Attention with Trained Embeddings Provably Selects Important Tokens

Author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2025

Abstract

Token embeddings play a crucial role in language modeling but, despite this practical relevance, their theoretical understanding remains limited. Our paper addresses the gap by characterizing the structure of embeddings obtained via gradient descent. Specifically, we consider a one-layer softmax attention model with a linear head for binary classification, i.e., $\texttt{Softmax}(p^\top \boldsymbol{E}_X^\top)\boldsymbol{E}_X v$, where $\boldsymbol{E}_X = [E_{x_1}, \dots, E_{x_T}]^\top$ contains the embeddings of the input sequence, p is the embedding of the $\langle \text{cls} \rangle$ token and v the output vector. First, we show that, already after a single step of gradient training with the logistic loss, the embeddings $\boldsymbol{E}_X$ capture the importance of tokens in the dataset by aligning with the output vector v proportionally to the frequency with which the corresponding tokens appear in the dataset. Then, after training p via gradient flow until convergence, the softmax selects the important tokens in the sentence (i.e., those that are predictive of the label), and the resulting $\langle \text{cls} \rangle$ embedding maximizes the margin for such a selection. Experiments on real-world datasets (IMDB, Yelp) exhibit a phenomenology close to that unveiled by our theory.

1. Introduction

The introduction of the attention mechanism [5, 41] marked a paradigm shift in the design of frontier machine learning models, leading to significant advances such as ChatGPT [2], Claude [3], AlphaFold [18], CLIP [30] and Dall-E [31]. This success prompted a surge of interest in understanding the structure and function of attention layers, with their optimization dynamics and inductive biases being object of extensive theoretical research [1, 7, 9, 24, 37, 42]. Embeddings are a crucial component of the attention mechanism [45], especially for downstream adaptation [13, 16, 19] with some works [20, 45] specifically highlighting their importance. However, despite the importance of learning embeddings, the existing analyses of transformer-like architectures either ignore the properties of embeddings by resorting to orthogonal structures [44], or omit embeddings completely by considering unprocessed inputs [39]. Our paper fills this gap by studying directly the embedding training dynamics. We aim to provide theoretical insight to the following questions:

What is the structure learnt by the embeddings during gradient descent training?
How is this structure related to the statistical properties of the data?

In Figure 1, we investigate these questions by analyzing the embeddings of a two-layer transformer trained on a sentiment analysis task on IMDB. The plots reveal a remarkable simplicity in the structure of the learned embeddings, which capture the frequency of appearance of tokens in the dataset. Specifically, the predictive mechanism (overlap with the regression coefficient v) favors the tokens which appear more frequently in the corresponding positive/negative context. A similar pattern emerges at the selection stage of the attention mechanism (overlap with the $\langle \text{cls} \rangle$ embedding

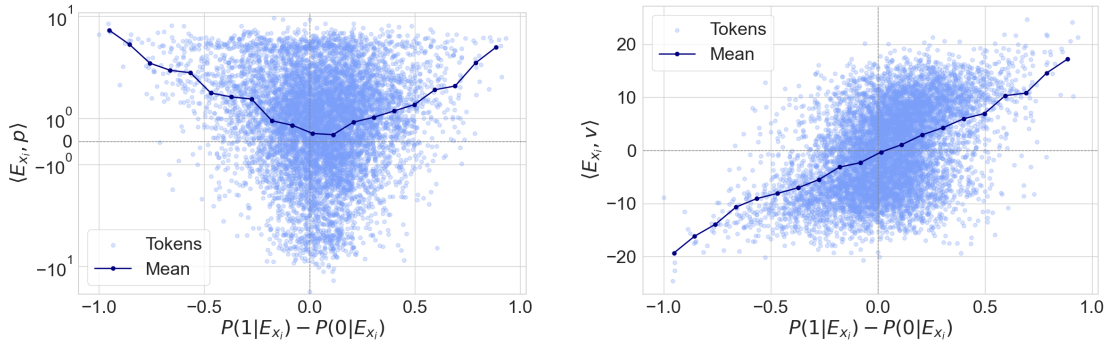

Figure 1: Dot-product of token embeddings with $\langle \text{cls} \rangle$ embedding p (left) and regression coefficients v (right), as a function of token-wise difference in posterior probabilities, for a two-layer attention model trained on the IMBD dataset (see (32) in Appendix E for details).

p), i.e., more frequent tokens have a higher attention score. Analogous results for Yelp data are reported in Figure 4 deferred to Appendix E.

