1,\mu_1\rangle\langle X_2,\mu_1\rangle)\\&\cdot\langle X_1,X_2\rangle|Z_1=z_1,Z_2=z_2].\end{aligned}$$

And then

$$\begin{aligned}&\frac{1}{4}\sum_{(z_1,z_2)\in\{0,1\}^2}\Upsilon_1(z_1,z_2)\\&=\frac{1}{4}(p_1(\mu_0,\mu_0,\mu_0^\star,\mu_0^\star)+p_1(\mu_0,\mu_1,\mu_0^\star,\mu_0^\star)+p_1(\mu_1,\mu_0,\mu_0^\star,\mu_0^\star)+p_1(\mu_1,\mu_1,\mu_0^\star,\mu_0^\star))\\&+\frac{1}{4}(p_1(\mu_0,\mu_0,\mu_1^\star,\mu_0^\star)+p_1(\mu_0,\mu_1,\mu_1^\star,\mu_0^\star)+p_1(\mu_1,\mu_0,\mu_1^\star,\mu_0^\star)+p_1(\mu_1,\mu_1,\mu_1^\star,\mu_0^\star))\\&+\frac{1}{4}(p_1(\mu_0,\mu_0,\mu_0^\star,\mu_1^\star)+p_1(\mu_0,\mu_1,\mu_0^\star,\mu_1^\star)+p_1(\mu_1,\mu_0,\mu_0^\star,\mu_1^\star)+p_1(\mu_1,\mu_1,\mu_0^\star,\mu_1^\star))\\&+\frac{1}{4}(p_1(\mu_0,\mu_0,\mu_1^\star,\mu_1^\star)+p_1(\mu_0,\mu_1,\mu_1^\star,\mu_1^\star)+p_1(\mu_1,\mu_0,\mu_1^\star,\mu_1^\star)+p_1(\mu_1,\mu_1,\mu_1^\star,\mu_1^\star)).\end{aligned}$$

Consequently,

$$\begin{aligned}(III_0)&=8\lambda^2\frac{(L-1)}{L^2}\mathbb{E}[(\langle X_1,\mu_0\rangle^2+\langle X_1,\mu_1\rangle^2)(\langle X_1,\mu_0\rangle\langle X_2,\mu_0\rangle+\langle X_1,\mu_1\rangle\langle X_2,\mu_1\rangle)\langle X_1,X_2\rangle]\\&=2\lambda^2\frac{(L-1)}{L^2}\sum_{(z_1,z_2)\in\{0,1\}^2}\Upsilon_1(z_1,z_2).\end{aligned}$$

To compute (I), we can use Lemma E.5 to obtain:

$$\begin{aligned}&\mathbb{E}[(\langle X_1,\mu_0\rangle\langle X_2,\mu_0\rangle+\langle X_1,\mu_1\rangle\langle X_2,\mu_1\rangle)\langle X_1,X_2\rangle]\\&=\frac{1}{4}\mathbb{E}[(\langle X_1,\mu_0\rangle\langle X_2,\mu_0\rangle+\langle X_1,\mu_1\rangle\langle X_2,\mu_1\rangle)\langle X_1,X_2\rangle|Z_1=0,Z_2=0]\\&+\frac{1}{4}\mathbb{E}[(\langle X_1,\mu_0\rangle\langle X_2,\mu_0\rangle+\langle X_1,\mu_1\rangle\langle X_2,\mu_1\rangle)\langle X_1,X_2\rangle|Z_1=0,Z_2=1]\\&+\frac{1}{4}\mathbb{E}[(\langle X_1,\mu_0\rangle\langle X_2,\mu_0\rangle+\langle X_1,\mu_1\rangle\langle X_2,\mu_1\rangle)\langle X_1,X_2\rangle|Z_1=1,Z_2=0]\\&+\frac{1}{4}\mathbb{E}[(\langle X_1,\mu_0\rangle\langle X_2,\mu_0\rangle+\langle X_1,\mu_1\rangle\langle X_2,\mu_1\rangle)\langle X_1,X_2\rangle|Z_1=1,Z_2=1]\\&=\frac{1+4\sigma^2}{4}(\kappa_0^2+\eta_0^2+\kappa_1^2+\eta_1^2)+\sigma^4(\|\mu_0\|^2+\|\mu_1\|^2).\end{aligned}$$

Thus, $(I)=\lambda\frac{L-1}{L}[(\kappa_0^2+\eta_0^2+\kappa_1^2+\eta_1^2)(1+4\sigma^2)+4\sigma^4(\|\mu_0\|^2+\|\mu_1\|^2)]$.

To compute (II), by defining $p_2(\mu_0,\mu_1,\mu_{Z_1}^\star,\mu_{Z_2}^\star)$ as in Lemma E.6, we obtain:

$$\begin{aligned}&\mathbb{E}[(\langle X_1,\mu_0\rangle\langle X_2,\mu_0\rangle+\langle X_1,\mu_1\rangle\langle X_2,\mu_1\rangle)^2\|X_2\|^2]\\&=\frac{1}{4}\sum_{(z_1,z_2)\in\{0,1\}^2}\Upsilon_2(z_1,z_2),\end{aligned}$$

where

$$\Upsilon_2(z_1,z_2)=\mathbb{E}[(\langle X_1,\mu_0\rangle\langle X_2,\mu_0\rangle+\langle X_1,\mu_1\rangle\langle X_2,\mu_1\rangle)^2\|X_2\|^2|Z_1=z_1,Z_2=z_2].$$

And then

$$\begin{aligned}
&\frac{1}{4}\sum_{(z_1,z_2)\in\{0,1\}^2}\Upsilon_2(z_1,z_2)\\
&=\frac{1}{4}(p_2(\mu_0,\mu_0,\mu_0^\star,\mu_0^\star)+2p_2(\mu_0,\mu_1,\mu_0^\star,\mu_0^\star)+p_2(\mu_1,\mu_1,\mu_0^\star,\mu_0^\star))\\
&+\frac{1}{4}(p_2(\mu_0,\mu_0,\mu_0^\star,\mu_1^\star)+2p_2(\mu_0,\mu_1,\mu_0^\star,\mu_1^\star)+p_2(\mu_1,\mu_1,\mu_0^\star,\mu_1^\star))\\
&+\frac{1}{4}(p_2(\mu_0,\mu_0,\mu_1^\star,\mu_0^\star)+2p_2(\mu_0,\mu_1,\mu_1^\star,\mu_0^\star)+p_2(\mu_1,\mu_1,\mu_1^\star,\mu_0^\star))\\
&+\frac{1}{4}(p_2(\mu_0,\mu_0,\mu_1^\star,\mu_1^\star)+2p_2(\mu_0,\mu_1,\mu_1^\star,\mu_1^\star)+p_2(\mu_1,\mu_1,\mu_1^\star,\mu_1^\star)).
\end{aligned}$$

So we obtain,

$$\begin{aligned}
(II)&=\frac{4\lambda^2(L-1)}{L^2}\mathbb{E}[(\langle X_1,\mu_0\rangle\langle X_2,\mu_0\rangle+\langle X_1,\mu_1\rangle\langle X_2,\mu_1\rangle)^2\|X_2\|^2]\\
&=\frac{\lambda^2(L-1)}{L^2}\sum_{(z_1,z_2)\in\{0,1\}^2}\Upsilon_2(z_1,z_2).
\end{aligned}$$

Finally, to compute (III), by defining $p_3(\mu_0,\mu_1,\mu_{Z_1}^\star,\mu_{Z_2}^\star,\mu_{Z_3}^\star)$ as in Lemma E.7, we get

$$\begin{aligned}
&\mathbb{E}[(\langle X_1,\mu_0\rangle\langle X_2,\mu_0\rangle+\langle X_1,\mu_1\rangle\langle X_2,\mu_1\rangle)(\langle X_1,\mu_0\rangle\langle X_3,\mu_0\rangle+\langle X_1,\mu_1\rangle\langle X_3,\mu_1\rangle)\langle X_2,X_3\rangle]\\
&=\frac{1}{8}\sum_{(z_1,z_2,z_3)\in\{0,1\}^3}\Upsilon_3(z_1,z_2,z_3),
\end{aligned}$$

where

$$\begin{aligned}
\Upsilon_3(z_1,z_2,z_3)=\mathbb{E}\big[&(\langle X_1,\mu_0\rangle\langle X_2,\mu_0\rangle+\langle X_1,\mu_1\rangle\langle X_2,\mu_1\rangle)\\
&\cdot(\langle X_1,\mu_0\rangle\langle X_3,\mu_0\rangle+\langle X_1,\mu_1\rangle\langle X_3,\mu_1\rangle)\\
&\cdot\langle X_2,X_3\rangle\;\big|\;Z_1=z_1,Z_2=z_2,Z_3=z_3\big].
\end{aligned}$$

Observing that

$$\sum_{(z_1,z_2,z_3)\in\{0,1\}^3}\Upsilon_3(z_1,z_2,z_3)=\sum_{(a,b,c,d,e)\in\{0,1\}^5}p_3(\mu_a,\mu_b,\mu_c^\star,\mu_d^\star,\mu_e^\star),$$

we get

$$\begin{aligned}
(III)&=\frac{8\lambda^2(L-1)(L-2)}{2L^2}\cdot\frac{1}{8}\sum_{(z_1,z_2,z_3)\in\{0,1\}^3}\Upsilon_3(z_1,z_2,z_3)\\
&=\frac{\lambda^2(L-1)(L-2)}{2L^2}\sum_{(a,b,c,d,e)\in\{0,1\}^5}p_3(\mu_a,\mu_b,\mu_c^\star,\mu_d^\star,\mu_e^\star).
\end{aligned}$$

Finally, putting everything together, recalling the notation introduced in (7), and inspecting the formulas given by Lemmas from E.2 to E.7 allows to conclude.

C.2 Proof of Lemma 3.2 (expression of the risk on the manifold, non-degenerate case)

We provide in Lemma C.1 a more precise version of Lemma 3.2, with explicit constants.

Lemma C.1. *Define* $c_1(n)=1+n\sigma^2$ *and* $c_2(n)=1+\sigma^2(d+n)$*, then the risk* $\mathcal{R}^<(\kappa_0,\kappa_1)$ *restricted to* $\mathcal{M}$ *has the form*

$$\mathcal{R}^<(\kappa_0,\kappa_1)=A(\kappa_0^4+\kappa_1^4)+B(\kappa_0^2+\kappa_1^2)+C\kappa_0^2\kappa_1^2+D,$$

where

$$\begin{aligned}
A &= \frac{2\lambda^2}{L^2}c_2(8) + \frac{2\lambda^2(L-1)}{L^2}c_1(5) + \frac{\lambda^2(L-1)}{L^2}c_2(4) + \frac{\lambda^2(L-1)(L-2)}{2L^2}c_1(4).\\
B &= -\frac{2\lambda}{L}c_2(4) + \frac{16\lambda^2\sigma^2}{L^2}c_2(6) + \frac{8\lambda^2\sigma^2(L-1)}{L^2}c_1(6) - \frac{\lambda(L-1)}{L}c_1(4)\\
&\quad + \frac{4\lambda^2\sigma^2(L-1)}{L^2}c_2(3) + \frac{\lambda^2\sigma^2(L-1)(L-2)}{L^2}c_1(6).\\
C &= \frac{4\lambda^2\sigma^2(L-1)}{L^2}.\\
D &= c_1(d) - \frac{8\lambda\sigma^2}{L}c_2(2) + \frac{32\lambda^2\sigma^4}{L^2}c_2(4) + \frac{64\lambda^2\sigma^6(L-1)}{L^2}\\
&\quad - \frac{8\lambda\sigma^4(L-1)}{L} + \frac{8\lambda^2\sigma^4(L-1)}{L^2}c_2(2) + \frac{8\lambda^2\sigma^6(L-1)(L-2)}{L^2}.
\end{aligned}$$

Proof of Lemma C.1. Using the decomposition obtained in the proof of Proposition 3.1, after simple algebraic manipulation we get that on this manifold:

- $(I_0) = \frac{2\lambda}{L}[(\kappa_0^2+\kappa_1^2)(1+\sigma^2(d+4)) + 4\sigma^2(1+\sigma^2(d+2))]$.
- $(II_0) = \frac{2\lambda^2}{L^2}[(\kappa_0^4+\kappa_1^4)(1+\sigma^2(d+8)) + 8\sigma^2(\kappa_0^2+\kappa_1^2)(1+\sigma^2(d+6)) + 16\sigma^4(1+\sigma^2(d+4))]$.
- $(III_0) = 2\lambda^2\frac{(L-1)}{L^2}[(\kappa_0^4+\kappa_1^4)(1+5\sigma^2) + 4\sigma^2(\kappa_0^2+\kappa_1^2)(1+6\sigma^2) + 2\sigma^2\kappa_0^2\kappa_1^2 + 32\sigma^6]$.
- $(I) = \lambda\frac{L-1}{L}[(\kappa_0^2+\kappa_1^2)(1+4\sigma^2) + 8\sigma^4]$.
- $(II) = \lambda^2\frac{(L-1)}{L^2}[(\kappa_0^4+\kappa_1^4)(1+\sigma^2(d+4)) + 4\sigma^2(\kappa_0^2+\kappa_1^2)(1+\sigma^2(d+3)) + 8\sigma^4(1+\sigma^2(d+2))]$.
- $(III) = \lambda^2\frac{(L-1)(L-2)}{2L^2}[(\kappa_0^4+\kappa_1^4)(1+4\sigma^2) + 2\sigma^2(\kappa_0^2+\kappa_1^2)(1+6\sigma^2) + 16\sigma^6]$.

We conclude by noting that the risk restricted to this manifold is

$$\mathcal{R}^<(\kappa_0,\kappa_1) = (1+d\sigma^2) - (I_0) + (II_0) + (III_0) - (I) + (II) + (III),$$

and properly factorizing the terms. □

C.3 Proof of Proposition 3.3 (global minima of the risk, non-degenerate case).

In what follows we provide an extended version of Proposition 3.3 with explicit constant, together with its proof.

