From softmax mixture ensembles to mixtures of experts, with applications to LLM output summarization

Florentina Bunea,

Department of Statistics and Data Science, Cornell University

Statistics + Data Science

S. Strimas-Mackey, Google; *X. Bing*, U. Toronto; *N. Bétache*, Cornell; *J. Niles-Weed*, NYU; *M. Wegkamp*, Cornell

Overview

1. Steady progress made in LLM text generation.

   Not this talk !

2. LLM evaluation lags behind. Many heuristics.

   - Comparing LLM output with (human) reference requires **corpora summarization.**
   - Quality of comparison hinges on quality of **summarization**.

     Challenge and goal: Do not use the billion parameter generative model !
   - Statistical models for corpora **summarization** are emerging.

     New metrics for corpora comparison based on these summaries are emerging. Stay tuned !

3. This talk: Interpretable corpora summarization via mixture ensembles.

   Corpus summary = mixing measure

From analog to LLM's

Topic models

Ensembles of generic discrete mixtures

2000 - 2020+

Softmax-parametrized topic models

Ensembles of softmax mixtures

2022 - 2025+

A simple version of mixture of experts ensembles

Ensembles of mixtures of continuous mixtures

2025+

What movie genres are they reviewing ?

Domain (To discover)	Interpretation (To discover)	Movie ID	Document excerpt
1	?	37,123	*This was an OK movie, at best, outside the context of the book. But having read and enjoyed the book quite a bit it was a real disappointment in comparison . . .*
		15,709	*This has always been one of my favourite books. I was thrilled when I saw that the book had been made into a movie, for the first time since it was written, over 50 years before. . .*
2	?	12,445	*This was the most disappointing films I have ever seen recently. And I really hardly believe that people say goods things about this very bottom film!. . .*
		32,442	*Acting was awful. Photography was awful. Dialogue was awful. Plot was awful. (I'm not being mean here...It really was this bad.). . .*
3	?	23,753	*This game really is worth the ridiculous prices out there. The graphics really are great for the SNES, though the magic spells don't look particularly great. . .*
		12,261	*I remember playing this game at a friend. Watched him play a bit solo until we decided to try play 2 and 2, which we found out how to do. . .*
4	?	29,114	*After watching such teen horror movies as Cherry Falls and I know what you did last summer, I expected this to be similar. . .*
		3,448	*Being a HUGE fan of the horror genre, I have come to expect and appreciate cheesey acted, plot-holes galore, bad scripts. . .*
?	?	xxxx	*Available text*

Topic model = a non-parametrized ensemble of related discrete mixtures

An "analog" story

- Data = Corpus = Ensemble of n documents
  - Document i = "Bag of words" representation
    = $Y^{(i)} \sim \text{Multinomial}_p(N, \pi^{(i)})$
  - True word distribution $\pi^{(i)} \in \Delta_p$ relative to given vocabulary of size p.
- Most basic model for the ensemble = Topic model
  - $\pi^{(i)} = \sum_{k=1}^{K} \alpha_k^{(i)} A_k, \quad i \in [n].$
  - Document word frequencies driven by corpus topics $A_k \in \Delta_p$.
  - Topics A_k are latent, and common to the corpus.

How to fit a topic model ? Many choices, 20+ years of research

- Bayesian methods
  Blei (2003++, 2022); Nguyen (2013); Ho et al (2016+)

- NMF-inspired methods
  - Π admits an NMF $\Pi = AT$ that **is** unique and learnable in poly time under (sufficient) restrictions on A and T.

    Most popular: topic models with anchor words or anchor documents.
  - Existing methods learn a decomposition of Π under these restrictions.
    - Simplex based methods. Arora et al (2012+) ; Bitorff et al, 2012; Ke et al.(2020+); Fan, J. et al (2021+); Klopp et al (2023+)
    - A specialized second moment matching method.
      Bing, Bunea, Wegkamp (2020 a, b)

Theoretical guarantees for computationally efficient estimates in identifiable topic models

1. Minimax optimal latent topic estimation

$$\max_{1\le k\le K} \|\widehat{A}_{\cdot k} - A_{\cdot k}\|_1 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{pK}{nN}}\right).$$

Bing, Bunea, Wegkamp (2020 a, b)

2. For each document i, the MLE $\widehat{\alpha}^{(i)}$ is sparse if $\alpha^{(i)}$ is sparse, w.h.p.

$$\text{If} \quad \widehat{\alpha}^{(i)} := \operatorname{argmax}_{\alpha\in\Delta_K} \sum_{j=1}^{p} Y_j^{(i)} \log\left(\widehat{A}_{j\cdot}^{\top}\alpha\right),$$

$$\text{then} \quad \operatorname{supp}(\widehat{\alpha}^{(i)}) \subseteq \operatorname{supp}(\alpha^{(i)}).$$

Bing, Bunea, Strimas-Mackey, Wegkamp (2022)

