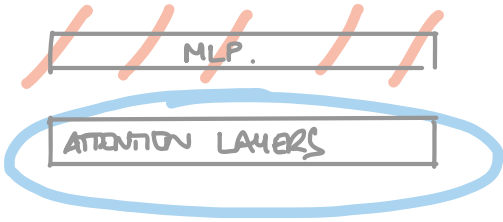
Attention-based Clustering

quantization ?

---

joint work with Pierre Marion & Rodrigo Maulen-Soto.

Context: transformer architecture



OMITTED.

FOCUS

Growing literature:

- Mainly in supervised learning
- Via in-context learning (see the work of Peter Bartlett's team)
- Other work with Pierre Marion, Raphaël Berthier and Gérard Biau where we defined a new statistical task called "single location regression" in which only one token is relevant to form the prediction, but its position is random (and therefore prompt-dependent)..

↦ some link with the ongoing work of Alberto Bietti last week.

Motivations:

- Understand the statistical abilities of attention layers in unsupervised scenarios
- Why? Powerful embedding performed by attention layers -
- Existing work by He et al (2025) showing that transformers

can mimic the EM algorithm . But their analysis requires the knowledge of the labels during training.

- "Clustering" phenomenon exhibited by the team of Letrouit, Jenkovski Polyanskiy & Rigollet. But this is a completely ≠ setting
  - ✗ they stack an infinite number of layers and study the continuous dynamics where time ≡ depth
  - ✗ no training of the parameters

THIS TALK: Focus on a single attention layer fully trained in an unsupervised way.

SETTING : MODEL-BASED CLUSTERING

- Input sequence of L tokens where each token $X_\ell \in \mathbb{R}^d$.

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

- Assumptions :

$$X_\ell \overset{iid}{\sim} \frac{1}{2}\mathcal{N}(\mu_0^*, \sigma^2 I) + \frac{1}{2}\mathcal{N}(\mu_1^*, \sigma^2 I) \quad (P_{\sigma^2})$$

Mixture model with balanced components

$$\pi_0 = \pi_1 = 1/2$$

centroids as unit vectors

$$\mu_0^*, \mu_1^* \in \mathbb{S}^{d-1}.$$
$$\|\mu_0^*\|_2 = \|\mu_1^*\| = 1$$

- Orthogonality assumption $\mu_0^* \perp \mu_1^*$ or $\langle \mu_0^*, \mu_1^* \rangle = 0$

- For each token X_ℓ, there exists an associated latent variable Z_ℓ such that $Z_\ell \sim \mathcal{B}(1/2)$ encodes the corresponding cluster of X_ℓ.

- Degenerate case : $\sigma^2 = 0$ $\quad X_\ell \sim \frac{1}{2}\delta_{\mu_0^*} + \frac{1}{2}\delta_{\mu_1^*}$.

ATTENTION-BASED PREDICTOR / EMBEDDING

Attention head : $H^{\text{soft},\lambda}(X) = \text{softmax}_\lambda\left(XQK^TX^T\right)XV$

$$K, Q, V \in \mathbb{R}^{d\times p}$$

Step 1 : Take $V = Id$. (no training for V)

Step 2 : $K = Q = \mu$ a column matrix $\mu^T \in \mathbb{R}^{1\times d}$

$$H^{\text{soft},\lambda,\mu}(X) = \text{softmax}_\lambda\left(X\mu\mu^TX^T\right)X$$

The ℓ-th output vector is therefore given by

$$H^{\text{soft},\lambda,\mu}(X)_\ell = \sum_{k=1}^{L} \text{softmax}_\lambda\left(X_\ell^T\mu\mu^TX\right)_k X_k$$

= we aggregate the X_k's when X_k and X_ℓ are simultaneously aligned with μ.