Proposition C.2. *Let us define*

$$\begin{aligned}
c_3(\sigma,L) &\overset{\text{def}}{=} 16\sigma^2c_2(6) + 8\sigma^2(L-1)c_1(6) + 4\sigma^2(L-1)c_2(3) + \sigma^2(L-1)(L-2)c_1(6) + 4c_2(8)\\
&\quad + 4(L-1)c_1(5) + 2(L-1)c_2(4) + (L-1)(L-2)c_1(4) + 4\sigma^2(L-1),
\end{aligned}$$

and consider $\mathcal{R}^<(\kappa_0,\kappa_1)$ with the following λ:

$$\lambda^\star(\sigma,L) = \frac{2Lc_2(4) + L(L-1)c_1(4)}{c_3(\sigma,L)}.$$

Then the points $(\pm1,\pm1)$ are global minimum of $\mathcal{R}^<(\kappa_0,\kappa_1)$.

Proof. Imposing first order conditions on $\mathcal{R}^<(\kappa_0,\kappa_1)$ from Lemma C.1, we obtain an explicit form of its critical points. From this expression, we note that the global minimum are the points $(\pm1,\pm1)$ if and only if $4A+2B+2C=0$. The function $\lambda\mapsto 2A(\lambda)+B(\lambda)+C(\lambda)$ is a quadratic which is negative for $0\le\lambda<\lambda^\star(\sigma,L)$, and vanishes at $\lambda=\lambda^\star(\sigma,L)$. □

C.4 Proof of Theorem 3.4

The proof of this result is built upon a series of intermediate results that progressively lead to the desired conclusion.

Lemma C.3. *At a point $(\kappa_0, \kappa_1, \eta_0, \eta_1, \xi)$ such that $\eta_0 = \eta_1 = \xi = 0$, we have $\partial_{\eta_0}\mathcal{R}^< = \partial_{\eta_1}\mathcal{R}^< = \partial_\xi \mathcal{R}^< = 0$.*

Proof. According to Proposition 3.1, we can directly obtain that

$$\mathcal{R}^<(\kappa_0, \kappa_1, \eta_0, \eta_1, \xi) = \mathcal{R}^<(\kappa_0, \kappa_1, -\eta_0, -\eta_1, -\xi).$$

Taking the partial derivative in η_0, we get

$$\partial_{\eta_0}\mathcal{R}^<(\kappa_0, \kappa_1, \eta_0, \eta_1, \xi) = -\partial_{\eta_0}\mathcal{R}^<(\kappa_0, \kappa_1, -\eta_0, -\eta_1, -\xi).$$

At a point such that $\eta_0 = \eta_1 = \xi = 0$, this gives $\partial_{\eta_0}\mathcal{R}^<(\kappa_0, \kappa_1, 0, 0, 0) = -\partial_{\eta_0}\mathcal{R}^<(\kappa_0, \kappa_1, 0, 0, 0)$, therefore $\partial_{\eta_0}\mathcal{R}^<(\kappa_0, \kappa_1, 0, 0, 0) = 0$, the proof for $\partial_{\eta_1}\mathcal{R}^<, \partial_\xi\mathcal{R}^<$ is analogous. □

Lemma C.4. *The manifold $\mathcal{M}$ is invariant under* (PGD) *dynamics, this is if $(\mu_0^k, \mu_1^k) \in \mathcal{M}$, then $(\mu_0^{k+1}, \mu_1^{k+1}) \in \mathcal{M}$.*

Proof. We apply the chain rule and Lemma C.3 to get:

$$\nabla_{\mu_0}\mathcal{R} = \partial_{\kappa_0}\mathcal{R}^< \mu_0^\star + \partial_{\eta_1}\mathcal{R}^< \mu_1^\star + \partial_\xi \mathcal{R}^< \mu_1 = \partial_{\kappa_0}\mathcal{R}^< \mu_0^\star, \tag{26}$$

$$\nabla_{\mu_1}\mathcal{R} = \partial_{\kappa_1}\mathcal{R}^< \mu_1^\star + \partial_{\eta_0}\mathcal{R}^< \mu_0^\star + \partial_\xi \mathcal{R}^< \mu_0 = \partial_{\kappa_1}\mathcal{R}^< \mu_1^\star. \tag{27}$$

We then follow the same ideas as in Marion et al. (2024, Lemma 4), where our Lemma B.2 takes the role of Marion et al. (2024, Lemma 14).

More concretely, let us consider $c_0 = \|\mu_0^k - \gamma(I_d - \mu_0^k(\mu_0^k)^\top)\nabla_{\mu_0}\mathcal{R}(\mu_0^k, \mu_1^k)\|_2$, and $c_1 = \|\mu_1^k - \gamma(I_d - \mu_1^k(\mu_1^k)^\top)\nabla_{\mu_1}\mathcal{R}(\mu_0^k, \mu_1^k)\|_2$, then recalling (PGD) updates, we have that if $(\mu_0^k, \mu_1^k) \in \mathcal{M}$, then

$$(\mu_1^\star)^\top \mu_0^{k+1} = \frac{(\mu_1^\star)^\top \mu_0^k - \gamma(\mu_1^\star)^\top (I_d - \mu_0^k(\mu_0^k)^\top)\partial_{\kappa_0}\mathcal{R}^<(\kappa_0^k, \kappa_1^k)\mu_0^\star}{c_0} = 0,$$
$$(\mu_0^\star)^\top \mu_1^{k+1} = \frac{(\mu_0^\star)^\top \mu_1^k - \gamma(\mu_0^\star)^\top (I_d - \mu_1^k(\mu_1^k)^\top)\partial_{\kappa_1}\mathcal{R}^<(\kappa_0^k, \kappa_1^k)\mu_1^\star}{c_1} = 0,$$

And

$$\begin{aligned}
&(\mu_1^{k+1})^\top \mu_0^{k+1} \\
&= \frac{(\mu_1^k)^\top \mu_0^k}{c_0 c_1} \\
&- \frac{\gamma(\partial_{\kappa_1}\mathcal{R}^<(\kappa_0^k, \kappa_1^k))((I_d - \mu_1^k(\mu_1^k)^\top)\mu_1^\star)^\top \mu_0^k - \gamma(\partial_{\kappa_0}\mathcal{R}^<(\kappa_0^k, \kappa_1^k))((I_d - \mu_0^k(\mu_0^k)^\top)\mu_0^\star)^\top \mu_1^k}{c_0 c_1} \\
&+ \frac{\gamma^2(\partial_{\kappa_0}\mathcal{R}^<(\kappa_0^k, \kappa_1^k))(\partial_{\kappa_1}\mathcal{R}^<(\kappa_0^k, \kappa_1^k))((I_d - \mu_0^k(\mu_0^k)^\top)\mu_0^\star)^\top (I_d - \mu_1^k(\mu_1^k)^\top)\mu_1^\star}{c_0 c_1} = 0,
\end{aligned}$$

where the last term is zero since

$$((I_d - \mu_0^k(\mu_0^k)^\top)\mu_0^\star)^\top (I_d - \mu_1^k(\mu_1^k)^\top)\mu_1^\star = 0.$$

Then $(\mu_0^{k+1}, \mu_1^{k+1}) \in \mathcal{M}$. □

Lemma C.5. *When initialized on the manifold $\mathcal{M}$, the iterations generated by* (PGD) *can be reformulated as follows:*

$$(\kappa_0^{k+1}, \kappa_1^{k+1}) = \varphi(\kappa_0^k, \kappa_1^k), \tag{28}$$

where $\varphi : [-1, 1]^2 \to [-1, 1]^2$ is given by

$$\varphi(\kappa_0, \kappa_1) = \left(\frac{\kappa_0 - \gamma(\partial_{\kappa_0}\mathcal{R}^<(\kappa_0, \kappa_1))(1 - \kappa_0^2)}{\sqrt{1 + \gamma^2(\partial_{\kappa_0}\mathcal{R}^<(\kappa_0, \kappa_1))^2(1 - \kappa_0^2)}}, \frac{\kappa_1 - \gamma(\partial_{\kappa_1}\mathcal{R}^<(\kappa_0, \kappa_1))(1 - \kappa_1^2)}{\sqrt{1 + \gamma^2(\partial_{\kappa_1}\mathcal{R}^<(\kappa_0, \kappa_1))^2(1 - \kappa_1^2)}} \right),$$

and $\mathcal{R}^<(\kappa_0, \kappa_1) \overset{\text{def}}{=} \mathcal{R}^<(\kappa_0, \kappa_1, 0, 0, 0)$ as in Lemma C.1.

Proof. By definition of the iterates and (26),(27), we have

$$
\begin{aligned}
\kappa_0^{k+1} = (\mu_0^{k+1})^\top \mu_0^\star &= \frac{\kappa_0^k - \gamma \partial_{\kappa_0} \mathcal{R}^<(\kappa_0,\kappa_1)((\mu_0^\star)^\top (I_d - \mu_0^k (\mu_0^k)^\top)\mu_0^\star)}{\sqrt{1+\gamma^2 (\partial_{\kappa_0}\mathcal{R}^<(\kappa_0,\kappa_1))^2 \|(I_d - \mu_0^k(\mu_0^k)^\top)\mu_0^\star\|_2^2}} \\
&= \frac{\kappa_0^k - \gamma \partial_{\kappa_0}\mathcal{R}^<(\kappa_0,\kappa_1)(1-(\kappa_0^k)^2)}{\sqrt{1+\gamma^2 \partial_{\kappa_0}\mathcal{R}^<(\kappa_0,\kappa_1)(1-(\kappa_0^k)^2)}}, \\
\kappa_1^{k+1} = (\mu_1^{k+1})^\top \mu_1^\star &= \frac{\kappa_1^k - \gamma \partial_{\kappa_1} \mathcal{R}^<(\kappa_0,\kappa_1)((\mu_1^\star)^\top (I_d - \mu_1^k (\mu_1^k)^\top)\mu_1^\star)}{\sqrt{1+\gamma^2 (\partial_{\kappa_1}\mathcal{R}^<(\kappa_0,\kappa_1))^2 \|(I_d - \mu_1^k(\mu_1^k)^\top)\mu_1^\star\|_2^2}} \\
&= \frac{\kappa_1^k - \gamma \partial_{\kappa_1}\mathcal{R}^<(\kappa_0,\kappa_1)(1-(\kappa_1^k)^2)}{\sqrt{1+\gamma^2 \partial_{\kappa_1}\mathcal{R}^<(\kappa_0,\kappa_1)(1-(\kappa_1^k)^2)}}.
\end{aligned}
$$

□

In the following propositions, we consider $\mathcal{R}^< : [-1,1]^2 \to \mathbb{R}_+$ defined as in Lemma C.1 with $\lambda \in]0, \lambda^*(\sigma, L)]$, where $\lambda^*(\sigma, L)$ is defined in Proposition C.2.

Proposition C.6. *Let $(\mu_0^0, \mu_1^0) \in \mathcal{M}$, consider (28) with initial conditions $\kappa_0^0 = \langle \mu_0^0, \mu_0^\star \rangle, \kappa_1^0 = \langle \mu_1^0, \mu_1^\star \rangle$. Then there exists $\bar{\gamma} > 0$ such that for every $0 < \gamma < \bar{\gamma}$, the risk $\mathcal{R}^<$ is decreasing along the iterates of (28). Besides, the distance between successive iterates tends to zero, and, if $(\kappa_0^\star, \kappa_1^\star)$ is an accumulation point of the sequence of iterates $(\kappa_0^k, \kappa_1^k)_{k\in\mathbb{N}}$, then*

$$
(1-(\kappa_0^\star)^2)\partial_{\kappa_0}\mathcal{R}^<(\kappa_0^\star,\kappa_1^\star) = 0, \quad (1-(\kappa_1^\star)^2)\partial_{\kappa_1}\mathcal{R}^<(\kappa_0^\star,\kappa_1^\star) = 0. \tag{29}
$$

Proof. The proof is identical to that of Marion et al. (2024, Proposition 8) and is therefore omitted. □

Proposition C.7. *The points $(\kappa_0, \kappa_1) \in [-1,1]^2$ satisfying (29) belong to the set*

$$
\mathscr{C} \overset{\text{def}}{=} \{(\pm 1, \pm 1), (0, \pm 1), (\pm 1, 0), (0,0)\}.
$$

Proof. We recall that by Lemma C.1, the risk $\mathcal{R}^<$ restricted to the manifold $\mathcal{M}$, has the following form

$$
\mathcal{R}^<(\kappa_0,\kappa_1) = A(\kappa_0^4 + \kappa_1^4) + B(\kappa_0^2 + \kappa_1^2) + C\kappa_0^2\kappa_1^2 + D.
$$

Then

$$
\begin{aligned}
\partial_{\kappa_0}\mathcal{R}^<(\kappa_0,\kappa_1) &= 4A\kappa_0^3 + 2B\kappa_0 + 2C\kappa_0\kappa_1^2, \\
\partial_{\kappa_1}\mathcal{R}^<(\kappa_0,\kappa_1) &= 4A\kappa_1^3 + 2B\kappa_1 + 2C\kappa_1\kappa_0^2.
\end{aligned}
$$

And we can rewrite equations (29) as

$$
\begin{aligned}
\kappa_0(1-\kappa_0^2)[2A\kappa_0^2 + B + C\kappa_1^2] = 0, \\
\kappa_1(1-\kappa_1^2)[2A\kappa_1^2 + B + C\kappa_0^2] = 0,
\end{aligned}
$$

Since each equation is a product of 3 terms, the general solution to this system occurs when at least in each equation is zero. By considering only the first two terms in each equation, we obtain the solution set $\{(\pm 1, \pm 1), (0, \pm 1), (\pm 1, 0), (0,0)\}$. Now we consider the case when $2A\kappa_0^2 + B + C\kappa_1^2 = 0$, this implies that:

- If $\kappa_1 = 0$, then $\kappa_0^2 = -\frac{B}{2A}$.
- If $\kappa_1^2 = 1$, then $\kappa_0^2 = -\frac{B+C}{2A}$.
- If $2A\kappa_1^2 + B + C\kappa_0^2 = 0$, then $2A\kappa_0^2 + C\kappa_1^2 = 2A\kappa_0^2 + C\kappa_1^2$, thus $(2A - C)(\kappa_0^2 - \kappa_1^2) = 0$. By inspection, $C < 2A$, hence we get $\kappa_0^2 = \kappa_1^2$ and $\kappa_0^2 = -\frac{B}{2A+C}$.