3. An appropriate one-step update $\widetilde{\alpha}^{(i)}$ of $\widehat{\alpha}^{(i)}$ is asymptotically normal.

$$\lim_{n,N\to\infty} \sqrt{N}\,\Sigma^{+\frac{1}{2}}\left(\widetilde{\alpha}^{(i)} - \alpha\right) \xrightarrow{d} \mathcal{N}_K\left(0, \begin{bmatrix} \boldsymbol{I}_{K-1} & 0 \\ 0 & 0 \end{bmatrix}\right).$$

Bing, Bunea, Niles-Weed (2024)

Intuition: the MLE of a sparse mixture weight vector is sparse w.p. 1

$$(Y_1, Y_2, Y_3) \sim \text{Multinomial}_3(N, \pi),$$

$$\pi = \sum_{k=1}^{2} \alpha_k A_k, \quad \text{with} \quad A_1 = (a_1, a_2, 0)^\top, \quad A_2 = (0, 0, 1)^\top, \quad a_1 + a_2 = 1, \quad a_1, a_2 \geq 0.$$

$$Y_1 + Y_2 \sim \text{Bin}(N, \alpha_1), \quad Y_3 = \text{Bin}(N, \alpha_2).$$

$$\widehat{\alpha}_{\text{mle},1} = \frac{Y_1 + Y_2}{N}, \quad \widehat{\alpha}_{\text{mle},2} = \frac{Y_3}{N}.$$

If $\alpha_k = 0$, for some $k \in \{1, 2\}$, then $\widehat{\alpha}_{\text{mle},k} = 0$, w.p. 1.

If data is generated from a sparse discrete mixture, $\widehat{\alpha}_{\rm mle}$ can have zero entries.

Use topic models to learn **broad** latent domain representations

Discovered domain	Discovered interpretation	Movie ID	Discovered movie cluster (review excerpt)
1	Book Adaptations	37,123	*This was an OK movie, at best, outside the context of the book. But having read and enjoyed the book quite a bit it was a real disappointment in comparison . . .*
		15,709	*This has always been one of my favourite books. I was thrilled when I saw that the book had been made into a movie, for the first time since it was written, over 50 years before. . .*
2	Negative reviews	12,445	*This was the most disappointing films I have ever seen recently. And I really hardly believe that people say goods things about this very bottom film!. . .*
		32,442	*Acting was awful. Photography was awful. Dialogue was awful. Plot was awful. (I'm not being mean here...It really was this bad.). . .*
3	Game-related	23,753	*I won't get into the computer game style of the movie ... i have never seen a movie that tried on the computer game-story and succeeded. . . .*
	(video and movies)	12,261	*This is the second best game ever only beaten by Final Fantasy VII. I have found this one of the more boring games the FMV's are even better than this game but there is no story. . . .*
4	Horror	29,114	*After watching such teen horror movies as Cherry Falls and I know what you did last summer, I expected this to be similar. . .*
		3,448	*Being a HUGE fan of the horror genre, I have come to expect and appreciate cheesy acted, plot-holes galore, bad scripts. . .*
5	TV Shows	32,315	*I used to watch this show when I was a little girl. . .*
		10,454	*I'v watched the TV show Hex twice over and I still can not get enough of it. The show is excellent. . .*
6	Historical	6,709	*Carlo Levi, an Italian who fought against the arrival of Fascism in his native Torino, was arrested for his activities. . .*

Curated Dataset:

Content Type	Count
Movie Reviews	58
Game Reviews	58
Total	**116**

Examples

- Movies: Final Fantasy film, Resident Evil movie, Street Fighter
- Games: Final Fantasy VIII, Metal Gear Solid, Tomb Raider III

A standard topic model with $K = 2$ on the curated data set **is not enough**

Topic 1 (mixed review types)

"Game" and "movie" co-occur.

Top Words:

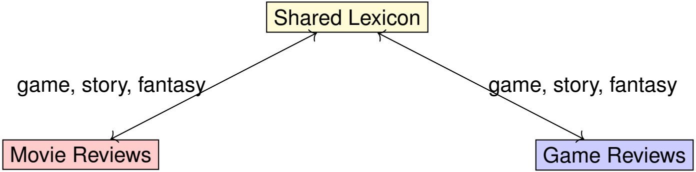
1. game (14.06%)
2. movie (8.13%)

Topic 2 (mixed review types)

"Game" and "movie" co-occur.

Top Words:

1. game (17.9%)
2. movie (5%)



Impossible to distinguish between the two topics based on word frequency alone.

Standard topic models: the good and the bad

1. The Good
   - Excellent tools for **preliminary** data understanding.
   - Induce a **summary** of the corpus, the mixing measure

$$\sum_{k=1}^{K} \bar{\alpha}_k \delta_{A_{\cdot k}}.$$

   - Extensive theoretical and computational understanding.
2. The Bad
   - By definition, they cannot incorporate covariates.
   - The "bag of words" representation loses document context.