Step 3: Getting rid of the softmax

$$H^{\text{lin},\mu}(X)_\ell = \frac{2}{L} \sum_{k=1}^{L} (\lambda X_\ell^T \mu \mu^T X_k) X_k$$

Fix $\mu = \mu_0^*$. When X_ℓ and X_k belong to the cluster 0, i.e., such that $z_\ell = z_k = 0$, then the vectors X_ℓ and X_k are likely to be aligned with μ_0^*

" $(X_\ell^T \mu_0^* \mu_0^{*T} X_k) X_k \simeq X_\ell$ "

Conversely, if X_ℓ and X_k are associated with $\neq$ clusters, i.e., $z_\ell \neq z_k$ then " $(X_\ell^T \mu_0^* \mu_0^{*T} X_k) X_\ell \simeq 0$ "

Remark: If $\mu = \mu_0^*$

$X_\ell \in$ cluster 0 $(z_\ell = 0)$

then $\sum_k \lambda (X_\ell^T \mu_0^* \mu_0^{*T} X_k) X_k$ aggregates the X_k's from the same cluster, whose expected number is $L/2$ $\rightsquigarrow$ renormalization factor

$H^{\text{lin},\mu_0^*}(X)_\ell \equiv$ "empirical mean" of the tokens belonging to the same cluster.

Assuming that the # clusters is known, it seems natural to consider an attention-based layer composed of 2 attention heads

$$T^{\text{lin},\mu_0,\mu_1}(X) = H^{\text{lin},\mu_0}(X) + H^{\text{lin},\mu_1}(X)$$

Metric loss As no label is available, one may consider

$$\mathcal{L}(T) = \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}\,\|X_\ell - T(X)_\ell\|_2^2$$

$$= \mathbb{E}\,\|X_1 - T(X)_1\|_2^2$$

- Note that if T was able to return for each token X_ℓ the corresponding centroid $\mu^*_{z_\ell}$, the risk would correspond to a QUANTIZATION ERROR.
- It also looks like the kind of loss used during self-supervised learning but $\neq$

Training of the attention layer Set $\boxed{\mathcal{R}(\mu_0,\mu_1) = \mathcal{L}(T^{\text{lin},\mu_0,\mu_1})}$

PGD iterates on $(\mathbb{S}^{d-1})^2$

$$\mu_0^{k+1} = \frac{\mu_0^k - \gamma(I - \mu_0^k(\mu_0^k)^\top)\nabla_{\mu_0}\mathcal{R}(\mu_0^k,\mu_1^k)}{\|\mu_0^k - \gamma(I - \mu_0^k(\mu_0^k)^\top)\nabla_{\mu_0}\mathcal{R}(\mu_0^k,\mu_1^k)\|_2}$$

$$\mu_1^{k+1} = \frac{\mu_1^k - \gamma(I - \mu_1^k(\mu_1^k)^\top)\nabla_{\mu_1} R(\mu_0^k, \mu_1^k)}{\|\mu_1^k - \gamma(I - \mu_1^k(\mu_1^k)^\top)\nabla_{\mu_1} R(\mu_0^k, \mu_1^k)\|_2}$$

TRAINING DYNAMICS IN THE DEGENERATE CASE.

- Reparameterization $\quad K_0 = \langle \mu_0, \mu_0^* \rangle \qquad K_1 = \langle \mu_1, \mu_1^* \rangle$

  $\eta_0 = \langle \mu_1, \mu_0^* \rangle \qquad \eta_1 = \langle \mu_0, \mu_1^* \rangle$

- Expression of the risk as a polynomial of K_0, K_1, η_0, η_1. (order 4)

- Characterization of global minima

  When $\lambda = \frac{L+1}{L+3}$*, argmin of the risk characterized by

  $$K_0^2 + \eta_0^2 = 1 \qquad K_1^2 + \eta_1^2 = 1 \qquad K_0\eta_1 + K_1\eta_0 = 0.$$

  * computation can be conducted to any value of λ.