We note the following relation

$$
-\frac{B+C}{2A} < -\frac{B}{2A+C} < -\frac{B}{2A}.
$$

Further remark by inspecting the proof of Proposition C.2 that for $\lambda \in]0, \lambda^*(\sigma, L)]$, we have

$$
1 \le -\frac{B+C}{2A}.
$$

Thus the only possible solution when $2A\kappa_0^2 + B + C\kappa_1^2 = 0$ is $\kappa_1^2 = 1$ and $\kappa_0^2 = -\frac{B+C}{2A} = 1$ Putting everything together, the solution set is precisely

$$\{(\pm 1, \pm 1), (0, \pm 1), (\pm 1, 0), (0, 0)\}.$$

□

Proposition C.8. *The fixed points of the dynamic can be classified as follows:*

1. *The points $(\kappa_0, \kappa_1) = (\pm 1, \pm 1)$ are global minima of $\mathcal{R}^<$ on $[-1, 1]^2$.*
2. *The points $(\kappa_0, \kappa_1) = (0, \pm 1)$ and $(\pm 1, 0)$ are strict saddle points of $\mathcal{R}^<$ on $[-1, 1]^2$.*
3. *The point $(\kappa_0, \kappa_1) = (0, 0)$ is a global maxima of $\mathcal{R}^<$ on $[-1, 1]^2$.*

Proof. The claim is trivial, as $\mathcal{R}^<$ on $[-1, 1]^2$ is a simple quartic function and can be verified directly. □

Proposition C.9. *Consider the context of Proposition C.6, then there exists $\bar{\gamma} > 0$ such that for any stepsize $0 < \gamma < \bar{\gamma}$, the iterates $(\kappa_0^k, \kappa_1^k)_{k\in\mathbb{N}}$ generated by* (28) *converge to an element of $\mathscr{C}$.*

Proof. By Proposition C.6, the distance between successive iterates $(\kappa_0^k, \kappa_1^k)_{k\in\mathbb{N}}$, then the set of accumulation points of the sequence is connected (Lange, 2013, Proposition 12.4.1). Since we have a finite number of accumulation points by Proposition C.7, the sequence has a unique accumulation point. Besides, the sequence belongs to the compact set $[-1, 1]^2$, then it converges and its limit is one of the nine fixed points. □

Proposition C.10. *Consider the context of Proposition C.6, then there exists $\bar{\gamma} > 0$ such that for any stepsize $0 < \gamma < \bar{\gamma}$, the set of initializations such that the iterates $(\kappa_0^k, \kappa_1^k)_{k\in\mathbb{N}}$ generated by* (28) *converge to $(0, \pm 1), (\pm 1, 0)$ or $(0, 0)$ has Lebesgue measure zero (with respect to the Lebesgue measure on the manifold $\mathcal{M}$).*

Proof. The point $(0, 0)$ is a maxima of the risk $\mathcal{R}^<$ on $[-1, 1]^2$ and the value of the risk decreases along the iterates of (PGD) by Proposition C.6. We follow the ideas presented in the proof of Marion et al. (2024, Proposition 12), we can conclude that φ is differentiable on $[-1, 1]^2$, and that its Jacobian is not degenerate, besides φ is a local diffeomorphism around $(0, \pm 1)$ and $(\pm 1, 0)$, whose Jacobian matrix in each point has one eigenvalue in $[0, 1[$ and one eigenvalue in $]1, \infty[$. The result follows from the Center-Stable Manifold Theorem (Shub, 1987, Theorem III.7), we refer to Marion et al. (2024, Proposition 13) for a detailed and analogous proof. □

D Proofs of Section 4

D.1 Proof of Proposition 4.1

We provide hereafter the proof of Proposition 4.1, which proof follows.

Proposition D.1. *Consider $(X_\ell)_{1\leq \ell \leq L}$ i.i.d. drawn from (P_σ). Consider also $Z_\ell \in \{0, 1\}$ the latent variable of X_ℓ, i.e. $X_\ell | Z_\ell \sim \mathcal{N}(\mu^\star_{Z_\ell}, \sigma^2 I_d)$, and*

$$T^{\mathrm{lin}, \mu_0^\star, \mu_1^\star}(\mathbb{X})_1 = \frac{2}{L} \sum_{k=1}^{L} (e_k(\mu_0^\star) + e_k(\mu_1^\star)) X_k,$$

where $e_k(\mu) \overset{\mathrm{def}}{=} \lambda \langle X_1, \mu \rangle \langle X_k, \mu \rangle$. Then

$$\mathbb{E}[T^{\mathrm{lin}, \mu_0^\star, \mu_1^\star}(\mathbb{X})_1 | Z_1 = c] = \mu_c^\star \frac{\lambda}{L}[(L + 1) + 2(L + 3)\sigma^2], \quad c = \{0, 1\}.$$

$$\mathbb{E}[T^{\mathrm{lin}, \mu_0^\star, \mu_1^\star}(\mathbb{X})_1] = \frac{\mu_0^\star + \mu_1^\star}{2} \frac{\lambda}{L}[(L + 1) + 2(L + 3)\sigma^2].$$

Moreover, when $\lambda = \frac{L}{(L+1)+2(L+3)\sigma^2}$, then the encoding is unbiased, this is

$$\mathbb{E}[T^{\mathrm{lin}, \mu_0^\star, \mu_1^\star}(\mathbb{X})_1 | Z_1 = c] = \mu_c^\star, \quad c = \{0, 1\}.$$

Proof. We decompose the following term as follows,

$$\sum_{k=1}^{L}\langle X_1,\mu_0^\star\rangle\langle X_k,\mu_0^\star\rangle X_k = \langle X_1,\mu_0^\star\rangle^2 X_1 + \sum_{k=2}^{L}\langle X_1,\mu_0^\star\rangle\langle X_k,\mu_0^\star\rangle X_k$$

Due to the independence of the variables,

$$\begin{aligned}\mathbb{E}[T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}(\mathbb{X})_1] = \frac{2\lambda}{L}(&\mathbb{E}[(\langle X_1,\mu_0^\star\rangle^2 + \langle X_1,\mu_1^\star\rangle^2)X_1] \\ &+ (L-1)\mathbb{E}[(\langle X_1,\mu_0^\star\rangle\langle X_2,\mu_0^\star\rangle + \langle X_1,\mu_1^\star\rangle\langle X_2,\mu_1^\star)X_2]).\end{aligned}$$

On the one hand we have

$$\mathbb{E}(\langle X_1,\mu_0^\star\rangle^2 X_1|Z_1) = \langle\mu_{Z_1}^\star,\mu_0^\star\rangle^2\mu_{Z_1}^\star + \sigma^2(\mu_{Z_1}^\star + 2\langle\mu_{Z_1}^\star,\mu_0^\star\rangle\mu_0^\star).$$

On the other hand,

$$\mathbb{E}[\langle X_1,\mu_0^\star\rangle\langle X_2,\mu_0^\star\rangle X_2|Z_1,Z_2] = \langle\mu_{Z_1}^\star,\mu_0^\star\rangle[\langle\mu_{Z_2}^\star,\mu_0^\star\rangle\mu_{Z_2}^\star + \sigma^2\mu_0^\star].$$

Therefore,

$$\begin{aligned}\mathbb{E}[T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}(\mathbb{X})_1|Z_1,Z_2] = \frac{2\lambda}{L}[&(\langle\mu_{Z_1}^\star,\mu_0^\star\rangle^2 + \langle\mu_{Z_1}^\star,\mu_1^\star\rangle^2)\mu_{Z_1}^\star + 2\sigma^2\mu_{Z_1}^\star \\ &+ 2\sigma^2(\langle\mu_{Z_1}^\star,\mu_0^\star\rangle\mu_0^\star + \langle\mu_{Z_1}^\star,\mu_1^\star\rangle\mu_1^\star) \\ &+ (L-1)(\langle\mu_{Z_1}^\star,\mu_0^\star\rangle\langle\mu_{Z_2}^\star,\mu_0^\star\rangle + \langle\mu_{Z_1}^\star,\mu_1^\star\rangle\langle\mu_{Z_2}^\star,\mu_1^\star\rangle)\mu_{Z_2}^\star \\ &+ (L-1)\sigma^2(\langle\mu_{Z_1}^\star,\mu_0^\star\rangle\mu_0^\star + \langle\mu_{Z_1}^\star,\mu_1^\star\rangle\mu_1^\star)].\end{aligned}$$

And then,

$$\begin{aligned}&\mathbb{E}[T^{\mathrm{lin},\mu_0,\mu_1}(\mathbb{X})_1|Z_1] \\ &\quad= \frac{2\lambda}{L}[(\langle\mu_{Z_1}^\star,\mu_0^\star\rangle^2 + \langle\mu_{Z_1}^\star,\mu_1^\star\rangle^2)\mu_{Z_1}^\star + 2\sigma^2\mu_{Z_1}^\star + 2\sigma^2(\langle\mu_{Z_1}^\star,\mu_0^\star\rangle\mu_0^\star + \langle\mu_{Z_1}^\star,\mu_1^\star\rangle\mu_1^\star) \\ &\quad+ (L-1)\left(\frac{1}{2}+\sigma^2\right)(\langle\mu_{Z_1}^\star,\mu_0^\star\rangle\mu_0^\star + \langle\mu_{Z_1}^\star,\mu_1^\star\rangle\mu_1^\star)].\end{aligned}$$

Which let us conclude that for $c\in\{0,1\}$,

$$\begin{aligned}\mathbb{E}[T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}(\mathbb{X})_1|Z_1=c] &= \frac{\mu_c^\star\lambda}{L}((L+1)+2(L+3)\sigma^2), \\ \mathbb{E}[T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}(\mathbb{X})_1] &= \frac{(\mu_0^\star+\mu_1^\star)\lambda}{2L}((L+1)+2(L+3)\sigma^2).\end{aligned}$$

□

Proposition D.2. *Consider $(X_\ell)_{1\le\ell\le L}$ i.i.d. drawn from (P_σ). Consider also $Z_\ell\in\{0,1\}$ the latent variable of X_ℓ, i.e. $X_\ell|Z_\ell\sim\mathcal{N}(\mu_{Z_\ell}^\star,\sigma^2 I_d)$, and*

$$T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}(\mathbb{X})_1 = \frac{2}{L}\sum_{k=1}^{L}(e_k(\mu_0^\star)+e_k(\mu_1^\star))X_k,$$

where $e_k(\mu)\overset{\mathrm{def}}{=}\lambda\langle X_1,\mu\rangle\langle X_k,\mu\rangle$. Then for $c\in\{0,1\}$,

$$\begin{aligned}&\mathbb{E}[\|T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}(\mathbb{X})_1\|^2|Z_1=c] \\ &= \frac{4\lambda^2}{L^2}[1+\sigma^2(d+16)+8\sigma^4(d+7)+8\sigma^6(d+4)] \\ &+ 2\lambda^2\frac{(L-1)}{L^2}[3+\sigma^2(d+28)+4\sigma^4(d+16)+4\sigma^6(d+10)] \\ &+ \lambda^2\frac{(L-1)(L-2)}{L^2}[1+6\sigma^2+12\sigma^4+8\sigma^6].\end{aligned}$$

And

$$
\begin{aligned}
&\mathrm{Var}[T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}(\mathbb{X})_1|Z_1=c]\\
&=\frac{4\lambda^2}{L^2}[1+\sigma^2(d+16)+8\sigma^4(d+7)+8\sigma^6(d+4)]\\
&+2\lambda^2\frac{(L-1)}{L^2}[3+\sigma^2(d+28)+4\sigma^4(d+16)+4\sigma^6(d+10)]\\
&+\lambda^2\frac{(L-1)(L-2)}{L^2}[1+6\sigma^2+12\sigma^4+8\sigma^6]\\
&-\frac{\lambda^2}{L^2}[(L+1)+2(L+3)\sigma^2]^2.
\end{aligned}
$$

When $\lambda=\frac{L}{(L+1)+2(L+3)\sigma^2}$*, the encoding is unbiased, with variance*

$$
\begin{aligned}
&\mathrm{Var}[T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}(\mathbb{X})_1|Z_1=c]\\
&=\frac{4}{[(L+1)+2(L+3)\sigma^2]^2}[1+\sigma^2(d+16)+8\sigma^4(d+7)+8\sigma^6(d+4)]\\
&+\frac{2(L-1)}{[(L+1)+2(L+3)\sigma^2]^2}[3+\sigma^2(d+28)+4\sigma^4(d+16)+4\sigma^6(d+10)]\\
&+\frac{(L-1)(L-2)}{[(L+1)+2(L+3)\sigma^2]^2}[1+6\sigma^2+12\sigma^4+8\sigma^6]-1.
\end{aligned}
$$

Besides, when $L\to\infty$ *we get that,*

$$\mathrm{Var}[T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}(\mathbb{X})_1|Z_1=c]\sim 2\sigma^2,\quad \lambda\sim\frac{1}{1+2\sigma^2}.$$

In general, if λ *is not fixed and* $L\to\infty$*, we get*

$$\mathrm{Var}[T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}(\mathbb{X})_1|Z_1=c]\sim 2\lambda^2\sigma^2(1+2\sigma^2)^2.$$

Remark D.3. *We recall that*

$$\mathrm{Var}[X_1\mid Z_1=c]=\sigma^2 d.$$

Notably, by selecting λ *to ensure an unbiased encoding, the variance becomes independent of the dimension* d *and equals* $2\sigma^2$*. This shows a variance reduction effect whenever the dimension* d *is bigger than 2. More generally,* λ *can be chosen independently of* d *such that*

$$2\lambda^2(1+2\sigma^2)^2\ll d.$$

In this regime, the encoding also asymptotically reduces the variance of X_1*, conditioned on its cluster assignment, as the number of components* $L\to\infty$*.*

Proof. We note that the needed computations were already stated in the proof of Proposition 3.1, we follow as in the proof of Lemma C.1, without loss of generality, we assume $Z_1=0$, then we get that for $\mu_0,\mu_1\in\mathcal{M}$:

$$
\begin{aligned}
&\mathbb{E}[\|T^{\mathrm{lin},\mu_0,\mu_1}(\mathbb{X})_1\|^2|Z_1=0]\\
&=\frac{4\lambda^2}{L^2}[\kappa_0^4(1+\sigma^2(d+8))+8\sigma^2\kappa_0^2(1+\sigma^2(d+6))+8\sigma^4(1+\sigma^2(d+4))]\\
&+4\lambda^2\frac{L-1}{L^2}[\kappa_0^4(1+5\sigma^2)+4\sigma^2\kappa_0^2(1+6\sigma^2)+\sigma^2\kappa_0^2\kappa_1^2+16\sigma^6]\\
&+2\lambda^2\frac{(L-1)}{L^2}[\kappa_0^4(1+\sigma^2(d+4))+4\sigma^2\kappa_0^2(1+\sigma^2(d+3))+4\sigma^4(1+\sigma^2(d+2))]\\
&+\lambda^2\frac{(L-1)(L-2)}{L^2}[\kappa_0^4(1+4\sigma^2)+2\sigma^2\kappa_0^2(1+6\sigma^2)+8\sigma^6].
\end{aligned}
$$