Softmax mixture ensembles

$Y^{(i)}$ discrete r.v. supported on $\{\text{word}_1, \ldots, \text{word}_p\}$, for each $i \in [n]$.

$$\mathbb{P}(\text{Word } j \text{ in doc } i) \underbrace{=}_{\textit{Topic model assumption}} \sum_{k=1}^{K} \mathbb{P}(\text{ Topic } k \text{ in doc } i\ )\, \mathbb{P}(\text{Word } j \mid \text{Topic } k)$$

$$\mathbb{P}(Y^{(i)} = \text{word}_j) =: \pi_j^{(i)} = \sum_{k=1}^{K} \alpha_k^{(i)} A_{jk}, \quad j \in [p], \quad i \in [n].$$

$Y^{(i)}$ discrete r.v. supported on given $\{x_1, \ldots, x_p\}$ in $\mathbb{R}^L$, for each $i \in [n]$.

$$\mathbb{P}(Y^{(i)} = x_j) =: \pi^{(i)}(\omega^{(i)} \mid x_j) = \sum_{k=1}^{K} \alpha_k^{(i)} A_k(x_j)$$

$$=: \sum_{k=1}^{K} \alpha_k^{(i)} \frac{\exp(x_j^\top \theta_k)}{\sum_{\ell=1}^{p} \exp(x_\ell^\top \theta_k)}, \qquad j \in [p], \quad i \in [n].$$

$\theta_1 \ldots, \theta_K \in \mathbb{R}^L$ are the latent domain parameters; $\quad L < p$.

Softmax mixture ensembles with embedded dictionary

1. Before 2017

   - Generic-context word embeddings (word2vec).
     Each **word** in a vocabulary mapped to a **unique vector in** $\mathbb{R}^L$.
   - Documents = (still) "bag of words" representation, now parametrized.
   - Gain: dimension reduction, with $L < p$.

2. After 2017 (Vaswani et. al. *Attention is all you need*)

   - Contextually embedded text. BERT, CAMEMBERT ...
   - Each vocabulary **word** will be mapped to many **different vectors in** $\mathbb{R}^L$;
     Each vocab. word embedding depends on its local context.
   - Different modeling strategy needed: MoE.
     Softmax mixture ensembles = the key building block. Coming up !

Quick account of existing results on softmax mixture (ensemble) models

- Theoretical understanding of softmax mixtures ($n \geq 1$) lacking.

  Classical results (Lindsay 95) on "quadratic variance exponential families" do not extend, especially for $L > 1$.

- Computationally efficient estimation of **mixture parameters**, by **any method** and with theory, generally **lacking**, for **any mixture** model, for $L > 1$.

  Scattered results, devoted exclusively to Gaussian mixtures (GMM).

- Most basic algorithm for mixture models: EM.

  Theoretical and practical understanding of EM, even in GMM: still emerging, for $K > 2$.

  No existing results for EM in softmax mixtures.

- Joint log-likelihood is **not** concave in parameters:

$$\ell(\boldsymbol{\omega}) \;\propto\; \sum_{i=1}^{n}\sum_{j=1}^{p}\widehat{\pi}_j^{(i)}\log\left\{\sum_{k=1}^{K}\alpha_k^{(i)}\frac{\exp(X_j^\top\boldsymbol{\theta}_k)}{\sum_{\ell=1}^{p}\exp(X_\ell^\top\boldsymbol{\theta}_k)}\right\}$$

- The surrogate objective function **is** concave in $\boldsymbol{\omega}$:

$$\widehat{Q}(\boldsymbol{\omega}\mid\boldsymbol{\omega}') := \sum_{i=1}^{n}\sum_{j=1}^{p}\widehat{\pi}_j^{(i)}\sum_{k=1}^{K}h_{\boldsymbol{\omega}'}(k,i,X_j)\left\{\log(\alpha_k^{(i)}) + X_j^\top\boldsymbol{\theta}_k - \log\left(\sum_{\ell=1}^{p}\exp(X_\ell^\top\boldsymbol{\theta}_k)\right)\right\}$$

where

$$h_{\boldsymbol{\omega}'}(k,i,x) = \frac{\alpha_k^{(i)'}A_{\boldsymbol{\theta}_k'}(x)}{\sum_{a=1}^{K}\alpha_a^{(i)'}A_{\boldsymbol{\theta}_a'}(x)}, \qquad \forall\, k\in[K],\; i\in[n],\; x\in\{X_1,\ldots,X_p\},$$

for $A_{\boldsymbol{\theta},j}(x) = \frac{\exp(X_j^\top\boldsymbol{\theta})}{\sum_{\ell=1}^{p}\exp(X_\ell^\top\boldsymbol{\theta})}$, for any $\theta\in\mathbb{R}^L$ and all $j\in[p]$.

Theoretical guarantees for ensemble EM with warm start

1. For warm start within a $r = O(L + \log p)^{-1/2}$ neighborhood of the target, the EM estimates satisfy

$$\max_k \|\widehat{\theta}_k - \theta_k\|_2 \lesssim \sqrt{\frac{L \log(nN)}{nN}}$$

w.h.p., after $O(\log(Nn/L))$ iterations .