- Convergence analysis

Thm: Under the degenerate mixture model, take $\lambda \in]0; \frac{\lambda+1}{\lambda+3}]$,

$\exists \bar{\gamma} > 0$ such that for any stepsize $0 < \gamma < \bar{\gamma}$

for a generic init° $(\mu_0^0, \mu_1^0) \in \tilde{\mathcal{M}} := \{(\mu_0, \mu_1) \in (\mathbb{S}^{d-1})^2 : \langle \mu_0^*, \mu_1 \rangle = 0, \langle \mu_1^*, \mu_0 \rangle = 0, \langle \mu_0, \mu_1 \rangle = 0\}$

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

- Despite the non-convexity of the pb, the key/query matrices trained via PGD align with the centroids of the underlying Dirac mixture.
- Initialization on $\tilde{\mathcal{M}}$ impractical !!
- In practice
  - (a) CV to the centroids observed for init° on $\tilde{\mathcal{M}}$
  - (b) Stagnation ——//—— $(\mathbb{S}^{d-1})^2$
  - (c) To mitigate this effect, we introduce a regularization

$$r(\mu_0, \mu_1) = \mathbb{E}\left[\langle \mu_0, X_1\rangle^2 \langle \mu_1, X_1\rangle^2\right]$$

so that training is done via PSGD for

$$\min_{\mu_0, \mu_1 \in (\mathbb{S}^{d-1})^2} \mathcal{R}(\mu_0, \mu_1) + \rho\, r(\mu_0, \mu_1)$$

ENFORCES THE ORTHOGONALITY AND HELPS THE ATTENTION PARAMETERS TO CONVERGE TO THE CENTROIDS WHEN INITIALIZED ON THE SPHERE

SKIPPED.

TRAINING DYNAMICS IN THE NON-DEGENERATE CASE

- Setting $\begin{cases} X_i \sim \frac{1}{2}\mathcal{N}(\mu_0^*, \sigma^2 I) + \frac{1}{2}\mathcal{N}(\mu_1^*, \sigma^2 I) \\ \mu_0^*, \mu_1^* \in \mathbb{S}^{d-1} \quad \text{and} \quad \langle \mu_0^*, \mu_1^* \rangle = 0 \end{cases}$

- Reparameterization $\quad K_0 = \langle \mu_0, \mu_0^* \rangle \qquad K_1 = \langle \mu_1, \mu_1^* \rangle$

$$R_0 = \langle \mu_1, \mu_0^* \rangle \qquad R_1 = \langle \mu_0, \mu_1^* \rangle$$

$$S = \langle \mu_1, \mu_0 \rangle$$

- Risk ≡ polynomial of order 4 of these quantities

- CV analysis

value of λ s.t $(\pm 1, \pm 1)$ are global min ↓

thm: Under the GMM, take $\lambda \in]0; \lambda^*(\sigma, L)]$

$\exists \bar{\gamma}$ s.t for any step size $0 < \gamma < \bar{\gamma}$, for a generic init° $(\mu_0^0, \mu_1^0) \in \mathcal{M} := \left\{ (\mu_0, \mu_1) \in (\mathbb{S}^{d-1})^2 : \begin{array}{l} \langle \mu_0^*, \mu_1 \rangle = 0 \\ \langle \mu_1^*, \mu_0 \rangle = 0 \\ \langle \mu_1, \mu_0 \rangle = 0 \end{array} \right\}$

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

✓ Attention layer can uncover and encode latent structure of the input distribution in a fully unsupervised setting

✓ Interpretability of the attention parameters.

- Despite the non-convexity of the pb, the key/query vectors trained via PGD align with the underlying centroids of the mixture.
- Initialization on $\mathcal{M}$ impractical !!
- In practice
  - (a) CV to the centroids observed for init° on $\mathcal{M}$
  - (b) ~~Stagnation~~ ——"—— $(\mathbb{S}^{d-1})^2$
  - (c) To mitigate this effect, we introduce a regularization

$$r(\mu_0, \mu_1) = \mathbb{E}\left[\langle \mu_0, x_1\rangle^2 \langle \mu_1, x_1\rangle^2\right]$$