Recalling that $\kappa_0=\langle\mu_0,\mu_0^\star\rangle$, $\kappa_1=\langle\mu_1,\mu_1^\star\rangle$, in order to compute $\mathbb{E}[\|T^{\mathrm{lin},\mu_0^*,\mu_1^*}(\mathbb{X})_1\|^2|Z_1=0]$, we just need to replace κ_0 and κ_1 by 1 in the previous expression, as follows

$$\mathbb{E}[\|T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}(\mathbb{X})_1\|^2|Z_1=0]$$

$$
\begin{aligned}
&= \frac{4\lambda^2}{L^2}[1+\sigma^2(d+8)+8\sigma^2(1+\sigma^2(d+6))+8\sigma^4(1+\sigma^2(d+4))] \\
&+ 4\lambda^2\frac{(L-1)}{L^2}[1+5\sigma^2+4\sigma^2(1+6\sigma^2)+\sigma^2+16\sigma^6] \\
&+ 2\lambda^2\frac{(L-1)}{L^2}[1+\sigma^2(d+4)+4\sigma^2(1+\sigma^2(d+3))+4\sigma^4(1+\sigma^2(d+2))] \\
&+ \lambda^2\frac{(L-1)(L-2)}{L^2}[1+4\sigma^2+2\sigma^2(1+6\sigma^2)+8\sigma^6] \\
&= \frac{4\lambda^2}{L^2}[1+\sigma^2(d+16)+8\sigma^4(d+7)+8\sigma^6(d+4)] \\
&+ 2\lambda^2\frac{(L-1)}{L^2}[3+\sigma^2(d+28)+4\sigma^4(d+16)+4\sigma^6(d+10)] \\
&+ \lambda^2\frac{(L-1)(L-2)}{L^2}[1+6\sigma^2+12\sigma^4+8\sigma^6].
\end{aligned}
$$

The expression of the variance comes from subtracting to this term the square of the conditional expectation given in Proposition 4.1. The asymptotic expressions are then straightforward to derive.

□

D.2 Proof of Proposition 4.2

Proof. Assume that the tokens are i.i.d., such that for any ℓ, $X_\ell \sim \frac{1}{2}\mathcal{N}(\mu_0^\star, \sigma^2 I) + \frac{1}{2}\mathcal{N}(\mu_1^\star, \sigma^2 I)$. The risk of the oracle predictor $T^{\text{lin},\mu_0^\star,\mu_1^\star}$ can be decomposed as follows

$$
\mathcal{L}(T^{\text{lin},\mu_0^\star,\mu_1^\star}) = (1+d\sigma^2) - (I_0) + (II_0) + (III_0) - (I) + (II) + (III), \tag{30}
$$

where, from the proof of Lemma C.1, by taking $\kappa_0 = \kappa_1 = 1$,

- $(I_0) = \frac{4\lambda}{L}\left[(1+\sigma^2(d+4)) + 2\sigma^2(1+\sigma^2(d+2))\right]$.
- $(II_0) = \frac{4\lambda^2}{L^2}[1+\sigma^2(d+16)+8\sigma^4(d+7)+8\sigma^6(d+4))]$.
- $(III_0) = 4\lambda^2\frac{L-1}{L^2}[1+10\sigma^2+24\sigma^4+16\sigma^6]$.
- $(I) = 2\lambda\frac{L-1}{L}[1+4\sigma^2+4\sigma^4]$.
- $(II) = 2\lambda^2\frac{(L-1)}{L^2}[1+\sigma^2(d+8)+4\sigma^4(d+7)+4\sigma^6(d+2)]$.
- $(III) = \lambda^2\frac{(L-1)(L-2)}{L^2}[1+6\sigma^2+12\sigma^4+8\sigma^6]$.

When L tends to ∞, only the first term together with (I) and (III) contribute. Therefore, we obtain that

$$
\mathcal{L}(T^{\text{lin},\mu_0^\star,\mu_1^\star}) \underset{L\to\infty}{\sim} (1+d\sigma^2) - \lambda[2+8\sigma^2+8\sigma^4] + \lambda^2[1+6\sigma^2+12\sigma^4+8\sigma^6].
$$

Choosing $\lambda = \frac{1+4\sigma^2+4\sigma^4}{1+6\sigma^2+12\sigma^4+8\sigma^6}$ (its value being independent of L) minimizes the equivalent bound obtained above. With such a choice, the equivalent becomes

$$
\begin{aligned}
\mathcal{L}(T^{\text{lin},\mu_0^\star,\mu_1^\star}) &\underset{L\to\infty}{\sim} (1+d\sigma^2) - \frac{(1+4\sigma^2+4\sigma^4)^2}{1+6\sigma^2+12\sigma^4+8\sigma^6} \\
&\underset{L\to\infty}{\sim} (1+d\sigma^2) - \frac{(1+2\sigma^2)^4}{(1+2\sigma^2)^3} \\
&\underset{L\to\infty}{\sim} (1+d\sigma^2) - 1 - 2\sigma^2 \\
&\underset{L\to\infty}{\sim} \sigma^2(d-2).
\end{aligned}
$$

□

Remark D.4. *In the case of the degenerate case ($\sigma = 0$), similar computations lead to*

$$
\mathcal{L}(T^{\text{lin},\mu_0^\star,\mu_1^\star}) = 1 - \frac{4\lambda}{L} + 4\frac{\lambda^2}{L^2} + 4\lambda^2\frac{L-1}{L^2} - 2\lambda\frac{L-1}{L} + 2\lambda^2\frac{L-1}{L^2} + \lambda^2\frac{(L-1)(L-2)}{L^2}
$$

$$= 1 - 2\lambda\frac{L+1}{L} + \lambda^2\frac{(L+3)}{L}$$

Optimizing this quantity w.r.t. λ *leads to choose* $\lambda = \frac{L+1}{L+3}$*, plugging this value for* λ *gives*

$$\mathcal{L}(T^{\mathrm{lin},\mu_0^\star,\mu_1^\star}) = 1 - \frac{(L+1)^2}{L(L+3)}.$$

In the degenerate case, we observe that as the sequence length tends to infinity, the risk of the attention-based predictor with oracle parameters converges to zero, matching that of the optimal quantizer.

E Technical results

This section gathers a series of technical results about Gaussian random variables, used to derive expression of the risk in the rest of the document.

Lemma E.1. *Consider* $G \sim \mathcal{N}(0, \sigma^2 I_d)$ *and* $\mu_0, \mu_1, \mu_2 \in \mathbb{R}^d$*, then*

1. $\mathbb{E}[\|G\|^2] = \sigma^2 d.$
2. $\mathbb{E}[\langle \mu_0, G\rangle] = 0.$
3. $\mathbb{E}(\langle \mu_0, G\rangle G) = \sigma^2 \mu_0.$
4. $\mathbb{E}[\langle \mu_0, G\rangle\langle \mu_1, G\rangle] = \sigma^2\langle \mu_0, \mu_1\rangle.$
5. $\mathbb{E}[\langle \mu_0, G\rangle^2\langle \mu_1, G\rangle^2] = \sigma^4(\|\mu_0\|^2\|\mu_1\|^2 + 2\langle \mu_0, \mu_1\rangle^2).$
6. $\mathbb{E}[\langle \mu_0, G\rangle\langle \mu_1, G\rangle^2\langle \mu_2, G\rangle] = \sigma^4(\|\mu_1\|^2\langle \mu_0, \mu_2\rangle + 2\langle \mu_0, \mu_1\rangle\langle \mu_1, \mu_2\rangle).$
7. $\mathbb{E}[\langle \mu_0, G\rangle\langle \mu_1, G\rangle\|G\|^2] = \sigma^4(d+2)\langle \mu_0, \mu_1\rangle.$
8. $\mathbb{E}[\langle \mu_0, G\rangle^2\langle \mu_1, G\rangle^2\|G\|^2] = \sigma^6(d+4)(\|\mu_0\|^2\|\mu_1\|^2 + 2\langle \mu_0, \mu_1\rangle^2).$

Lemma E.2. *Consider* $X \sim \mathcal{N}(\mu^\star, \sigma^2 I_d)$ *where* $\|\mu^\star\| = 1$ *and* $\mu_0 \in \mathbb{R}^d$*, then*

$$\mathbb{E}[\langle X, \mu_0\rangle^2\|X\|^2] = \langle \mu^\star, \mu_0\rangle^2(1 + \sigma^2(d+4)) + \sigma^2\|\mu_0\|^2(1 + \sigma^2(d+2)).$$

Lemma E.3. *Let* $X \sim \mathcal{N}(\mu^\star, \sigma^2 I_d)$*, where* $\|\mu^\star\| = 1$ *and* $\mu_0, \mu_1 \in \mathbb{R}^d$*, then*

$$p_0(\mu_0, \mu_1, \mu^\star) \overset{\mathrm{def}}{=} \mathbb{E}(\langle X, \mu_0\rangle^2\langle X, \mu_1\rangle^2\|X\|^2)$$

can be expressed as

$$\begin{aligned}
p_0(\mu_0, \mu_1, \mu^\star) = \langle \mu^\star, \mu_0\rangle^2\langle \mu^\star, \mu_1\rangle^2 \\
&+ \sigma^2\left(\langle \mu^\star, \mu_1\rangle^2\|\mu_0\|^2 + 4\langle \mu^\star, \mu_0\rangle\langle \mu^\star, \mu_1\rangle\langle \mu_0, \mu_1\rangle + \langle \mu^\star, \mu_0\rangle^2\|\mu_1\|^2\right) \\
&+ \sigma^2(d+8)\langle \mu^\star, \mu_0\rangle^2\langle \mu^\star, \mu_1\rangle^2 \\
&+ \sigma^4(\|\mu_0\|^2\|\mu_1\|^2 + 2\langle \mu_0, \mu_1\rangle^2 + (d+6)(\|\mu_0\|^2\langle \mu^\star, \mu_1\rangle^2 + \|\mu_1\|^2\langle \mu^\star, \mu_0\rangle^2)) \\
&+ 4\sigma^4(d+6)\langle \mu^\star, \mu_0\rangle\langle \mu^\star, \mu_1\rangle\langle \mu_0, \mu_1\rangle + \sigma^6(d+4)(\|\mu_0\|^2\|\mu_1\|^2 + 2\langle \mu_0, \mu_1\rangle^2).
\end{aligned}$$

Lemma E.4. *Let* Z_1 *and* $Z_2 \in \{0, 1\}$ *be fixed. Consider two independent* $\mathbb{R}^d-$*valued random variables* X_1 *and* X_2*, such that*

$$X_i | Z_i \sim \mathcal{N}(\mu^\star_{Z_i}, \sigma^2 I_d), \quad \text{for each } i = \{1, 2\},$$

where the unit vectors $\mu_0^\star, \mu_1^\star$ *(i.e.,* $\|\mu_0^\star\| = \|\mu_1^\star\| = 1$*) are orthogonal. For* $\mu_0, \mu_1, \mu_2 \in \mathbb{R}^d$*, define*

$$p_{1,0}(\mu_0, \mu_1, \mu^\star_{Z_1}, \mu_2) \overset{\mathrm{def}}{=} \mathbb{E}[\langle X_1, \mu_0\rangle^2\langle X_1, \mu_1\rangle\langle X_1, \mu_2\rangle | Z_1].$$

This quantity satisfies

$$\begin{aligned}p_{1,0}(\mu_0, \mu_1, \mu^\star_{Z_1}, \mu_2) = {} & \langle \mu^\star_{Z_1}, \mu_0\rangle^2 \langle \mu^\star_{Z_1}, \mu_1\rangle \langle \mu^\star_{Z_1}, \mu_2\rangle \\ & + \sigma^2 \left[\|\mu_0\|^2 \langle \mu^\star_{Z_1}, \mu_1\rangle\langle \mu^\star_{Z_1}, \mu_2\rangle + 2\langle \mu^\star_{Z_1}, \mu_0\rangle(\langle \mu^\star_{Z_1}, \mu_2\rangle\langle \mu_0, \mu_1\rangle + \langle \mu^\star_{Z_1}, \mu_1\rangle\langle \mu_0, \mu_2\rangle)\right] \\ & + \sigma^2 \left[\langle \mu^\star_{Z_1}, \mu_0\rangle^2 \langle \mu_1, \mu_2\rangle\right] \\ & + \sigma^4(\|\mu_0\|^2\langle \mu_1, \mu_2\rangle + 2\langle \mu_0, \mu_1\rangle\langle \mu_0, \mu_2\rangle).\end{aligned}$$

Moreover, we define

$$p_1(\mu_0, \mu_1, \mu^\star_{Z_1}, \mu^\star_{Z_2}) \overset{\text{def}}{=} \mathbb{E}[\langle X_1, \mu_0\rangle^2\langle X_1, \mu_1\rangle\langle X_2, \mu_1\rangle\langle X_1, X_2\rangle | Z_1, Z_2],$$

which satisfies

$$p_1(\mu_0, \mu_1, \mu^\star_{Z_1}, \mu^\star_{Z_2}) = \langle \mu^\star_{Z_2}, \mu_1\rangle p_{1,0}(\mu_0, \mu_1, \mu^\star_{Z_1}, \mu^\star_{Z_2}) + \sigma^2 p_{1,0}(\mu_0, \mu_1, \mu^\star_{Z_1}, \mu_1).$$

Lemma E.5. *Let $Z_1, Z_2 \in \{0, 1\}$ be fixed. Consider two independent $\mathbb{R}^d-$valued random variables X_1 and X_2, such that*

$$X_i | Z_i \sim \mathcal{N}(\mu^\star_{Z_i}, \sigma^2 I_d), \quad \textit{for each } i = \{1, 2\},$$

where the unit vectors $\mu^\star_0, \mu^\star_1$ (i.e., $\|\mu^\star_0\| = \|\mu^\star_1\| = 1$) are orthogonal. For $\mu_0, \mu_1 \in \mathbb{R}^d$, we get that

$$\begin{aligned}&\mathbb{E}[(\langle X_1, \mu_0\rangle\langle X_2, \mu_0\rangle + \langle X_1, \mu_1\rangle\langle X_2, \mu_1\rangle)\langle X_1, X_2\rangle | Z_1, Z_2] \\ &= \langle \mu^\star_{Z_1}, \mu^\star_{Z_2}\rangle(\langle \mu^\star_{Z_1}, \mu_0\rangle\langle \mu^\star_{Z_2}, \mu_0\rangle + \langle \mu^\star_{Z_1}, \mu_1\rangle\langle \mu^\star_{Z_2}, \mu_1\rangle) \\ &+ \sigma^2(\langle \mu_0, \mu^\star_{Z_1}\rangle^2 + \langle \mu_0, \mu^\star_{Z_2}\rangle^2 + \langle \mu_1, \mu^\star_{Z_1}\rangle^2 + \langle \mu_1, \mu^\star_{Z_2}\rangle^2) \\ &+ \sigma^4(\|\mu_0\|^2 + \|\mu_1\|^2)\end{aligned}$$