For the theoretical study of this emergent structure, we focus on a one-layer softmax attention model. Namely, for an input sequence $X = [x_1, \cdots, x_T]$, the output of the model is given by

$$f(X; p, \boldsymbol{E}) = \texttt{Softmax}(p^\top \boldsymbol{E}_X^\top) \boldsymbol{E}_X v, \tag{1}$$

where $\boldsymbol{E}_X = [E_{x_1}, \ldots, E_{x_T}]^\top$ contains the embeddings of the input X, p is the embedding of the $\langle \text{cls} \rangle$ token and v is the final regression vector. Our main results are summarized below:

- We show that, already after a single step of gradient training with the standard logistic loss, the embeddings $\boldsymbol{E}_X$ capture the importance of tokens in the dataset by aligning with the output vector v proportionally to the corresponding empirical frequencies (Lemma 2).
- In case, each sequence contains a single important token, the $\langle \text{cls} \rangle$ embedding obtained from gradient flow must select all important tokens. We characterize all the possible directions that the $\langle \text{cls} \rangle$ embedding may converge to, which are the max-margin solutions associated to feasible token selections (Theorem 5). While in general the $\langle \text{cls} \rangle$ embedding may select irrelevant tokens, we identify sufficient conditions leading to the selection only of important tokens.

Related work. The implicit bias literature has been instrumental in understanding the behavior of neural networks or overparameterized models optimized by gradient methods [4, 8, 26]. A key phenomenon is that gradient descent on separable data with logistic loss directionally converges to the max-margin separator [15, 35]. More recently, a series of works [17, 21, 22, 32, 34, 36, 37, 40] has established an equivalence between the optimization geometry of self-attention and a hard-margin SVM problem selecting a subset of tokens via linear constraints on the outer-products of token pairs. Compared to these works that mostly focus on the training of single-layer attention weights, we point out two differences. First, we study the role of embeddings and their joint training with the $\langle \text{cls} \rangle$ token. Second, under our data model, we establish benign properties of the solution reached at convergence (which may not hold for arbitrary datasets [37]). Additional related work on the theory of attention layers is discussed in Appendix A.

2. Preliminaries

Data and model. We focus on binary text classification problems. We consider a vocabulary set $\mathcal{S}$ with size $|\mathcal{S}|$ and a $\langle \text{cls} \rangle$ token for classification. Let $(X_i, y_i)_{i=1}^n$ be the dataset containing n context

sequences, where $y_i \in \{-1, 1\}$ and each context sequence $X \in \mathcal{X}_n := \{X_1, \dots, X_n\}$ contains T tokens, i.e., $X = [x_1, \dots, x_T]$ with $x_i \in \mathcal{S}$. W.l.o.g., we let $\mathcal{S}$ be the set of tokens that appears in $\mathcal{X}_n$, as the embeddings of the remaining tokens are not trained and are not relevant for the problem at hand. We consider a one-layer softmax attention model with a linear head for classification. First, we append a $\langle \text{cls} \rangle$ token at the end of the sequence X, and then we embed each token into a vector of dimension d. Namely, after the embedding layer, we have $\boldsymbol{E}_X = [E_{x_1}, \dots, E_{x_T}]^\top \in \mathbb{R}^{T \times d}$, where $E_s \in \mathbb{R}^d$ denotes the embedding of the token s. We let $\boldsymbol{E} \in \mathbb{R}^{|\mathcal{S}| \times d}$ be the embedding matrix of all context tokens and $p \in \mathbb{R}^d$ the embedding of the $\langle \text{cls} \rangle$ token.

We focus on the architecture (1) where, given $a \in \mathbb{R}^T$, $[\texttt{Softmax}(a)]_i := \frac{\exp(a_i)}{\sum_{j=1}^T \exp(a_j)}$ for $i \in \{1, \dots, T\}$. The same model is also studied in [32, 37]. In practice, it is common to include the W_{KQ} matrix and consider a model with output $f(X; p, W_{KQ}, \boldsymbol{E}) = \texttt{Softmax}(p^\top W_{KQ} \boldsymbol{E}_X^\top) \boldsymbol{E}_X v$. Since $p^\top W_{KQ}$ plays the same role as p and one can easily reconstruct W_{KQ} from p in each gradient update as discussed in [37], we use the model in (1) for simplicity. The output vector v is fixed and all the embedding vectors $p, \boldsymbol{E}$ are trained to minimize with the standard logistic loss:

$$\mathcal{L}(\boldsymbol{E}, p) = \frac{1}{n} \sum_{k=1}^n \log(1 + \exp(-y_k f(X_k; \boldsymbol{E}, p))) = \widehat{\mathbb{E}} \left[\log(1 + \exp(-y f(X; \boldsymbol{E}, p)))\right], \quad (2)$$

where the notation $\widehat{\mathbb{E}}$ is a shorthand for the average over the dataset $\mathcal{D} = \{(X_k, y_k)\}_{k=1}^n$.

Empirical statistics of each token in the dataset. The goal of the paper is to characterize the structure of the embeddings $\boldsymbol{E}, p$ obtained by optimizing the objective (2) via gradient descent, and we show that such structure is related to the empirical statistics of the tokens in the dataset. Specifically, after training, the softmax attention learns to select tokens that are more correlated to the labels based on the dataset. To quantify the correlation between a token s and the label y, we define the *average signed frequency* of a token as:

$$\alpha_s := \frac{1}{nT} \sum_{(X,y) \in \mathcal{D}} \left(y \sum_{i=1}^T \mathbb{1}_{x_i = s} \right) = \frac{1}{T} \cdot \widehat{\mathbb{E}} \left[y \sum_{i=1}^T \mathbb{1}_{x_i = s} \right]. \quad (3)$$

In words, α_s is obtained by taking the number of occurrences of s in sequences with a positive label, subtracting the number of occurrences of s in sequences with a negative label, and finally dividing by the total number of tokens nT. As such, it provides an average of the signed frequency of s, where the sign comes from the label of the sequences in which the token appears.

Definition 1 (Positive, negative and irrelevant tokens) *We say that a token s is* (i) *positive if $\alpha_s > 0$,* (ii) *negative if $\alpha_s < 0$, and* (iii) *irrelevant if $\alpha_s = 0$. Moreover, a token s is completely positive (*completely negative*) if it appears only in sequences with label 1 (-1).*

3. Main results

One step of gradient descent learns the importance of the tokens. We initialize v with any unit-norm vector and $E_s^0, p^0 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, d^{-1} I)$ for all $s \in \mathcal{S}$. Then, we perform one step of gradient descent with step size η_0 on all trainable embeddings:

$$p^1 = p^0 - \eta_0 \nabla_p \mathcal{L}(\boldsymbol{E}^0, p^0), \qquad E_s^1 = E_s^0 - \eta_0 \nabla_{E_s} \mathcal{L}(\boldsymbol{E}^0, p^0), \quad \text{for all } s \in \mathcal{S}. \quad (4)$$

Lemma 2 *For any $\delta > 0$, let $d \geq \text{polylog}(|\mathcal{S}|, \delta^{-1})$. Then, after the first step of gradient descent in (4) for any $s \in \mathcal{S}$, the following holds*

$$E_s^1 = E_s^0 + \eta_0/2 \cdot \alpha_s v + \mathit{err}_s, \qquad p^1 = p^0 + \mathit{err}_p, \quad (5)$$

where the error terms $\mathit{err}_s, \mathit{err}_p$ are bounded with probability at least $1 - \delta$ as

$$\max\{\max_{s \in \mathcal{S}} \|\mathit{err}_s\|_2, \|\mathit{err}_p\|_2\} \leq 11 \eta_0 d^{-\frac{1}{4}}. \quad (6)$$

Lemma 2 implies that after one step of training, the embedding vector E_s of each token s learns the empirical importance of the tokens by adding a vector in the direction of the output vector v with magnitude proportional to α_s. The proof is deferred to Appendix D.1.

The decomposition in (5) also implies that the overlap between the $\langle \text{cls} \rangle$ embedding vector p and E_s does not improve after the first step. Thus, we study the training dynamics of p, characterizing its implicit bias. Specifically, we fix the context embedding matrix to $\boldsymbol{E}^1$ (after the first gradient step) and train the $\langle \text{cls} \rangle$ embedding vector p with gradient flow initialized at p^1 (obtained after the first step):

$$\frac{\mathrm{d}}{\mathrm{d}t} p_t = -\nabla_p \mathcal{L}(\boldsymbol{E}^1, p_t). \tag{7}$$

We consider gradient flow for technical convenience, and all results in this section can be readily extended to gradient descent with small enough step size. We will refer to the embeddings in $\boldsymbol{E}^1$ as E_s and not E_s^1, omitting the superscript to favor readability.