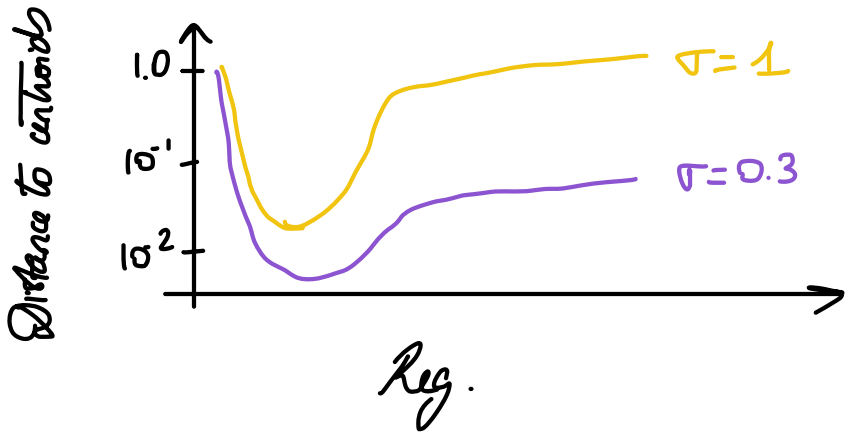
suggested by the analysis in the degenerate case

so that training is done via PGD for

$$\min_{\mu_0, \mu_1 \in (\mathbb{S}^{d-1})^2} \mathcal{R}(\mu_0, \mu_1) + \rho\, r(\mu_0, \mu_1)$$

ENFORCES THE ORTHOGONALITY AND HELPS THE ATTENTION PARAMETERS TO CONVERGE TO THE CENTROIDS WHEN INITIALIZED ON THE SPHERE

- XP



ATTENTION-BASED LAYERS AS APPROXIMATE QUANTIZERS

- Motivations: understanding the statistical properties of an attention layer when parameters have converged to the true centroids.

- Consider the "optimal" quantizer $T^*(X)_l = \mu^*_{z_l}$.

- The risk of the optimal quantizer $\mathcal{L}(T^*) = \mathbb{E}\left[\|X_1 - \mu^*_{z_1}\|^2\right] = d\sigma^2$.

Lemma: Under the GMM, it holds that

$$\mathbb{E}\left[T^{\mathrm{lin},\mu_0^*,\mu_1^*}(X)_1 \mid z_1 = c\right] = \mu_c^* \, \frac{\lambda}{L}\left[(L+1) + 2(L+3)\sigma^2\right]$$

Choosing $\lambda = \frac{L}{(L+1)+2(L+3)\sigma^2}$ gives $\mathbb{E}\left[T^{\mathrm{lin},\mu_0^*,\mu_1^*}(X)_1 \mid z_1 = c\right] = \mu_c^*$. (*)

✓ The l-th output aligns on average with the centroid of the cluster to which the l-th token belongs.

Prop: Under the GMM, fix $\lambda = \frac{1+4\sigma^2+4\sigma^4}{1+6\sigma^2+12\sigma^4+8\sigma^6}$,

then

$$\mathcal{L}\left(T^{lin,\mu_0^*,\mu_1^*}\right) \underset{L\to+\infty}{\sim} (d-2)\sigma^2$$

Striking obs°

$$\frac{\mathcal{L}\left(T^{lin,\mu_0^*,\mu_1^*}\right)}{\mathcal{L}\left(T^*\right)} \underset{L\to+\infty}{\sim} 1-\frac{2}{d}.$$

!! For long input sequences, attention layer can achieve a better risk than the optimal quantizer

!! But the comparison is not entirely fair:

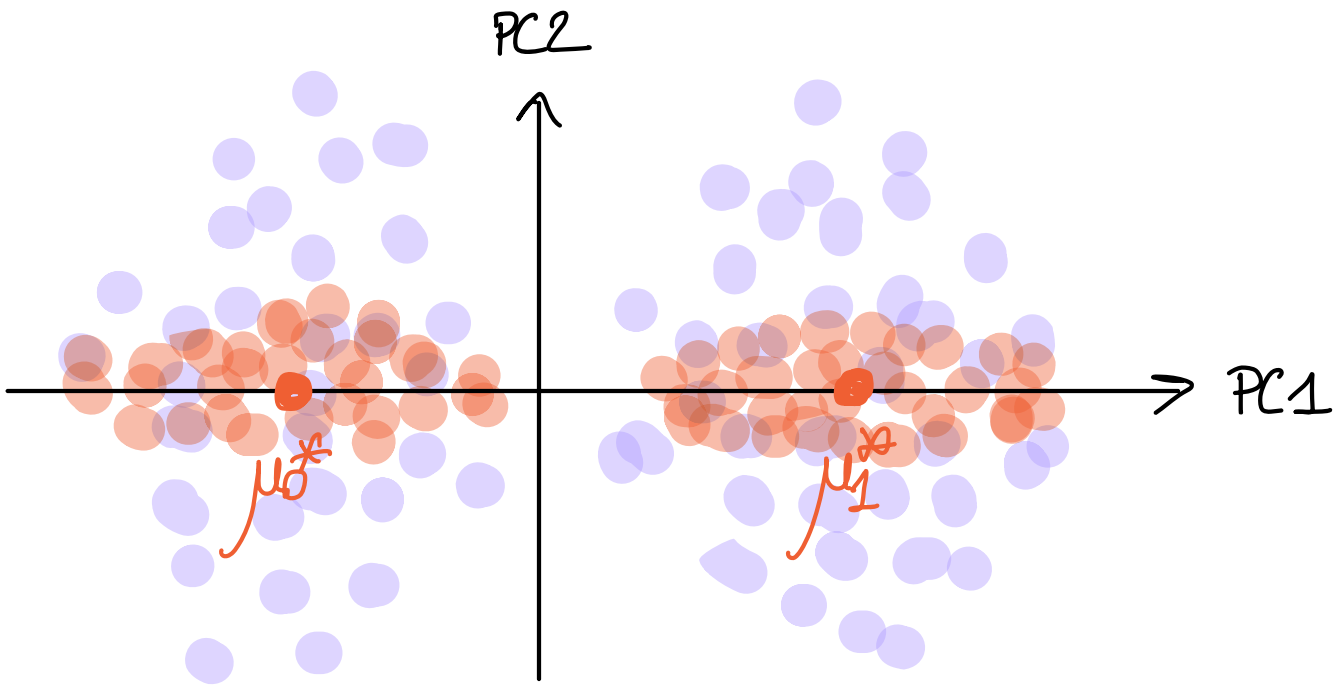
① T^{lin,μ_0^*,μ_1^*} is not a quantizer per se

② The optimal quantizer T^* relies on a single input token to predict the associated centroid (being oracle w.r.t centroids)

(vs. the oracle attention layer T^{lin,μ_0^*,μ_1^*} uses the info of the underlying centroids but benefits from a growing sequence of random var., all drawn from the same Gaussian mixture. ⇝ VARIANCE REDUCTION MECHANISM.

(*) with the same choice of λ: $\mathrm{Var}\left[T^{lin,\mu_0^*,\mu_1^*}(X)_1 \mid z_1\right] = 2\sigma^2$ instead of d

Illustration of the variance reduction mechanism.



EXPLORATION : IN-CONTEXT CLUSTERING / QUANTIZATION

- Up to now, with 2 heads, the transformed 1st token is

$$T^{lin,\mu_0,\mu_1}(X)_1 = \frac{2}{L}\sum_{\ell=1}^{L} a\left(x_1^T \mu_0 \mu_0^T x_\ell\right)x_\ell + a\left(x_1^T \mu_1 \mu_1^T x_\ell\right)x_\ell .$$

$$= \frac{2}{L}\sum_{\ell=1}^{L} a\left[x_1^T \underbrace{\left(\mu_0\mu_0^T + \mu_1\mu_1^T\right)}_{\text{when } \mu_0 \perp \mu_1 \text{ this is a rank-2 matrix}} x_\ell\right]x_\ell$$

- Remark : we just showed that we could train rank-2 key / query head by leveraging the non-convex optimization of 2 simple raw structured heads .