Lemma E.6. *Let $Z_1, Z_2 \in \{0, 1\}$ be fixed. Consider two independent $\mathbb{R}^d-$valued random variables X_1 and X_2, such that*

$$X_i | Z_i \sim \mathcal{N}(\mu^\star_{Z_i}, \sigma^2 I_d), \quad \textit{for each } i = \{1, 2\},$$

where the unit vectors $\mu^\star_0, \mu^\star_1$ (i.e., $\|\mu^\star_0\| = \|\mu^\star_1\| = 1$) are orthogonal. For $\mu_0, \mu_1 \in \mathbb{R}^d$, define also:

$$\begin{aligned}p_{2,0}(\mu_0, \mu_1, \mu^\star_{Z_1}) &\overset{\text{def}}{=} \mathbb{E}[\langle X_1, \mu_0\rangle\langle X_1, \mu_1\rangle | Z_1] \\ p_{2,1}(\mu_0, \mu_1, \mu^\star_{Z_2}) &\overset{\text{def}}{=} \mathbb{E}[\langle X_2, \mu_0\rangle\langle X_2, \mu_1\rangle\|X_2\|^2 | Z_2].\end{aligned}$$

These quantities satisfy

$$\begin{aligned}p_{2,0}(\mu_0, \mu_1, \mu^\star_{Z_1}) &= \langle \mu^\star_{Z_1}, \mu_0\rangle\langle \mu^\star_{Z_1}, \mu_1\rangle + \sigma^2\langle \mu_0, \mu_1\rangle, \\ p_{2,1}(\mu_0, \mu_1, \mu^\star_{Z_2}) &= \langle \mu^\star_{Z_2}, \mu_0\rangle\langle \mu^\star_{Z_2}, \mu_1\rangle + \sigma^2((d+4)\langle \mu^\star_{Z_2}, \mu_0\rangle\langle \mu^\star_{Z_2}, \mu_1\rangle + \langle \mu_0, \mu_1\rangle) \\ &\quad + \sigma^4(d+2)\langle \mu_0, \mu_1\rangle.\end{aligned}$$

Moreover, we define

$$p_2(\mu_0, \mu_1, \mu^\star_{Z_1}, \mu^\star_{Z_2}) \overset{\text{def}}{=} \mathbb{E}[\langle X_1, \mu_0\rangle\langle X_2, \mu_0\rangle\langle X_1, \mu_1\rangle\langle X_2, \mu_1\rangle\|X_2\|^2 | Z_1, Z_2],$$

which satisfies

$$p_2(\mu_0, \mu_1, \mu^\star_{Z_1}, \mu^\star_{Z_2}) = p_{2,0}(\mu_0, \mu_1, \mu^\star_{Z_1})p_{2,1}(\mu_0, \mu_1, \mu^\star_{Z_2}).$$

Lemma E.7. *Let $Z_1, Z_2, Z_3 \in \{0,1\}$ be fixed. Consider three independent $\mathbb{R}^d-$valued random variables X_1, X_2, X_3, where*

$$X_i|Z_i \sim \mathcal{N}(\mu^\star_{Z_i}, \sigma^2 I_d), \quad \textit{for each } i = \{1,2,3\},$$

such that $\mu^\star_0, \mu^\star_1$ unit vectors (i.e., $\|\mu^\star_0\| = \|\mu^\star_1\| = 1$) are orthogonal. For $\mu_0, \mu_1 \in \mathbb{R}^d$, define also:

$$\begin{aligned} p_{3,0}(\mu_0, \mu_1, \mu^\star_{Z_1}) &\overset{\text{def}}{=} \mathbb{E}[\langle X_1, \mu_0\rangle\langle X_1, \mu_1\rangle | Z_1], \\ p_{3,1}(\mu_0, \mu_1, \mu^\star_{Z_2}, \mu^\star_{Z_3}) &\overset{\text{def}}{=} \mathbb{E}[\langle X_2, \mu_0\rangle\langle X_3, \mu_1\rangle\langle X_2, X_3\rangle | Z_2, Z_3]. \end{aligned}$$

These quantities satisfy

$$\begin{aligned} p_{3,0}(\mu_0, \mu_1, \mu^\star_{Z_1}) &= \langle \mu^\star_{Z_1}, \mu_0\rangle\langle \mu^\star_{Z_1}, \mu_1\rangle + \sigma^2\langle \mu_0, \mu_1\rangle, \\ p_{3,1}(\mu_0, \mu_1, \mu^\star_{Z_2}, \mu^\star_{Z_3}) &= \langle \mu^\star_{Z_2}, \mu_0\rangle\langle \mu^\star_{Z_3}, \mu_1\rangle\langle \mu^\star_{Z_2}, \mu^\star_{Z_3}\rangle \\ &\quad + \sigma^2(\langle \mu^\star_{Z_2}, \mu_0\rangle\langle \mu^\star_{Z_2}, \mu_1\rangle + \langle \mu^\star_{Z_3}, \mu_0\rangle\langle \mu^\star_{Z_3}, \mu_1\rangle) \\ &\quad + \sigma^4\langle \mu_0, \mu_1\rangle. \end{aligned}$$

Moreover, we define

$$p_3(\mu_0, \mu_1, \mu^\star_{Z_1}, \mu^\star_{Z_2}, \mu^\star_{Z_3}) \overset{\text{def}}{=} \mathbb{E}[\langle X_1, \mu_0\rangle\langle X_2, \mu_0\rangle\langle X_1, \mu_1\rangle\langle X_3, \mu_1\rangle\langle X_2, X_3\rangle | Z_1, Z_2, Z_3],$$

which satisfies

$$p_3(\mu_0, \mu_1, \mu^\star_{Z_1}, \mu^\star_{Z_2}, \mu^\star_{Z_3}) = p_{3,0}(\mu_0, \mu_1, \mu^\star_{Z_1}) p_{3,1}(\mu_0, \mu_1, \mu^\star_{Z_2}, \mu^\star_{Z_3}).$$

F Experimental details

This section provides algorithmic details, choices of parameters, and settings used for the plots displayed in Sections A and 3.

F.1 Projected Stochastic Gradient Descent

We formally define the method Projected Stochastic Gradient Descent (PSGD), which we run for our numerical experiments.

PSGD iterates for linear attention heads. Given the objective function $\mathcal{R}^\rho : (\mathbb{S}^{d-1})^2 \to \mathbb{R}$ defined in ($\mathcal{P}_\rho$), we define $h : (\mathbb{S}^{d-1})^2 \times \mathbb{R}^{L\times d}$ as

$$h(\mu_0, \mu_1, \mathcal{X}) = \left\| X_1 - \frac{2}{L}\sum_{k=1}^{L} \lambda[X_1^\top(\mu_0\mu_0^\top + \mu_1\mu_1^\top)X_k]X_k \right\|_2^2 + \rho\langle \mu_0, X_1\rangle^2\langle \mu_1, X_1\rangle^2,$$

where X_i is the $i-$th row of the matrix $\mathcal{X}$. Consequently we can write

$$\mathcal{R}^\rho(\mu_0, \mu_1) = \mathbb{E}_{\mathbb{X}\sim\mathcal{D}}[h(\mu_0, \mu_1, \mathbb{X})],$$

where $\mathcal{D}$ is the distribution over $\mathbb{R}^{L\times d}$ where each row is i.i.d. according to

$$\frac{1}{2}\mathcal{N}(\mu^\star_0, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\mu^\star_1, \sigma^2 I_d).$$

Then, given and an initialization $(\mu_0^0, \mu_1^0) \in (\mathbb{S}^{d-1})^2$, a stepsize γ, we define $(\mu_0^k, \mu_1^k) \in (\mathbb{S}^{d-1})^2$ recursively by:

$$\begin{aligned} g_0^k &= \frac{1}{M}\sum_{i=1}^{M} \nabla_{\mu_0} h(\mu_0^k, \mu_1^k, \xi_i^k), \\ g_1^k &= \frac{1}{M}\sum_{i=1}^{M} \nabla_{\mu_1} h(\mu_0^k, \mu_1^k, \xi_i^k), \\ \mu_0^{k+1} &= \frac{\mu_0^k - \gamma(I_d - \mu_0^k(\mu_0^k)^\top)g_0^k}{\|\mu_0^k - \gamma(I_d - \mu_0^k(\mu_0^k)^\top)g_0^k\|_2}, \\ \mu_1^{k+1} &= \frac{\mu_1^k - \gamma(I_d - \mu_1^k(\mu_1^k)^\top)g_1^k}{\|\mu_1^k - \gamma(I_d - \mu_1^k(\mu_1^k)^\top)g_1^k\|_2}, \end{aligned} \tag{PSGD}$$

where M is called the batch size, and for each $k \in \mathbb{N}$, $(\xi_i^k)_{i=\{1,\ldots,M\}}$ are M independents samples of $\mathcal{D}$.

PSGD iterates for softmax attention heads. Given the objective function $\mathcal{R}^{\mathrm{soft},\rho_0} : (\mathbb{S}^{d-1})^2 \times \mathbb{R}^2 \to \mathbb{R}$ defined in $(\mathcal{P}_{\rho_0})$, for simplicity, we note that for an appropriate h_0, we can write

$$\mathcal{R}^{\mathrm{soft},\rho_0}(\mu_0, \mu_1, \psi, \lambda) = \mathbb{E}_{\mathbb{X}\sim\mathcal{D}}[h_0(\mu_0, \mu_1, \psi, \lambda, \mathbb{X})],$$

where $\mathcal{D}$ is the distribution over $\mathbb{R}^{L\times d}$ where each row is i.i.d. according to

$$\frac{1}{2}\mathcal{N}(\mu_0^\star, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\mu_1^\star, \sigma^2 I_d).$$

Then, given and an initialization $(\mu_0^0, \mu_1^0) \in (\mathbb{S}^{d-1})^2$, $(\psi^0, \lambda^0) = (2, 3)$, a stepsize γ, we define $(\mu_0^k, \mu_1^k) \in (\mathbb{S}^{d-1})^2$ and $(\psi^k, \lambda^k) \in \mathbb{R}^2$ recursively by:

$$\begin{aligned}
g_0^k &= \frac{1}{M}\sum_{i=1}^{M} \nabla_{\mu_0} h_0(\mu_0^k, \mu_1^k, \psi^k, \lambda^k, \xi_i^k),\\
g_1^k &= \frac{1}{M}\sum_{i=1}^{M} \nabla_{\mu_1} h_0(\mu_0^k, \mu_1^k, \psi^k, \lambda^k, \xi_i^k),\\
\mu_0^{k+1} &= \frac{\mu_0^k - \gamma(I_d - \mu_0^k(\mu_0^k)^\top)g_0^k}{\|\mu_0^k - \gamma(I_d - \mu_0^k(\mu_0^k)^\top)g_0^k\|_2},\\
\mu_1^{k+1} &= \frac{\mu_1^k - \gamma(I_d - \mu_1^k(\mu_1^k)^\top)g_1^k}{\|\mu_1^k - \gamma(I_d - \mu_1^k(\mu_1^k)^\top)g_1^k\|_2},\\
\psi^{k+1} &= \psi^k - \gamma\frac{1}{M}\sum_{i=1}^{M} \partial_\psi h_0(\mu_0^k, \mu_1^k, \psi^k, \lambda^k, \xi_i^k),\\
\lambda^{k+1} &= \lambda^k - \gamma\frac{1}{M}\sum_{i=1}^{M} \partial_\lambda h_0(\mu_0^k, \mu_1^k, \psi^k, \lambda^k, \xi_i^k),
\end{aligned} \tag{PSGD$_{\mathrm{soft}}$}$$

where M is called the batch size, and for each $k \in \mathbb{N}$, $(\xi_i^k)_{i=\{1,\dots,M\}}$ are M independents samples of $\mathcal{D}$.

Remark F.1. *Gradient computations in the numerical experiments were carried out using JAX (Bradbury et al., 2018).*

F.2 Experimental details

In the following, we provide the experimental setup corresponding to Sections A and 3.

We use input sequences of length $L = 30$ of 5-dimensional tokens ($d = 5$), and define the true centroids as $\mu_0^\star = (0, 0, 0, 0, 1)$ and $\mu_1^\star = (-1, 0, 0, 0, 0)$. We recall that the metric used to quantify the distance to the centroids (up to a sign) is defined in (11).

Experimental details of Section A. Regarding the experiment on the manifold, i.e., Figure 4a, we perform 10^4 (PSGD) iterations without regularization ($\rho = 0$) with a learning rate of $\gamma = 0.01$, $\lambda = 0.6$, batch size $M = 256$. The experiment is repeated across 10 independent runs, each initialized randomly on the manifold $\widetilde{\mathcal{M}}$.

For the rest of the experiments of this section, we adopt the same setup as before, with the exception that each run is randomly initialized on the unit sphere. In Figure 4b, we perform 10^4 iterations to observe that without adding a regularization term, we only get partial alignment of the Transformer parameters towards the true centroids.

Then, in Figure 5a we perform 5×10^3 iterations of (PSGD) to minimize the regularized risk $\mathcal{R}^\rho$ for 15 values of the regularization strength ρ, linearly spaced in $[0, 0.3]$. Finally, in Figure 5b we choose $\rho = 0.1$ and perform 10^4 (PSGD) iterations.

Experimental details of Section 3. Regarding the experiment on the manifold, i.e., Figure 1, we run the algorithm for 10^4 iterations without regularization ($\rho = 0$), with a learning rate of $\gamma = 0.01$, batch size $M = 256$, and choosing $\lambda = 0.6$ for $\sigma = 0.3$, and $\lambda = 0.2$ for $\sigma = 1$. The experiment is repeated across 10 independent runs, each initialized randomly on the manifold $\mathcal{M}$.