Max-margin token selection. We denote the set of tokens in X selected by p as

$$\mathcal{S}_X(p) = \{\hat{s} : \hat{s} = \arg\max_{s \in X} p^\top E_s\}, \tag{8}$$

and we define $\overline{\mathcal{S}_X(p)} = X \setminus \mathcal{S}_X(p)$. Intuitively, given a sequence X, the selected tokens in X have the largest softmax weight (proportional to $\exp\big(p^\top E_{x_i}\big)$). Note that, for $p' \neq p$, we may have that $\mathcal{S}_X(p') = \mathcal{S}_X(p)$ for all X. Thus, we define the equivalence relation: $p \cong p' \iff \mathcal{S}_X(p) = \mathcal{S}_X(p')$, for all $X \in \mathcal{X}_n$. Intuitively, two vectors p, p' are equivalent under the above relation if they select the same tokens for all the sequences. Given a vector $p_\circ$, we denote by $\mathcal{P}_{p_\circ}$ its equivalence class, and we define the set of max-margin directions among all vectors in $\mathcal{P}_{p_\circ}$ as

$$\begin{aligned} \mathcal{P}_*(p_\circ) = \Bigg\{ \frac{\hat{p}}{\|\hat{p}\|_2} : \hat{p} = \arg\min_{p \in \mathcal{P}_{p_\circ}} \|p\|_2 \\ \text{s.t.} \quad p^\top (E_s - E_{s'}) \geq 1, \quad \forall s \in \mathcal{S}_X(\mathcal{P}_{p_\circ}),\ \forall s' \in \overline{\mathcal{S}_X(\mathcal{P}_{p_\circ})},\ \forall X \in \mathcal{X}_n \Bigg\}. \end{aligned} \tag{9}$$

We first show in the lemma below that the max-margin problem in (9) always has a unique solution, which means that $\mathcal{P}_*(p_\circ)$ is always a singleton. Thus, later on, we will use $\hat{p}(p_\circ)$ as the solution to (9), and $p_*(p_\circ) = \frac{\hat{p}(p_\circ)}{\|\hat{p}(p_\circ)\|_2}$. We drop the dependency on $p_\circ$ when there is no confusion.

Lemma 3 *For any $p_\circ \neq 0$, the max margin problem in (9) has a unique solution denoted as $\hat{p}$. Furthermore, for any $\delta > 0$, pick $d \geq \max\{\text{polylog}(|\mathcal{S}|, \delta^{-1}), \text{poly}(|\mathcal{S}|, \eta_0)\}$. Let N be the number of constraints in (9) and let the i-th constraint be $p^\top (E_{s_i} - E_{s_i'}) \geq 1$. Then, with probability at least $1 - \delta$, we have: $\hat{p} = \boldsymbol{M}^\dagger \mathbf{1}_N$, $\boldsymbol{M} = [E_{s_1} - E_{s_1'}, \ldots, E_{s_N} - E_{s_N'}]^\top \in \mathbb{R}^{N \times d}$, where $\boldsymbol{M}^\dagger$ denotes the pseudo-inverse of $\boldsymbol{M}$ and $\mathbf{1}_N$ a vector of N ones.*

Lemma 3 (proved in Appendix D.2) implies that with high probability, the solution $\hat{p}$ of the max-margin problem in (9) makes all the constraints tight.

Implicit bias of gradient flow. While Lemma 2 holds for any data, we need an extra assumption to analyze the gradient flow, due to the complex loss landscape caused by softmax attention.

Assumption 4 *Each sequence in $\mathcal{X}_n$ contains either a* single completely positive *token or a* single completely negative *token, and all remaining tokens are* irrelevant.

Assumption 4 implies that all sequences in the dataset contain precisely one relevant token, and the relevant token also aligns with the label. We remark that datasets containing only one relevant token have been also considered in prior work, see [36, Theorem 1] and [25]. We further denote by $\mathcal{S}_c$ the set containing all completely positive and all completely negative tokens.