- Challenge : IN-CONTEXT FRAMEWORK

Imagine now that for each input sequence, the centroids μ_0^* and μ_1^* are random .

If they are random but still aligned with particular directions μ_0^{**} and μ_1^{**}, then T^{lin,μ_0,μ_1} should remain a good "quantizer" / "embedding" candidate .

Imagine now that for n sequences of L tokens, the centroids are randomly drawn on the sphere :

$$\mu_{i0}^* \sim \mathcal{U}(\mathbb{S}^{d-1}) \qquad \mu_{i1}^* \,|\, \mu_{i0}^* \text{ distributed on } \mathbb{S}^{d-1} \cap (\mu_{0i}^*)^{\perp}.$$

✗ T^{lin,μ_0,μ_1} will struggle to capture the changing structure of the data because of the use of only 2 param. $\mu_0, \mu_1 \in \mathbb{S}^{d-1}$

Idea : Increase the degrees of freedom of the attention layer

take $\mu_1, \dots \mu_d \in \mathbb{S}^{d-1}$ of | unit-norm
| mutually orthogonal

!!

In such a case, the attention layer has no more parameters to learn !!

$$T^{ctx}(X)_\ell = \sum_{c=1}^{d} H^{lin,\mu_c}(X)_1 = \frac{2\lambda}{L} \sum_{c=1}^{d} \sum_{k=1}^{L} \left(X_\ell^T \mu_c \mu_c^T X_k\right) X_k$$

$$= \frac{2\lambda}{L} \sum_{k=1}^{L} X_\ell^T \left(\sum_{c=1}^{d} \mu_c \mu_c^T\right) X_k X_k$$

$$= \frac{2\lambda}{L} \sum_{k=1}^{L} X_\ell^T X_k X_k. \qquad = \frac{2\lambda}{L} \left[\sum_{k=1}^{L} X_k X_k^T\right] X_\ell$$

($\sum_{c=1}^{d} \mu_c \mu_c^T = Id$; $\sum_{k=1}^{L} X_k X_k^T = \hat{\Sigma}$)

(To simplify, we could remove the "autocorrelated" term

$$T^{ctx}(x)_\ell = \frac{2\lambda}{L} \sum_{k \neq \ell} x_\ell^T x_k x_k \cdot$$

)

- What can be said about this embedding?

$$\mathbb{E}\left[T^{ctx}(x)_\ell \mid \mu_1^*, \mu_0^*, z_\ell = c \right] = \frac{2\lambda}{L} C_{4,d,\sigma^2} \mu_c^* \cdot$$

$C_{4,d,\sigma^2} \to 1 + (d+2)\sigma^2 + (d+1)\left(\frac{1}{2} + \sigma^2\right)$

✓ Aligned with $\mu_{z_\ell}^*$ in expectation

✓ One can choose λ to get "unbiased" encoding with centroids

$$\mathrm{Var}\left[T(x)_\ell \mid \mu_1^*, \mu_0^*, z_\ell = c \right] \underset{L \to \infty}{\sim} 2\lambda^2\sigma^2\left(1 + 4\sigma^2 + 2\sigma^4\right)$$

Choosing $\lambda = 1/1 + 2\sigma^2$ (unbiased encoding when $L \to +\infty$)

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

With an appropriate choice of the temperature σ, one can show that

$$\frac{1+2\sigma^2}{1+6\sigma^2+12\sigma^4+4d\sigma^6} \leq 1.$$

$$\lim_{L\to+\infty} \frac{\mathcal{L}(T^{ctx})}{\mathcal{L}(T^*)} = \left(1-\frac{2}{d}\right) C_{\sigma,d} \leq 1-\frac{2}{d}.$$

$C_{\sigma,d} \leq 1$

Better quantization risk than the optimal quantizer and the improvement factor is now dimension-dependent.

Identification of a similar phenomenon

- in a more complex setting (in-context)
- with a simpler architecture (no trained parameters)

This simple (non-parametric) encoder manages to capture structure of the data