For the rest of the experiments of this section, we adopt the same setup as before, with the exception that each run is randomly initialized on the unit sphere. In Figure 6a we perform 5×10^3 iterations of (PSGD) to minimize the regularized risk $\mathcal{R}^\rho$ for 30 values of the regularization strength ρ, linearly spaced in $[0, 3]$. Finally, in Figure 6b we choose $\rho = 0.2$ and perform 10^4 (PSGD) iterations.

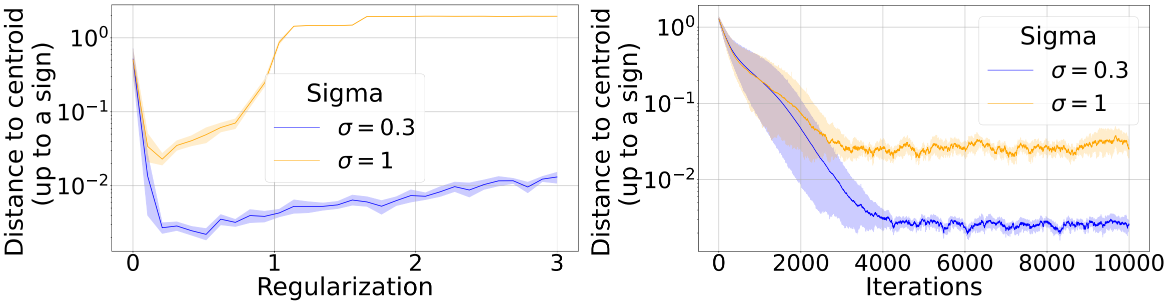
Remark F.2. *All experiments in Section A and 3 can be run on a standard laptop. Most complete within a few minutes, with the exception of those in Figures 5a and 6a, which require approximately 20 minutes and up to an hour, respectively, due to repeated problem-solving across a grid of regularization strengths.*

G Additional numerical experiments

G.1 Regularizing the risk to handle arbitrary initialization

Consider the experimental setting described in Section 3, i.e., the GMM model has 2 components with orthogonal centroids, and σ can take the values 0.3 or 1.

When initialization is performed outside the manifold, a small regularization term (of the order of 10^{-1}) substantially improves the accuracy of the recovered centroids, reducing the error below 10^{-2}, as shown in Figure 6a. However, as the strength of the regularization increases, it gradually overrides the original objective and impairs the alignment of the head parameters with the true centroids —an effect that becomes more pronounced at higher noise levels. In Figure 6b, we fix the regularization strength and observe linear convergence towards the centroids over the course of 4×10^3 iterations. The error eventually plateaus near 10^{-3} for $\sigma = 0.3$ and near 10^{-1} for $\sigma = 1$. This shows that the regularization strategy inspired by the analysis of the simplified Dirac mixture case remains effective in the more realistic setting of Gaussian mixtures. In this context as well, it enhances the interpretability of the attention parameters by encouraging their alignment with the unknown components of the underlying mixture.



(a) Distance to centroids after 5000 (PSGD) iterations vs regularization strength ρ for the minimization of $\mathcal{R}^\rho$, with an initialization on the unit sphere.

(b) Distance to centroids vs (PSGD) iterations for the minimization of $\mathcal{R}^\rho$, with an initialization on the unit sphere and regularization $\rho = 0.2$.

Figure 6: Performance of (PSGD), with data drawn from (P_σ). 10 runs, 95% percentile intervals are plotted.

In what follows, we first present numerical experiments in dimension 100. We then vary the dimension from 4 to 200. Results are shown only for the linear approach, as the softmax variant exhibits numerical instability in higher dimensions.

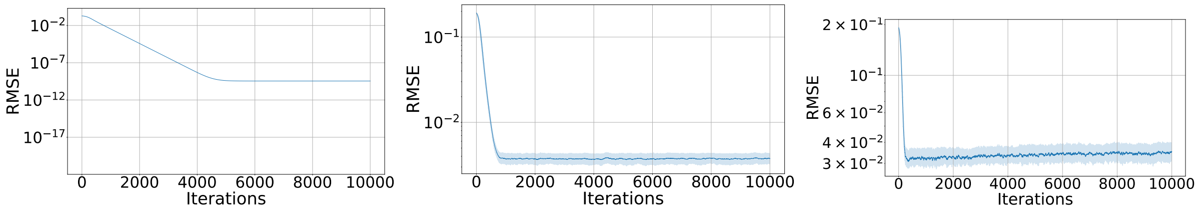
G.2 Influence of the dimension

Experiments in $\mathbb{R}^{100}$. We use input sequences of length $L = 30$ in $\mathbb{R}^{100}$, where we define two centroids: $\mu_0^\star = (\underbrace{0, \dots, 0}_{99 \text{ times}}, 1)$ and $\mu_1^\star = (-1, \underbrace{0, \dots, 0}_{99 \text{ times}})$. The model is trained using (PSGD) with an online batch sampling strategy, with a batch size of 256, and a learning rate of 0.01. Due to the big dimensionality of the problem, we modify the concept introduced as distance to the centroid up to a sign by the concept of minimal root mean squared error, which is nothing but the distance to the centroid (up to a sign) divided by the square root of the dimension, i.e.,

$$\text{Minimal RMSE} = \frac{\sqrt{\min\{\text{dist}_1, \text{dist}_2\}}}{\sqrt{d}},$$

where

$$\text{dist}_1 = \min\{\|\hat{\mu}_0 - \mu_0^\star\|^2, \|\hat{\mu}_0 + \mu_0^\star\|^2\} + \min\{\|\hat{\mu}_1 - \mu_1^\star\|^2, \|\hat{\mu}_1 + \mu_1^\star\|^2\},$$
$$\text{dist}_2 = \min\{\|\hat{\mu}_0 - \mu_1^\star\|^2, \|\hat{\mu}_0 + \mu_1^\star\|^2\} + \min\{\|\hat{\mu}_1 - \mu_0^\star\|^2, \|\hat{\mu}_1 + \mu_0^\star\|^2\},$$

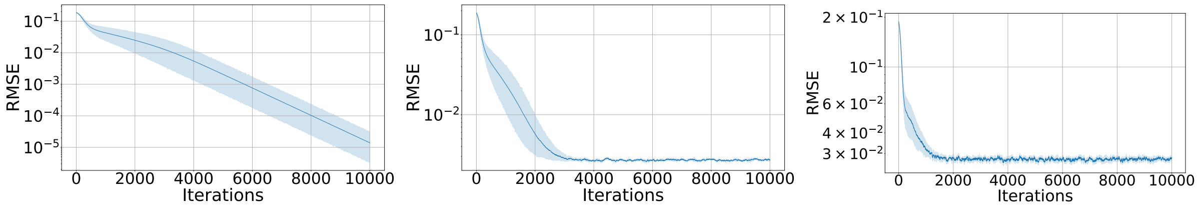


(a) $\sigma = 0$.

(b) $\sigma = 0.3$.

(c) $\sigma = 1$.

Figure 7: Minimal RMSE vs Iterations, Initialization on the manifold in dimension 100.



(a) $\sigma = 0$.

(b) $\sigma = 0.3$.

(c) $\sigma = 1$

Figure 8: Minimal RMSE vs Iterations in dimension 100, Regularization $\rho = 0.1$ for $\sigma = 0$, $\rho = 0.2$ for $\sigma > 0$, Initialization on the unit sphere.

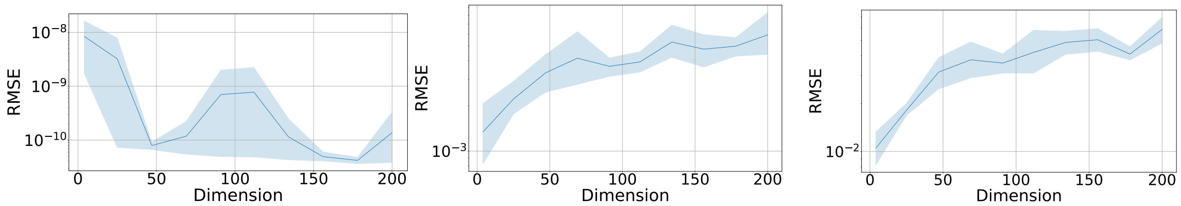
and $\mu_0^\star, \mu_1^\star$ are the true centroids, and $\hat{\mu}_0, \hat{\mu}_1$ are the returned parameters from (PSGD). In Figures 7, 8 we can observe the behavior of the RMSE over the iterations for different levels of noise σ. We remark that in Figure 7 we initialize on the manifold $\mathcal{M}$, and there is no regularization term (i.e. $\rho = 0$), in Figure 8 we initialize randomly over the unit sphere and we set $\rho = 0.2$. In both experiments we set $\lambda = 0.6$ for the case $\sigma = 0$ and $\sigma = 0.3$, and $\lambda = 0.2$ for the case $\sigma = 1$.

In each experiment, the RMSE is of the order 10^{-2}, which can be interpreted as, on average per coordinate, the estimators $\hat{\mu}_0, \hat{\mu}_1$ are missing the true parameters $\mu_0^\star, \mu_1^\star$ by 10^{-2}, suggesting a high level of accuracy in the estimation procedure.

Making d vary. We repeat the same experiment as before, just varying the dimension of the problem, the two centroids in $\mathbb{R}^d$ are defined by $\mu_0^\star = (\underbrace{0, \dots, 0}_{\text{d-1 times}}, 1)$ and $\mu_1^\star = (-1, \underbrace{0, \dots, 0}_{\text{d-1 times}})$. For d ranging between 4 and 200, we show in Figures 9 and 10, on the x-axis the dimension of the problem and on the y-axis the minimal RMSE after 5000 iterations.

Regardless of the initialization regime, in the noiseless case ($\sigma = 0$) the minimal RMSE decreases as the problem dimension d grows. By contrast, for any strictly positive noise level, the minimal RMSE increases slowly with d— for $\sigma = 0.3$ it remains of order 10^{-3}, and for $\sigma = 1$ of order 10^{-2}. This reflects the growing difficulty of the problem as both dimensionality and noise increase. We recall that the minimal RMSE can be interpreted as the average discrepancy per coordinate between the estimated parameters and the true centroids, suggesting a high level of accuracy in each experiment.

Regarding running times, the experiments run in Figures 7 and 8 take approximately two hours on a standard laptop, while that of Figures 9 and 10 may require up to 12 hours, to cover the dimension grid.

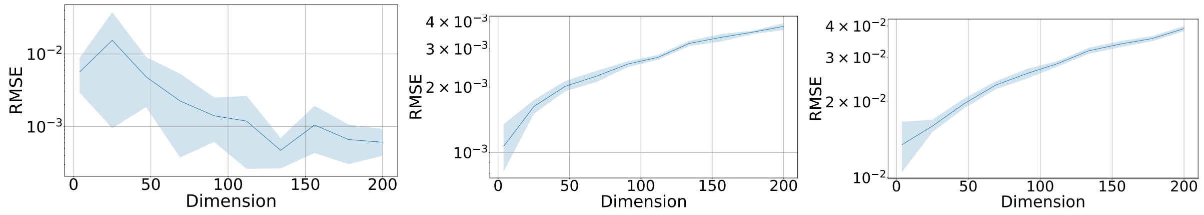


(a) $\sigma = 0$.

(b) $\sigma = 0.3$.

(c) $\sigma = 1$.

Figure 9: Minimal RMSE vs Dimensionality, Initialization on the manifold $\mathcal{M}$.problem on the y-axis: we do not know the scale



(a) $\sigma = 0$. (b) $\sigma = 0.3$. (c) $\sigma = 1$.

Figure 10: Minimal RMSE vs Dimensionality, Regularization $\rho = 0.1$ for $\sigma = 0$, $\rho = 0.2$ for $\sigma > 0$, Initialization on the unit sphere.

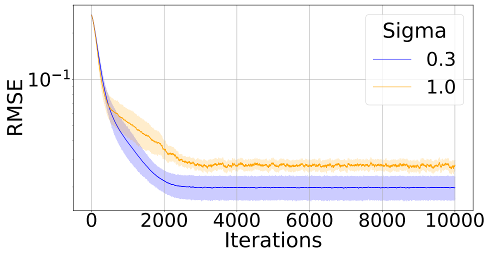


Figure 11: Minimal RMSE vs Iterations in dimension 500, Random initialization on the unit sphere of the centroids and of initial guesses, Regularization $\rho = 0.2$, 10 runs, 95% percentile intervals are plotted.

G.3 Relaxing the orthogonality assumption

We replicate the experiments and parameter-selection procedure from Section G.2, but this time initializing the centroids and the initial points uniformly at random on the sphere $\mathbb{S}^{d-1}$ in each run. Figure 11 illustrates the algorithm's convergence behavior over 10000 iterations in the case $d = 50$. In contrast, Figure 12 shows the minimal RMSE of (PSGD) after 5000 iterations for dimensions d ranging from 4 to 100.

We observed the expected behavior: as the dimension increases, randomly initializing centroids on the sphere makes them more likely to be orthogonal, and thus yields better results at higher dimensions. This effect is stronger at lower noise levels and becomes noticeably clearer beyond 40 dimensions.

Regarding running times, the experiments run in Figure 11 take approximately one hour on a standard laptop, while that of Figures 12 may require up to 7 hours, to cover the dimension grid.