Theorem 5 *Under Assumption 4, for any $\delta > 0$, let $\eta_0 \geq 4n^2T^2$, $d \geq \mathrm{poly}(|\mathcal{S}|, \log \delta^{-1}, \eta_0)$. Let p_t be the solution of the gradient flow (7). Then, with probability at least $1-\delta$, we have that $\|p_t\|_2 \to \infty$. Furthermore, assuming that $p_\infty := \lim_{t\to+\infty} p_t/\|p_t\|_2$ exists, the limiting direction p_∞ satisfies the following properties with probability at least $1-\delta$:*

1. *p_∞ selects all completely positive and completely negative tokens, i.e., $\mathcal{S}_c \subseteq \bigcup_X \mathcal{S}_X(p_\infty)$.*
2. *p_∞ is the max-margin direction for such a selection, i.e., $p_\infty = p_*(p_\infty)$.*

Theorem 5 (proved in Appendix D.3) shows that, if p_t converges in direction, it must converge to the max-margin direction that selects all the completely positive/negative token. We now highlight some differences w.r.t. [37]. Theorem 3 in [37] shows that gradient descent on p converges to a locally optimal max-margin solution when initialized close enough to such solution, and Theorem 4 in [37] shows that the regularization path can only converge to locally max-margin solutions. However, these results do not exclude the possibility of the gradient flow converging to directions that are *not* locally optimal and *not* the max-margin direction. In contrast, we characterize all possible directions the gradient flow converges to, showing that these are max-margin directions that select all completely positive/negative tokens. Furthermore, we do so without starting from an initialization that is close enough to such solution. This requires a different proof strategy as compared to [37].

Characterization of the max-margin solution. Theorem 5 still does not exclude the possibility that gradient flow also selects some irrelevant tokens. We address this point with the result below.

Lemma 6 *Suppose $\hat{p}$ selects all the tokens, i.e., $\mathcal{S}_X(\hat{p}) = \mathcal{S}$. Then, $p_\infty \neq \hat{p}$.*

Lemma 6 (proved in Appendix D.4) shows that the directional limit p_∞ (when it exists) cannot select all tokens and, as it selects all important ones, it must be biased towards them. As an application, consider the case where there is only one irrelevant token in the vocabulary. Then, the combination of Theorem 5 and Lemma 6 gives that only the completely positive/negative tokens are selected by gradient flow. Going beyond the case where there is a single irrelevant token, Lemma 19 in Appendix D.5 provides a sufficient condition for gradient flow to select only important tokens. This sufficient condition requires the max-margin direction that does not select irrelevant tokens to have a larger margin than any other max-margin solution associated to a different token selection. We expect this to be the case e.g. for datasets where all the completely positive/negative tokens have the same α_s.

Concluding remarks. We have studied how the embedding vectors trained via gradient methods capture the importance of different tokens in the dataset. Specifically, we have characterized *(i)* the context embedding E_s after one gradient step, and *(ii)* the implicit bias of the $\langle \mathrm{cls} \rangle$ embedding p after training with gradient flow until convergence. Experiments on synthetic and realistic datasets demonstrate the generality of our findings: Figure 1 considers a two-layer attention model trained on the IMBD datasets, and Figures 2, 3 in Appendix E show a similar behavior for the model (1) considered in the theoretical analysis trained on IMBD, Yelp and synthetic data.

The characterization we put forward is only in terms of the first-order statistics of the tokens (i.e., the frequencies with which they occur in the dataset), and it does not describe how the model learns the causal structure between tokens. In practice, both first-order statistics and causal structure are expected to be crucial for the model to "understand" a text. While our theory assumes a one-layer attention model, the numerical results of Figure 1 suggests that a similar qualitative picture holds more generally. This prompts us to conjecture that in deeper attention models with multiple heads, the earlier layers form induction heads [28] which learn the causal structure between tokens, and later layers perform classification based on the empirical statistics of the resulting k-tuples. We regard this investigation as an exciting future direction.

References

[1] Emmanuel Abbe, Samy Bengio, Aryo Lotfi, Colin Sandon, and Omid Saremi. How far can transformers reason? The globality barrier and inductive scratchpad. *NeurIPS*, 2024.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint*, 2023.

[3] Anthropic. Claude language model. https://www.anthropic.com/, 2025. Accessed via https://www.anthropic.com/.

[4] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *NeurIPS*, 2019.

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

[6] Simone Bombari and Marco Mondelli. Towards understanding the word sensitivity of attention layers: A study via random features. *International Conference on Machine Learning*, 2024.

[7] Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. *International Conference on Learning Representations*, 2024.

[8] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *Conference on Learning Theory*, 2020.