One sample ($n = 1$) theory: Bing, Bunea, Niles-Weed, Wegkamp (2025+)

Ensemble ($n \geq 1$) theory: stay tuned, 2025+ !

2. Cross-entropy (re-fitted) document mixture weight estimates $\widetilde{\alpha}^{(i)}$ are sparse !

Standard topic models, sparsity: Bing, Bunea, Strimas-Mackey, Wegkamp (2022);
Standard topic models, inference for sparse weights: Bing, Bunea, Niles-Weed, (2024).

Softmax topic models: stay tuned, 2025+ !

Main technical conditions for EM algorithmic-statistical success

1. Well behaved Fisher information matrix of **each** softmax mixture component.

2. Atom separation: $\min_{k \neq k'} \|\theta_k - \theta_{k'}\|_2^2 \geq C \log K$

   Weakest known separation is $\min\{K, L\}$, for EM convergence to a **global maximum** in GMM, $K \geq 3$.
   Yan et. al (2017); Zhao et al. (2020).

3. Initialization: there exist $\theta_1^0, \ldots, \theta_K^0$ s.t. $\max_{k \in [K]} \|\theta_k - \theta_k^0\|_2 \leq r$.

   Explicit characterization of warm start radius: $r = O((L + \log p)^{-1/2})$.

One sample ($n = 1$) theory: Bing, Bunea, Niles-Weed, Wegkamp (2024, 2025+)

Ensemble ($n \geq 1$): stay tuned, 2025+ !

Warm start construction for EM in softmax mixtures with random design

1. Classic with a twist ($n \geq 1$): New MoM parameter estimates.
   Bing, B., Niles-Weed and Wegkamp; 2025+

   - Estimate **many latent moments** of mixing measures $\rho := \sum_{k=1}^{K} \alpha_k \delta_{\theta_k}$ via softmax mixture-tailored functionals of the data (New !).

     Use Lindsay's results (89, 93) to find MoM latent **parameter** estimates.

   - Parametric rates if atoms θ_k differ in their first coordinate. Warm start !

   - Behavior of MoM parameter estimates deteriorates rapidly when K increases, even for moderate L.

2. Commonly used ($n \geq 1$): multiple random initializations. B, B, N-W, W; 2025+

   - Estimate only second-order latent moments of the mixing measure.

   - **Estimate K-dim sub-space** in $\mathbb{R}^L$ spanned by $\theta_1, \ldots, \theta_K$, using latent **moment** estimates.

     Low dimensional Initialization; success in only $\exp(K)$ draws.
     Recommended !

Specific warm-start construction for EM in ensembles with pure documents

1. Specific to softmax mixtures ensembles ($n > 1$): pure document assumption. Computational savings relative to random start.

2. *The pure document assumption.*

   For each domain $k \in [K]$, there exists a document $i \in [n]$ covering only that domain, $\|\alpha^{(i)}\|_0 = 1$: $\alpha_k^{(i)} = 1$, and $\alpha_l^{(i)} = 0$, $\forall\, l \neq k$.

3. Doc i pure = Doc distribution is a mixture with *one component.*

   Akin to having observations from *each individual* mixture component.

   - Estimate the pure document index set. Find its K clusters.
   - Warm start for θ_k: average of *individual* softmax (MLE) estimate.
   - Warm start for weight $\alpha_k^{(i)}$: Preliminary standard (MLE) estimates.

The "pure document" assumption is testable

- An if and only if characterization of pure document *i* based on second-order latent moments Uspensky (1937); Bing, Bunea, Wegkamp (2025+, to come !)

$$\|\alpha^{(i)}\|_0 = 1 \iff \mathrm{tr}(\Gamma^{(i)}) > 0;$$

$$\Gamma^{(i)} := \Gamma_1^{(i)} + \Gamma_2^{(i)} =: \sum_{k=1}^{K} \alpha_k^{(i)} \theta_k \theta_k^\top - \left(\sum_{k=1}^{K} \alpha_k^{(i)} \theta_k\right) \left(\sum_{k=1}^{K} \alpha_k^{(i)} \theta_k\right)^\top$$

- Estimable first and second-order latent moments tailored to softmax mixtures.

$$\widehat{\Gamma}_1^{(i)} = \frac{1}{N} \sum_{\ell=1}^{N} \widehat{\Sigma}^{-1} Y_\ell^{(i)} (Y_\ell^{(i)})^\top \widehat{\Sigma}^{-1} - \widehat{\Sigma}^{-1}; \quad \widehat{\Sigma} = p^{-1} X^\top X.$$

Document *i*: $Y_1^{(i)}, \ldots, Y_\ell^{(i)}, \ldots, Y_N^{(i)}$ in $\mathbb{R}^L$. The rows of X are $N(0, \Sigma)$.
Bing, Bunea, Niles-Weed, Wegkamp (2024, 2025 +)

- The pure document assumption is testable !

Stay tuned ! 2025+

Instances of latent moment estimates in softmax mixtures

- Latent moments of the first coordinate = moments of $Z^{(i)} \sim \sum_{k=1} \alpha_k^{(i)} \delta_{\theta_{1k}}$.