G.4 Extension to Gaussian mixture model with three components

We propose an extension of our work to the case where the mixture counts three orthonormal centroids. We believe that the approach described below would further generalize to the case of K orthonormal centroids with $K < d$. Specifically, we assume that the tokens are i.i.d. drawn from the mixture model

$$X_\ell \sim \frac{1}{3}\mathcal{N}(\mu_0^\star, \sigma^2 I_d) + \frac{1}{3}\mathcal{N}(\mu_1^\star, \sigma^2 I_d) + \frac{1}{3}\mathcal{N}(\mu_2^\star, \sigma^2 I_d), \tag{P_σ}$$

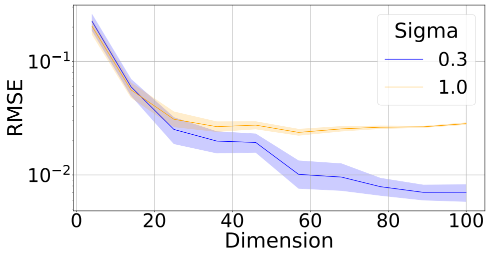
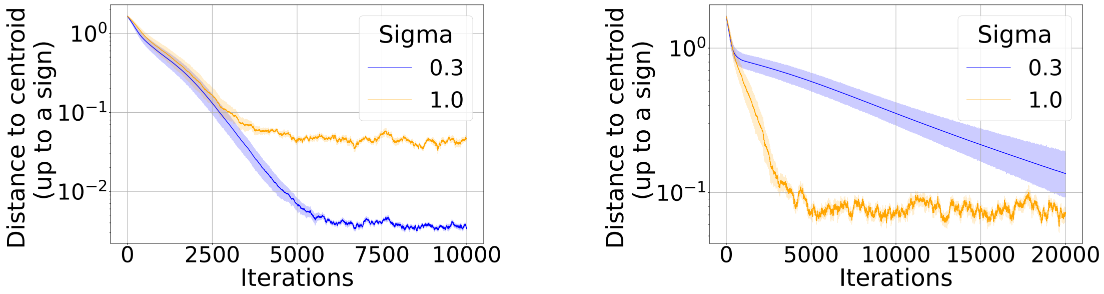


Figure 12: Minimal RMSE vs Dimensionality, Random initialization on the unit sphere of the centroids and of initial guesses, Regularization $\rho = 0.2$, 10 runs, 95% percentile intervals are plotted.



(a) With regularization $r^{(1)}$.

(b) With regularization $r^{(2)}$.

Figure 13: Distance to centroids vs number of iterations, with regularization strength $\rho = 0.2$, initialization on the unit sphere. 10 runs, 95% percentile intervals are plotted.

where $\mu_0^\star, \mu_1^\star, \mu_2^\star$ are orthonormal vectors. It is natural to consider an attention-based predictor composed of three attention heads, parameterized by $\mu_0, \mu_1, \mu_2 \in \mathbb{R}^d$,

$$T^{\text{lin},\mu_0,\mu_1,\mu_2}(\mathbb{X}) = H^{\text{lin},\mu_0}(\mathbb{X}) + H^{\text{lin},\mu_1}(\mathbb{X}) + H^{\text{lin},\mu_2}(\mathbb{X}). \tag{31}$$

The associated risk is $\mathcal{R}(\mu_0, \mu_1, \mu_2) = \mathbb{E}[\|X_1 - T^{\text{lin},\mu_0,\mu_1,\mu_2}(\mathbb{X})_1\|_2^2]$. There are two natural generalizations of the regularization term to this case:

$$r^{(1)}(\mu_0, \mu_1, \mu_2) = \sum_{0 \le i < j \le 2} \langle \mu_i, X_1 \rangle^2 \langle \mu_j, X_1 \rangle^2,$$
$$r^{(2)}(\mu_0, \mu_1, \mu_2) = \prod_{i=1}^{3} \langle \mu_i, X_1 \rangle^2.$$

The first one promotes pairwise orthogonality while the second one promotes mutual orthogonality.

We use input sequences of length $L = 30$ in $\mathbb{R}^6$, where we define the three centroids $\mu_0^\star = (1,0,0,0,0,0), \mu_1^\star = (0,0,0,1,0,0), \mu_2^\star = (0,0,0,0,0,1)$. The model is trained with an online batch sampling strategy (similar to PSGD, changing the data distribution and the regularization term), with a batch size of 256, and a learning rate of 0.01. We take $\lambda = 0.6$ for $\sigma = 0.3$, and $\lambda = 0.2$ for $\sigma = 1$. Since any parameter could learn any centroid up to a sign, we introduce the following distance to the centroid (up to a sign):

$$\min_{\pi \in S_3} \min_{s \in \{-1,1\}^3} \sqrt{\sum_{i=1}^{3} \|\hat{\mu}_{\pi(i)} - s_i \mu_i^\star\|^2},$$

where S_3 is the symmetric group of order 3 and $\hat{\mu}_0, \hat{\mu}_1, \hat{\mu}_2$ are the parameters returned by the algorithm. We present the results in Figure 13. We observe that the regularization $r^{(1)}$ outperforms $r^{(2)}$, since it explicitly includes all pairwise terms to enforce orthogonality. However, we note that the number of regularization terms grows quadratically with the number of centroids.

Regarding running times, the experiments run in Figure 13 take approximately 15 minutes on a standard laptop.

H Softmax attention layers and clustering

In this section, we assess the abilities of attention-based predictors involving a softmax activation in a clustering context.

H.1 Problem setting

An attention-based learner with softmax activation. We recall that an attention head made of a self-attention layer can be written as follows:

$$H^{\text{soft}_\lambda}(\mathbb{X}) = \text{softmax}_\lambda\left(\mathbb{X}QK^\top\mathbb{X}^\top\right)\mathbb{X}V$$

where the softmax of temperature $\lambda > 0$ is applied row-wise, and the matrices $K, Q, V \in \mathbb{R}^{d\times p}$ are usually referred to as keys, queries and values. As in Section 2, we assume that the values are taken as identity, meaning that the attention head simply outputs combinations of the initial tokens weighted by attention scores. Furthermore, we assume that the key and query matrices are equal to the same row matrix $\mu^\top \in \mathbb{R}^{1\times d}$, we obtain

$$H^{\mathrm{soft}_\lambda,\mu}(\mathbb{X}) = \mathrm{softmax}_\lambda\left(\mathbb{X}\mu\mu^\top\mathbb{X}^\top\right)\mathbb{X}. \tag{32}$$

With such an architecture, the ℓ-th output vector is therefore given by

$$H^{\mathrm{soft}_\lambda,\mu}(\mathbb{X})_\ell = \sum_{k=1}^{L} \mathrm{softmax}_\lambda\left(X_\ell^\top\mu\mu^\top\mathbb{X}^\top\right)_k X_k, \tag{33}$$

which corresponds to aggregating the X_k's when X_k and X_ℓ are simultaneously aligned with μ. This head should be a good candidate to estimate a centroid of a mixture model. In the case where the mixture involves two components, one could train two attention heads:

$$(\hat{\mu}_0, \hat{\mu}_1) \in \mathrm{argmin}_{\mu_0,\mu_1\in\mathbb{S}^{d-1}}\mathcal{R}^{\mathrm{soft}}(\mu_0,\mu_1), \tag{34}$$

where

$$\begin{aligned}\mathcal{R}^{\mathrm{soft}}(\mu_0,\mu_1) &= \frac{1}{L}\mathbb{E}\left[\left\|\mathbb{X} - (H^{\mathrm{soft}_\lambda,\mu_0} + H^{\mathrm{soft}_\lambda,\mu_1})(\mathbb{X})\right\|_F^2\right] \\ &= \frac{1}{L}\mathbb{E}\left[\sum_{\ell=1}^{L}\left\|X_\ell - (H^{\mathrm{soft}_\lambda,\mu_0} + H^{\mathrm{soft}_\lambda,\mu_1})(\mathbb{X})_\ell\right\|_2^2\right].\end{aligned} \tag{35}$$

Remark H.1 (The attention heads are biased)**.** *As the tokens $(X_\ell)_\ell$ are independent, we have that*

$$\mathcal{R}^{\mathrm{soft}}(\mu_0,\mu_1) = \mathbb{E}[\|X_1 - (H^{\mathrm{soft}_\lambda,\mu_0} + H^{\mathrm{soft}_\lambda,\mu_1})(\mathbb{X})_1\|_2^2].$$

We note that $H^{\mathrm{soft},\mu}(\mathbb{X})_1 = \mathrm{softmax}_\lambda(\langle X_1,\mu\rangle v)\mathbb{X}$, where the vector v is L-dimensional with components $v_\ell = \langle X_\ell, \mu\rangle$. In the idealized case where $\sigma^2 = 0$, then each token X_ℓ is sampled according to a mixture of Dirac masses given by $\frac{1}{2}\delta_{\mu_0^\star} + \frac{1}{2}\delta_{\mu_1^\star}$. Therefore, if we evaluate $H^{\mathrm{soft}_\lambda,\mu}(\mathbb{X})_1$ on $\mu = \mu_0^\star$, we observe the following:

- *Conditionally to $X_1 = \mu_1^\star$, then*

$$H^{\mathrm{soft}_\lambda,\mu_0^\star}(\mathbb{X})_1 = \left(\frac{1}{L},\dots,\frac{1}{L}\right)\mathbb{X}.$$

  This implies that even in a completely misaligned set-up (i.e., $\mu = \mu_0^\star, X_1 = \mu_1^\star$), the proposed attention head will return, as the transformation of the first token X_1, the empirical mean of the sequence tokens. This highlight the bias introduced by such an attention head, which should be handled through the use of centering techniques.

- *Conditionally to $X_1 = \mu_0^\star$, then*

$$H^{\mathrm{soft}_\lambda,\mu_0^\star}(\mathbb{X})_1 = \mathrm{softmax}_\lambda(v)\mathbb{X} \approx \frac{\exp(\lambda)\mu_0^\star + \mu_1^\star}{\exp(\lambda)+1}.$$

  This suggests that in the perfectly aligned case (i.e., $\mu = \mu_0^\star, X_1 = \mu_0^\star$), selecting a sufficiently large softmax temperature λ will cause the model to assign negligible weight to the misaligned components –a desirable property.

Debiasing and disentangling heads. To handle the bias introduced by the attention heads, discussed in Remark H.1, we propose to consider centered heads instead, leading to the following modified version of the risk

$$\mathcal{R}^{\mathrm{soft}}(\mu_0,\mu_1,\lambda,\psi) = \frac{1}{L}\mathbb{E}\left[\sum_{\ell=1}^{L}\left\|X_\ell - (H^{\mathrm{soft}_\lambda,\mu_0} + H^{\mathrm{soft}_\lambda,\mu_1})(\mathbb{X})_\ell + \frac{\psi}{L}\sum_{k=1}^{L}X_k\right\|_2^2\right]. \tag{36}$$

Considering such a risk is equivalent to using heads where a term proportional to $\frac{1}{L}\sum_{k=1}^{L} X_k$ is substracted. This type of head is known as shaped attention (Noci et al., 2023; He and Hofmann, 2024). For instance, initializing $\psi = 2$ debiases both attention heads independently, without considering their interaction. Using heads with oracle parameters, one would expect that a single head provides all the necessary information, making it sufficient to debias only that head (i.e. $\psi = 1$). In that case, one should obtain:

$$\mathcal{R}^{\text{soft}}(\mu_0^\star, \mu_1^\star, \lambda^\star, 1) \approx \min \mathcal{R}^{\text{soft}}. \tag{37}$$

However, when using non-oracle parameters μ_0 and μ_1 within the debiased heads, the risk function may admit global minima where the heads align with zero, one, or both centroids, which is undesirable for the clustering purpose. Therefore, we must enforce a constraint ensuring that each head aligns with exactly one centroid. To achieve this, we introduce the regularization term:

$$r_0(\mu_0, \mu_1) \overset{\text{def}}{=} \mathbb{E}[(\langle \mu_0, X_1\rangle - 1)^2(\langle \mu_1, X_1\rangle - 1)^2] + \langle \mu_0, \mu_1\rangle,$$

leading to the following regularized optimization problem

$$\min_{\mu_0,\mu_1\in\mathbb{S}^{d-1}} \mathcal{R}^{\text{soft},\rho_0}(\mu_0, \mu_1, \lambda, \psi) \overset{\text{def}}{=} \mathcal{R}^{\text{soft}}(\mu_0, \mu_1, \lambda, \psi) + \rho_0 r_0(\mu_0, \mu_1), \tag{$\mathcal{P}_{\rho_0}$}$$

where $\rho_0 > 0$.

H.2 Numerical experiments

We run Projected Stochastic Gradient Descent (see Appendix F.1) to learn the centroids $\mu_0^\star$ and $\mu_1^\star$ as well as the weights ψ and λ.

In this experiment, we use input sequences of length $L = 30$ of 5-dimensional tokens ($d = 5$), drawn from a 2-component Gaussian mixture of centroids $\mu_0^\star = (0, 0, 0, 0, 1)$ and $\mu_1^\star = (-1, 0, 0, 0, 0)$. The variance of each component is either set to $\sigma = 0.3$ (low interference) or to $\sigma = 1$ (high interference).

The model based on two softmax attention heads parameterized by μ_0 and μ_1 is trained using ($\text{PSGD}_{\text{soft}}$) with an online batch sampling strategy, with a batch size of 256, a learning rate of $\gamma = 0.01$, and running for a total of 3000 iterations. Additionally, we initialize with λ set to 3 and a centering value ψ of 2. Here we use the metric *distance to the centroids*, given by

$$\sqrt{\min\{\text{dist}_1, \text{dist}_2\}}, \tag{38}$$

where

$$\text{dist}_1 = \|\hat{\mu}_0 - \mu_0^\star\|^2 + \|\hat{\mu}_1 - \mu_1^\star\|^2,$$
$$\text{dist}_2 = \|\hat{\mu}_0 - \mu_1^\star\|^2 + \|\hat{\mu}_1 - \mu_0^\star\|^2,$$

and $\mu_0^\star, \mu_1^\star$ denote the true centroids, respectively, while $\hat{\mu}_0, \hat{\mu}_1$ are the parameters returned by ($\text{PSGD}_{\text{soft}}$). We remark that this distance is finer than the one defined in 11, as it does not disregard sign flips. The results are visualized in Figure 14a, we observe that a regularization term substantially improves the accuracy of the recovered solutions. However, as the strength of the regularization increases, it gradually overrides the original objective and impairs the alignment of the head parameters with the true centroids —an effect that becomes more pronounced at higher noise level, an effect also noticed in Section 3.

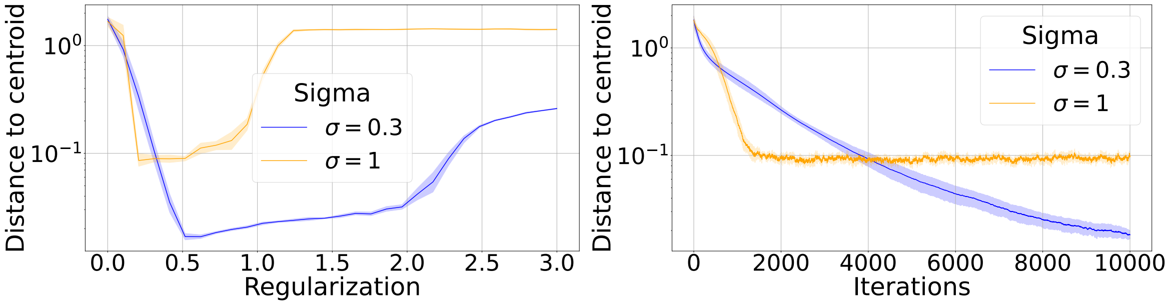
In Figure 14b, we set the regularization parameter ρ_0 to 0.5, and run $\text{PSGD}_{\text{soft}}$ for 10^4 iterations. We observe that the model yields accurate solutions under low interference ($\sigma = 0.3$); however, as the interference increases ($\sigma = 1$), the ability of the softmax attention heads to align with the underlying centroids is progressively impaired. A similar loss in alignment accuracy is observed as the dimensionality increases.