$m_r = \sum_{k=1}^K \alpha_k^{(i)} \theta_{1k}^r,$ for each $0 \leq r \leq 2K-1$.

- Latent moment estimator, for each $0 \leq r \leq 2K-1$

$$\widehat{m}_r = \frac{1}{N} \sum_{\ell=1}^{N} h_r(Y_\ell^{(i)}),$$

for $Y_1^{(i)}, \ldots, Y_\ell^{(i)}, \ldots Y_N^{(i)}$ i.i.d. $\pi^{(i)}(\cdot \mid x)$, the softmax mixture of doc. i .

If support points $x_1, \ldots, x_p$ i.i.d. $N(0, \Sigma)$, then

$$h_r(x) := h_r(x; \Sigma) = \|\Sigma^{-1/2} \boldsymbol{e}_1\|_2^r H_r\left(x^\top \Sigma^{-1} \boldsymbol{e}_1 / \|\Sigma^{-1/2} \boldsymbol{e}_1\|_2\right).$$

$$H_r(x) = r! \sum_{b=0}^{\lfloor r/2 \rfloor} \frac{(-1)^b}{b!(r-2b)!2^b} x^{r-2b}, \qquad \forall\, x \in \mathbb{R}.$$

Identifiability of ensembles of softmax mixtures vs single mixture

1. Single softmax mixtures ($n = 1$) are identifiable, when p is large, if:

   - $\min_{k \neq k'} \|\theta_k - \theta_{k'}\|_2^2 \geq C \log K$;
   - $\min_k \alpha_k > 0$;
   - $\min_{k \neq k'} |\theta_{1k} - \theta_{1k'}| \geq \Delta$.

   Constructive proof, via a population-level EM algorithm with population MoM warm start.
   Bing, Bunea, Niles-Weed, Wegkamp, 2025+

2. Ensembles of softmax mixtures ($n > 1$) are identifiable, when p is large, if:

   - $\min_{k \neq k'} \|\theta_k - \theta_{k'}\|_2 > 0$;
   - Ensemble has pure documents.

     To come, 2025+, stay tuned !

3. The power of the ensemble:
   Separation conditions ($n = 1$) replaced by a testable condition, when $n > 1$.
   Mixtures can be sparse.

Into the LLM era

Document $i \rightarrow N$ tokens $\rightarrow$ Contextually embedded into $v_1^{(i)}, \ldots, v_N^{(i)} \in \mathbb{R}^L$.

Embedding dimension L very large: 384 or 768, or **much larger** .

1. Vocabulary = a collection of agreed upon tokens.
2. Collection of documents = collection of samples in $\mathbb{R}^L$.
3. Major gain: each document contains N **contextually embedded** vectors.
   BERT, ROBERTA, CAMEMBERT, ...

A simple **interpretable** MoE for embedded corpora

Document i representation: $v_1^{(i)}, \ldots, v_N^{(i)} \in \mathbb{R}^L$.

Summarize document i probability distribution: continuous "topic model".

$$\sum_{k=1}^{K} \alpha_k^{(i)} \left(\sum_{j=1}^{q} \frac{\exp \boldsymbol{\theta}_k^\top \mu_j}{\sum_{j=1}^{q} \exp \boldsymbol{\theta}_k^\top \mu_j} \mathrm{N}(\mu_j, \sigma_j^2 \mathrm{I}) \right) = \sum_{j=1}^{q} \left(\sum_{k=1}^{K} \alpha_k^{(i)} \frac{\exp \boldsymbol{\theta}_k^\top \mu_j}{\sum_{j=1}^{q} \exp \boldsymbol{\theta}_k^\top \mu_j} \right) \mathrm{N}(\mu_j, \sigma_j^2 \mathrm{I})$$

$$=: \sum_{j=1}^{q} \pi_j^{(i)} \mathrm{N}(\mu_j, \sigma_j^2 \mathrm{I})$$

- Take q large: universal approximation results for infinite Gaussian mixtures.
- Parametrize $\pi_j^{(i)}$ to reflect the relative alignment of μ_j's with direction θ_k.
- With q large:
  - θ_k re-groups the many Gaussian means.
  - Each group around θ_k has weight $\alpha_k^{(i)}$.

Embedded corpus = nN contextually embedded vectors in $\mathbb{R}^L$.

1. Fit a Gaussian mixture model with a very **large** number of components q. Set $x_j =: \widehat{\mu}_j$.

   - $x_1, \ldots, x_q \in \mathbb{R}^L$ yield a corpus-specific **new** dictionary.
   - Tokens in cluster j name the **new word** $x_j \in \mathbb{R}^L$.