The experiments in Figure 14b run in a few minutes on a standard laptop, whereas those in Figure 14a may take up to two hours to cover the grid in the regularization hyperparameter.

I Proofs of Section 5

Proof of Proposition 5.1. We have

$$\mathbb{E}\left[\left\|X_1 - \frac{2\lambda}{L}\sum_{\ell=1}^{L}\langle X_1, X_\ell\rangle X_\ell\right\|^2 \middle| \mu_1^\star, \mu_0^\star\right]$$



(a) Distance to the centroids after 5000 PSGD iterations vs regularization strength ρ for the minimization of $\mathcal{R}^{\mathrm{soft},\rho_0}$, with an initialization on the unit sphere, with data drawn from the non-degenerate case (P_σ). 10 runs, 95% percentile intervals are plotted

(b) Distance to centroids vs PSGD iterations for the minimization of $\mathcal{R}^{\mathrm{soft},\rho_0}$, with an initialization on the unit sphere and regularization $\rho_0 = 0.5$, with data drawn from the non-degenerate case (P_σ). 10 runs, 95% percentile intervals are plotted.

Figure 14: Performance of ($\mathrm{PSGD}_{\mathrm{soft}}$), with data drawn from the the non-degenerate case (P_σ). 10 runs, 95% percentile intervals are plotted.

$$
\begin{aligned}
&= \mathbb{E}\left[\|X_1\|^2 \middle| \mu_1^\star, \mu_0^\star\right] - \frac{4\lambda}{L}\mathbb{E}[\|X_1\|^4|\mu_1^\star, \mu_0^\star] - \frac{4\lambda}{L}\sum_{\ell=2}^{L}\mathbb{E}\left[\langle X_1, X_\ell\rangle^2|\mu_1^\star, \mu_0^\star\right] \\
&+ \frac{4\lambda^2}{L^2}\mathbb{E}[\|X_1\|^6|\mu_1^\star, \mu_0^\star] + \frac{4\lambda^2}{L^2}\sum_{\ell=2}^{L}\mathbb{E}[\|\langle X_1, X_\ell\rangle X_\ell\|^2|\mu_1^\star, \mu_0^\star] \\
&+ \frac{8\lambda^2}{L^2}\sum_{\ell=2}^{L}\mathbb{E}[\|X_1\langle X_1, X_\ell\rangle\|^2||\mu_1^\star, \mu_0^\star] + \frac{8\lambda^2}{L^2}\sum_{2\le \ell<k\le L}\mathbb{E}[\langle X_1, X_\ell\rangle\langle X_1, X_k\rangle\langle X_\ell, X_k\rangle|\mu_1^\star, \mu_0^\star].
\end{aligned}
$$

Furthermore we have the following, (P: *proofs?*)

$$
\begin{aligned}
\mathbb{E}[\|X_1\|^2|\mu_1^\star, \mu_0^\star] &= 1 + \sigma^2 d, \\
\mathbb{E}\left[\langle X_1, X_2\rangle^2|\mu_1^\star, \mu_0^\star\right] &= \frac{1}{2} + 2\sigma^2 + d\sigma^4, \\
\mathbb{E}[\|\langle X_1, X_2\rangle X_2\|^2|\mu_1^\star, \mu_0^\star] &= \frac{1}{2} + \frac{d+8}{2}\sigma^2 + 3(d+2)\sigma^4 + d(d+2)\sigma^6, \\
\mathbb{E}[\|X_1\|^4|\mu_1^\star, \mu_0^\star] &= 1 + 2(d+2)\sigma^2 + d(d+2)\sigma^4, \\
\mathbb{E}[\|X_1\|^6|\mu_1^\star, \mu_0^\star] &= 1 + 3(d+4)\sigma^2 + 3(d+2)(d+4)\sigma^4 + d(d+2)(d+4)\sigma^6, \\
\mathbb{E}[\langle X_1, X_2\rangle\langle X_1, X_3\rangle\langle X_2, X_3\rangle|\mu_1^\star, \mu_0^\star] &= 2\left(\sigma^2 + \frac{1}{2}\right)^3 + (d-2)\sigma^6.
\end{aligned}
$$

Since no expression depends on $\mu_1^\star, \mu_0^\star$, we have

$$
\mathbb{E}\left[\left\|X_1 - \frac{2\lambda}{L}\sum_{\ell=1}^{L}\langle X_1, X_\ell\rangle X_\ell\right\|^2\right] = \mathbb{E}\left[\left\|X_1 - \frac{2\lambda}{L}\sum_{\ell=1}^{L}\langle X_1, X_\ell\rangle X_\ell\right\|^2 \middle| \mu_1^\star, \mu_0^\star\right].
$$

And

$$
\begin{aligned}
&\mathbb{E}\left[\left\|X_1 - \frac{2\lambda}{L}\sum_{\ell=1}^{L}\langle X_1, X_\ell\rangle X_\ell\right\|^2\right] \\
&= 1 + \sigma^2 d - \frac{4\lambda}{L}(1 + 2(d+2)\sigma^2 + d(d+2)\sigma^4) - \frac{4\lambda}{L}(L-1)\left(\frac{1}{2} + 2\sigma^2 + d\sigma^4\right) \\
&+ \frac{4\lambda^2}{L^2}(1 + 3(d+4)\sigma^2 + 3(d+2)(d+4)\sigma^4 + d(d+2)(d+4)\sigma^6) \\
&+ \frac{12\lambda^2}{L^2}(L-1)\left(\frac{1}{2} + \frac{d+8}{2}\sigma^2 + (3d+6)\sigma^4 + d(d+2)\sigma^6\right) \\
&+ \frac{8\lambda^2}{L^2}\frac{(L-1)(L-2)}{2}\left(2\left(\sigma^2 + \frac{1}{2}\right)^3 + (d-2)\sigma^6\right).
\end{aligned}
$$

When $L \to \infty$, we obtain

$$\mathbb{E}\left[\left\|X_1 - \frac{2\lambda}{L}\sum_{\ell=1}^{L}\langle X_1, X_\ell\rangle X_\ell\right\|^2\right] =(1+\sigma^2 d) - 2\lambda(1+4\sigma^2+2d\sigma^4) + 4\lambda^2\left(2\left(\sigma^2+\frac{1}{2}\right)^3 + (d-2)\sigma^6\right).$$

And we can choose $\lambda = \frac{1+4\sigma^2+2d\sigma^4}{4\left(2\left(\sigma^2+\frac{1}{2}\right)^3+(d-2)\sigma^6\right)}$ to get

$$\begin{aligned}\mathbb{E}\left[\left\|X_1 - \frac{2\lambda}{L}\sum_{\ell=1}^{L}\langle X_1, X_\ell\rangle X_\ell\right\|^2\right] &= (1+\sigma^2 d) - \frac{(1+4\sigma^2+2d\sigma^4)^2}{4\left(2\left(\sigma^2+\frac{1}{2}\right)^3+(d-2)\sigma^6\right)}\\ &= \sigma^2(d-2)\frac{1+2\sigma^2}{1+6\sigma^2+12\sigma^4+4d\sigma^6}\\ &\leq \sigma^2(d-2).\end{aligned}$$

□

Proof of Proposition 5.2. For $c \in \{0,1\}$, one has

$$\begin{aligned}\mathbb{E}\left[\frac{2\lambda}{L}\sum_{\ell=1}^{L}\langle X_1, X_\ell\rangle X_\ell \middle| \mu_1^\star, \mu_0^\star, Z_1 = c\right] &= \frac{2\lambda}{L}\mathbb{E}[\|X_1\|^2 X_1 | \mu_1^\star, \mu_0^\star, Z_1 = c]\\ &+ \frac{2\lambda(L-1)}{L}\mathbb{E}[\langle X_1, X_2\rangle X_2 | \mu_1^\star, \mu_0^\star, Z_1 = c]\\ &= \frac{2\lambda}{L}\left[(1+(d+2)\sigma^2) + (L-1)\left(\frac{1}{2}+\sigma^2\right)\right]\mu_c^\star.\end{aligned}$$

We remark that choosing $\lambda = \frac{L}{2}\frac{1}{1+(d+2)\sigma^2+(L-1)\left(\frac{1}{2}+\sigma^2\right)}$ we get that the encoding is unbiased.

And

$$\left\|\mathbb{E}\left[\frac{2\lambda}{L}\sum_{\ell=1}^{L}\langle X_1, X_\ell\rangle X_\ell \middle| \mu_1^\star, \mu_0^\star, Z_1 = c\right]\right\|^2 = \frac{4\lambda^2}{L^2}\left[1+(d+2)\sigma^2 + (L-1)\left(\frac{1}{2}+\sigma^2\right)\right]^2.$$

Besides,

$$\mathbb{E}\left[\frac{2\lambda}{L}\sum_{\ell=1}^{L}\langle X_1, X_\ell\rangle X_\ell \middle| \mu_1^\star, \mu_0^\star\right] = \frac{\lambda}{L}\left[1+(d+2)\sigma^2 + (L-1)\left(\frac{1}{2}+\sigma^2\right)\right](\mu_0^\star+\mu_1^\star)$$

Also,

$$\left\|\mathbb{E}\left[\frac{2\lambda}{L}\sum_{\ell=1}^{L}\langle X_1, X_\ell\rangle X_\ell \middle| \mu_1^\star, \mu_0^\star\right]\right\|^2 = \frac{2\lambda^2}{L^2}\left[1+(d+2)\sigma^2 + (L-1)\left(\frac{1}{2}+\sigma^2\right)\right]^2$$

On the other hand,

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{2\lambda}{L}\sum_{\ell=1}^{L}\langle X_1, X_\ell\rangle X_\ell\right\|^2 \middle| \mu_1^\star, \mu_0^\star, Z_1 = c\right] &= \frac{4\lambda^2}{L^2}\mathbb{E}[\|X_1\|^6 | \mu_1^\star, \mu_0^\star, Z_1 = c]\\ &+ \frac{12\lambda^2}{L^2}(L-1)\mathbb{E}[\|\langle X_1, X_2\rangle X_2\|^2 | \mu_1^\star, \mu_0^\star, Z_1 = c]\\ &+ \frac{8\lambda^2}{L^2}\frac{(L-1)(L-2)}{2}\mathbb{E}[\langle X_1, X_2\rangle\langle X_1, X_3\rangle\langle X_2, X_3\rangle | \mu_1^\star, \mu_0^\star, Z_1 = c].\end{aligned}$$

Recalling the expressions stated at the beginning of the proof of Proposition 5.1 for moments of Gaussian r.v. , we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{2\lambda}{L}\sum_{\ell=1}^{L}\langle X_1, X_\ell\rangle X_\ell\right\|^2 \middle| \mu_1^\star, \mu_0^\star, Z_1 = c\right] \\
&= \frac{4\lambda^2}{L^2}(1 + 3(d+4)\sigma^2 + 3(d+2)(d+4)\sigma^4 + d(d+2)(d+4)\sigma^6) \\
&+ \frac{12\lambda^2}{L^2}(L-1)\left(\frac{1}{2} + \frac{(d+8)}{2}\sigma^2 + 3(d+2)\sigma^4 + d(d+2)\sigma^6\right) \\
&+ \frac{8\lambda^2}{L^2}\frac{(L-1)(L-2)}{2}\left(2\left(\sigma^2 + \frac{1}{2}\right)^3 + (d-2)\sigma^6\right).
\end{aligned}$$

And

$$\begin{aligned}
&\mathrm{Var}\left[\frac{2\lambda}{L}\sum_{\ell=1}^{L}\langle X_1, X_\ell\rangle X_\ell \middle| \mu_1^\star, \mu_0^\star, Z_1 = c\right] \\
&= \mathbb{E}\left[\left\|\frac{2\lambda}{L}\sum_{\ell=1}^{L}\langle X_1, X_\ell\rangle X_\ell\right\|^2 \middle| \mu_1^\star, \mu_0^\star, Z_1 = c\right] - \left\|\mathbb{E}\left[\frac{2\lambda}{L}\sum_{\ell=1}^{L}\langle X_1, X_\ell\rangle X_\ell \middle| \mu_1^\star, \mu_0^\star, Z_1 = c\right]\right\|^2 \\
&= \frac{4\lambda^2}{L^2}(1 + 3(d+4)\sigma^2 + 3(d+2)(d+4)\sigma^4 + d(d+2)(d+4)\sigma^6) \\
&+ \frac{12\lambda^2}{L^2}(L-1)\left(\frac{1}{2} + \frac{(d+8)}{2}\sigma^2 + 3(d+2)\sigma^4 + d(d+2)\sigma^6\right) \\
&+ \frac{8\lambda^2}{L^2}\frac{(L-1)(L-2)}{2}\left(2\left(\sigma^2 + \frac{1}{2}\right)^3 + (d-2)\sigma^6\right) \\
&- \frac{4\lambda^2}{L^2}\left[1 + (d+2)\sigma^2 + (L-1)\left(\frac{1}{2} + \sigma^2\right)\right]^2.
\end{aligned}$$

When $L \to \infty$,

$$\begin{aligned}
&\mathrm{Var}\left[\frac{2\lambda}{L}\sum_{\ell=2}^{L}\langle X_1, X_\ell\rangle X_\ell \middle| \mu_1^\star, \mu_0^\star, Z_1 = c\right] \\
&\sim 4\lambda^2\left(2\left(\sigma^2 + \frac{1}{2}\right)^3 + (d-2)\sigma^6\right) - 4\lambda^2\left(\sigma^2 + \frac{1}{2}\right)^2 \\
&= 2\lambda^2\sigma^2(1 + 4\sigma^2 + 2d\sigma^4).
\end{aligned}$$

Choosing the $\lambda = \frac{1}{1+2\sigma^2}$, we have an unbiased encoding with variance

$$2\sigma^2\frac{1 + 4\sigma^2 + 2d\sigma^4}{(1+2\sigma^2)^2}$$

□