2. Conditionally on $x_1, \ldots, x_q$, use EM to optimize

$$\sum_{i=1}^{n} \sum_{j=1}^{p} \widehat{\pi}_j^{(i)} \log \left\{ \sum_{k=1}^{K} \alpha_k^{(i)} \frac{\exp(x_j^\top \boldsymbol{\theta}_k)}{\sum_{\ell=1}^{p} \exp(x_\ell^\top \boldsymbol{\theta}_k)} \right\}$$

   with Gaussian mixture proportions $\widehat{\pi}_j^{(i)}$, $j \in [p]$, $i \in [n]$.

   Estimation of this MoE is estimation in softmax mixture ensembles !

3. Theory carries over, conditionally on $x_1, \ldots, x_q$.
   (Bing, Bunea, Wegkamp, 2025 +. To Come !)

Curated Dataset:

Content Type	Count
Movie Reviews	58
Game Reviews	58
Total	**116**

- Total number of tokens $nN = 18,988$.
- New dictionary size $q = 5000$.
- Embedding dimension $L = 768$.

Instances of reviews

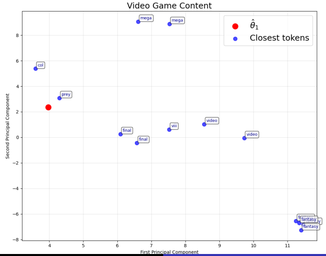
- Movies: Final Fantasy film, Resident Evil movie, Street Fighter
- Games: Final Fantasy VIII, Metal Gear Solid, Tomb Raider III

Create new, contextually meaningful, embedded dictionary.

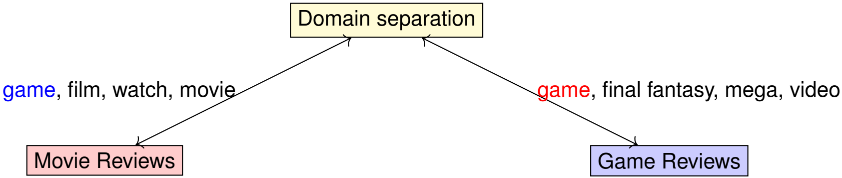
Instances of new "words".

- game$_1, \ldots,$ game$_M \in \mathbb{R}^L$: cluster centers of game as in game reviews.
- game$_1, \ldots,$ game$_D \in \mathbb{R}^L$: cluster centers of game as in movie reviews.

EM on softmax mixture ensembles on new dictionary.



Domain recovery via MoE on contextually embedded text



$\widehat{\theta}_1$: Video game reviews.

```
['final', 'final', 'fantasy', 'fantasy', 'fantasy', 'mega', 'viii',
'fantasy', 'video', 'video', 'prey', 'col', 'mega']
```

$\widehat{\theta}_2$: Movie (based on video-games) reviews.

```
[ 'see watch seeing', 'rendered backgrounds roles', 'movie', 'film', 'story
wrote stories', 'movie film', 'game' ]
```

1. **Corpus summarization via topic models in the LLM era: still valid !**

   - Standard (non-parametrized): mixing measure $\sum_{k=1}^{K} \bar{\alpha}_k A_k$, with $A_k \in \Delta_p$. Broad strokes summary.
   - Softmax-parametrized: mixing measure $\sum_{k=1}^{K} \bar{\alpha}_k \delta_{\theta_k}$, with $\theta_k \in \mathbb{R}^L$. Broad strokes, dimension stabilized summary. No local context.
   - Automatic document clustering by domain: mixture weight sparsity.

2. **Corpus summarization via softmax-mixture ensembles within MoE's**:

   - **Contextually interpretable latent domain** parameters θ_k.
   - **Automatic domain cluster building**: mixture weight sparsity.
   - Contextually-interpretable mixing measure-type summaries

$$\sum_{k=1}^{K} \bar{\alpha}_k \delta_{\theta_k}.$$

   Key to developing new metrics for comparing LLM and Human corpora.

   Bunea, Manole, Wegkamp, Thickstun et al. ; 2025+ To come !

Bibliography

- *Learning large softmax mixtures with warm start EM*, 2024/2025+
  ArXiv; Bing, X.; Bunea, F., Niles-Weed. J., Wegkamp, M.

- *Estimation and Inference for the Wasserstein distance between mixing measures in topic models* ; 2024
  Bernoulli ; Bing, X.; Bunea, F., Niles-Weed. J.

- *Likelihood estimation of sparse topic distributions in topic models and its applications to Wasserstein document distance calculations*; 2022,
  Annals of Stats. Bing, X., B., F., Strimas-Mackey, S., Wegkamp, M.

- *A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics;* 2020 (a).
  Bernoulli, Bing, X., B., F., Wegkamp, M.

- *Optimal estimation of sparse topic models*; 2020 (b).
  JMLR, Bing, X., B., F., Wegkamp, M.

To Come !

- *Ensembles of softmax mixtures with applications to LLM output summarization*, 2025+
  Bing, X.; Bunea, F., Betache, N, Wegkamp, M; Wu, W.

- *New metrics for LLM evaluation*, 2025+
  Bunea, F., Manole, T., Wegkamp, M., Thickstun, J., et al.

Thank you !

Minimax adaptive estimation of mixture components $A_{\cdot k} \in \Delta_p$

In a topic model with anchor words we can construct estimators such that, under assumptions,

$$\max_{1 \le k \le K} \|\widehat{A}_{\cdot k} - A_{\cdot k}\|_1 \lesssim \sqrt{\frac{pK}{nN}},$$

w.h.p., for p and K allowed to grow with n.

B., Bing, Wegkamp (2018).

Theory for mixture weight estimators in topic models

- Enough to provide theory for each document.
- Assume that the word count vector in one document of length N is

$$Y \sim \mathrm{Multinomial}_p(N; \pi = \sum_{k=1}^{K} \alpha_k A_{\cdot k}).$$

- Given minimax optimal estimators $\widehat{A}_{\cdot k}$, construct

$$\widehat{\alpha}_{\mathrm{mle}} := \mathrm{argmax}_{\alpha \in \Delta_K} \sum_{j=1}^{p} Y_j \log\left(\widehat{A}_{j\cdot}^{\top} \alpha\right),$$

the MLE (cross-entropy) estimator of the mixture weight vector α.

It is automatically sparse if α is sparse !

Intuition: the MLE of a sparse mixture weight vector is sparse w.p. 1

$$(Y_1, Y_2, Y_3) \sim \text{Multinomial}_3(N, \pi),$$

$$\pi = \sum_{k=1}^{2} \alpha_k A_k, \quad \text{with} \quad A_1 = (a_1, a_2, 0)^\top, \quad A_2 = (0, 0, 1)^\top, \quad a_1 + a_2 = 1, \quad a_1, a_2 \geq 0.$$

$$Y_1 + Y_2 \sim \text{Bin}(N, \alpha_1), \quad Y_3 = \text{Bin}(N, \alpha_2).$$

$$\widehat{\alpha}_{\text{mle},1} = \frac{Y_1 + Y_2}{N}, \quad \widehat{\alpha}_{\text{mle},2} = \frac{Y_3}{N}.$$

If $\alpha_k = 0$, for some $k \in \{1, 2\}$, then $\widehat{\alpha}_{\text{mle},k} = 0$, w.p. 1.

If data is generated from a sparse discrete mixture, $\widehat{\alpha}_{\text{mle}}$ can have zero entries.

Sparsity and sparse rate adaptation of the MLE

If the mixture components $\widehat{A}_{\cdot k};\ k \in [K]$ satisfy an incoherence-type condition then:

(1) The MLE $\widehat{\alpha}_{\mathrm{mle}}$ has at least as many zeroes as the true mixture weight vector α :

$$\mathrm{supp}(\widehat{\alpha}_{\mathrm{mle}}) \subseteq \mathrm{supp}(\alpha) =: s, \ \text{w.h.p.}$$

(2) The MLE $\widehat{\alpha}_{\mathrm{mle}}$ is minimax-rate adaptive

$$\|\widehat{\alpha}_{\mathrm{mle}} - \alpha\|_1 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{s \log K}{N}} + \sqrt{\frac{pK}{nN}}\right).$$

A one step update of the MLE $\widehat{\alpha}_{\mathrm{mle}} \in \Delta_K$ for efficient estimation of (sparse) $\alpha \in \Delta_K$:

$$\widetilde{\alpha} = \widehat{\alpha}_{\mathrm{mle}} + \widehat{V}^{+} \Psi(\widehat{\alpha}_{\mathrm{mle}}).$$

ASN estimators of sparse mixture weights

Under the General Regime,

$$\lim_{n,N\to\infty} \sqrt{N}\, \Sigma_{\alpha}^{+\frac{1}{2}} \left(\widetilde{\alpha} - \alpha\right) \xrightarrow{d} \mathcal{N}_K \left(0, \begin{bmatrix} \boldsymbol{I}_{K-1} & 0 \\ 0 & 0 \end{bmatrix}\right).$$

$\Sigma_{\alpha} \geq 0$ reduces to the asymptotically efficient covariance matrix under the Classical Regime.

$$\widehat{V} := \sum_{j\in\widehat{J}} \frac{1}{\widehat{A}_{j\cdot}^{\top} \widehat{\alpha}_{\mathrm{mle}}} \widehat{A}_{j\cdot} \widehat{A}_{j\cdot}^{\top} \qquad \Psi(\widehat{\alpha}_{\mathrm{mle}}) := \sum_{j\in\widehat{J}} \frac{\widehat{\pi}_j - \widehat{A}_{j\cdot}^{\top} \widehat{\alpha}_{\mathrm{mle}}}{\widehat{A}_{j\cdot}^{\top} \widehat{\alpha}_{\mathrm{mle}}} \widehat{A}_{j\cdot} \qquad \widehat{J} =: \mathrm{supp}(\widehat{\pi}).$$

Classical Regime	**General Regime**
(a) A is known	(a') A is estimated from n samples of size N each
(b) p is independent of N	(b') p can grow with N (and n)
(c) $0 < \alpha_k < 1$, for all $k \in [K]$	(c') $\alpha \in \Delta_K$ can be sparse.
(d) $\pi_j > 0$, for all $j \in [p]$	(d') $\pi \in \Delta_p$ can be sparse.

EM initialization in softmax mixture ensembles with random design

1. Classic: MoM parameter estimates. Bing, B., Niles-Weed and Wegkamp; 2024+
   - Estimate many latent moments of mixing measures $\rho := \sum_{k=1}^{K} \alpha_k \delta_{\theta_k}$ via softmax mixture-tailored functionals of the data.

     Use Lindsay's Lemma to find MoM **parameter estimates.**
   - Parametric rates if atoms θ_k differ in their first coordinate. Warm start !
   - Behavior deteriorates rapidly when K increases, even for moderate L.
2. Commonly used: multiple random initializations. B, B, N-W, W; 2025+
   - Estimate only second-order latent moments of the mixing measure.
   - **Estimate K-dim sub-space** in $\mathbb{R}^L$ spanned by $\theta_1, \ldots, \theta_K$.

     Initialize in this sub-space; success in only $\exp(K)$ draws.
3. Specific to softmax mixtures ensembles: anchor document assumption.

$m_r = \sum_{k=1}^K \alpha_k \theta_{1k}^r, \; 0 \le r \le 2K-1$ and $m_{r1;i} := \sum_{k=1}^K \alpha_k (\theta_{1k})^r \theta_{ik}, \; 0 \le r \le K-1$ and $2 \le i \le L$.

Lemma (Lindsay '89, Lindsay and Basak '93)

If $\min_{k \neq k'} |\theta_{1k} - \theta_{1k'}| > 0$ *, then the moments and mixed moments of the* **mixing measure** $\rho := \sum_{k=1}^K \alpha_k \delta_{\theta_k}$ *uniquely determine* $\boldsymbol{\alpha} \in \Delta_K$ *and* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K \in \mathbb{R}^L$*, up to some permutation, as shown below.*

1. *The first coordinates* $\theta_{11}, \theta_{12}, \ldots, \theta_{1K}$ *are the unique* K *roots of the degree* K *polynomial* $P(x)$*, in one variable, given by*

$$P(x) := \det \begin{pmatrix} 1 & m_1 & \cdots & m_K \\ m_1 & m_2 & \cdots & m_{K+1} \\ \vdots & \vdots & & \vdots \\ m_{K-1} & m_K & \cdots & m_{2K-1} \\ 1 & x & \cdots & x^K \end{pmatrix}.$$

2. *For each* $k \in [K]$*, the remaining* $L-1$ *coordinates* $\theta_{ik}, i \in \{2, \ldots, L\}$ *are uniquely given by*

$$\theta_{ik} = \begin{pmatrix} m_{01;i} \\ \vdots \\ m_{(K-1)1;i} \end{pmatrix}^\top \begin{pmatrix} 1 & m_1 & \cdots & m_{K-1} \\ m_1 & m_2 & \cdots & m_K \\ \vdots & \vdots & & \vdots \\ m_{K-1} & m_K & \cdots & m_{2K-2} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ \theta_{1k} \\ \vdots \\ \theta_{1k}^{K-1} \end{pmatrix}.$$

3. *The mixture weight vector* $\boldsymbol{\alpha}$ *is uniquely given by*

$$\boldsymbol{\alpha} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \theta_{11} & \theta_{12} & \cdots & \theta_{1K} \\ \vdots & \vdots & & \vdots \\ \theta_{11}^{K-1} & \theta_{12}^{K-1} & \cdots & \theta_{1K}^{K-1} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ m_1 \\ \vdots \\ m_{K-1} \end{pmatrix}.$$

Topic model = a non-parametrized ensemble of discrete mixtures

Data: Independent $Y^{(i)} =: (Y_1^{(i)}, \ldots, Y_p^{(i)}) \sim \text{Multinomial}_p(N, \pi^{(i)}), \qquad i \in [n].$

$$\mathbb{P}(\text{Word } j \text{ in doc } i) \underbrace{=}_{\textit{Topic model assumption}} \sum_{k=1}^{K} \mathbb{P}(\text{ Topic } k \text{ in doc } i\ )\, \mathbb{P}(\text{Word } j \mid \text{Topic } k)$$

$$\pi_j^{(i)} = \sum_{k=1}^{K} \alpha_k^{(i)} A_{jk}, \quad j \in [p], \quad i \in [n]$$

$$\pi^{(i)} = \sum_{k=1}^{K} \alpha_k^{(i)} A_{\cdot k}, \quad i \in [n]$$

$$\underbrace{\Pi}_{p \times n} = \underbrace{A}_{p \times K} \underbrace{T}_{K \times n}.$$

- $n =$ number of docs in the corpus; $p =$ dictionary size.
- $\pi^{(i)} \in \Delta_p$ true word distribution in document i.
  - $\alpha^{(i)} \in \Delta_K$: Document specific mixture weights vector.
  - $A_{\cdot k} \in \Delta_p$ mixture component vector, common to the corpus.