[9] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *NeurIPS*, 2023.

[10] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *arXiv preprint*, 2023.

[11] Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model. *arXiv preprint*, 2024.

[12] Halil Alperen Gozeten, M Emrullah Ildiz, Xuechen Zhang, Mahdi Soltanolkotabi, Marco Mondelli, and Samet Oymak. Test-time training provably improves transformers as in-context learners. *International Conference on Machine Learning*, 2025.

[13] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. *International Conference on Machine Learning*, 2019.

[14] M Emrullah Ildiz, Yixiao Huang, Yingcong Li, Ankit Singh Rawat, and Samet Oymak. From self-attention to markov models: Unveiling the dynamics of generative transformers. *International Conference on Machine Learning*, 2024.

[15] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. *Conference on Learning Theory*, 2019.

[16] Albert Q Jiang, Alicja Ziarko, Bartosz Piotrowski, Wenda Li, Mateja Jamnik, and Piotr Miłoś. Repurposing language models into embedding models: Finding the compute-optimal recipe. *NeurIPS*, 2024.

[17] Addison Kristanto Julistiono, Davoud Ataee Tarzanagh, and Navid Azizan. Optimizing attention with mirror descent: Generalized max-margin token selection. *Workshop Contribution at NeurIPS*, 2024.

[18] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 2021.

[19] Jannik Kossen, Mark Collier, Basil Mustafa, Xiao Wang, Xiaohua Zhai, Lucas Beyer, Andreas Steiner, Jesse Berent, Rodolphe Jenatton, and Effrosyni Kokiopoulou. Three towers: Flexible contrastive learning with pretrained image models. *NeurIPS*, 2023.

[20] Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, et al. Cell2sentence: Teaching large language models the language of biology. *International Conference on Machine Learning*, 2024.

[21] Yingcong Li, Yixiao Huang, Muhammed E Ildiz, Ankit Singh Rawat, and Samet Oymak. Mechanics of next token prediction with self-attention. *International Conference on Artificial Intelligence and Statistics*, 2024.

[22] Roey Magen, Shuning Shang, Zhiwei Xu, Spencer Frei, Wei Hu, and Gal Vardi. Benign overfitting in single-head attention. *arXiv preprint*, 2024.

[23] Ashok Vardhan Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Hyeji Kim, Michael Gastpar, and Chanakya Ekbote. Local to global: Learning dynamics and effect of initialization for transformers. *NeurIPS*, 2024.

[24] Ashok Vardhan Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Martin Jaggi, Hyeji Kim, and Michael Gastpar. Attention with markov: A curious case of single-layer transformers. *International Conference on Learning Representations*, 2025.

[25] Pierre Marion, Raphaël Berthier, Gérard Biau, and Claire Boyer. Attention layers provably solve single-location regression. *International Conference on Learning Representations*, 2025.

[26] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *Workshop Contribution at International Conference on Learning Representations*, 2015.

[27] Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. *International Conference on Machine Learning*, 2024.

[28] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint*, 2022.

[29] Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. *International Conference on Machine Learning*, 2023.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 2021.

[31] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *International Conference on Machine Learning*, 2021.

[32] Keitaro Sakamoto and Issei Sato. Benign or not-benign overfitting in token selection of attention mechanism. *arXiv preprint*, 2024.

[33] Clayton Sanford, Daniel J Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. *NeurIPS*, 2023.

[34] Heejune Sheen, Siyu Chen, Tianhao Wang, and Harrison H Zhou. Implicit regularization of gradient flow on one-layer softmax attention. *Workshop Contribution at International Conference on Learning Representations*, 2025.

[35] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 2018.

[36] Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *Workshop Contribution at NeurIPS*, 2023.

[37] Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. *NeurIPS*, 2023.

[38] Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Shaolei Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *NeurIPS*, 2023.

[39] Lorenzo Tiberi, Francesca Mignacco, Kazuki Irie, and Haim Sompolinsky. Dissecting the interplay of attention paths in a statistical mechanics theory of transformers. *NeurIPS*, 2024.

[40] Bhavya Vasudeva, Puneesh Deora, and Christos Thrampoulidis. Implicit bias and fast convergence rates for self-attention. *Transactions on Machine Learning Research*, 2025.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[42] James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. A mathematical theory of attention. *arXiv preprint*, 2020.

[43] Zixuan Wang, Stanley Wei, Daniel Hsu, and Jason D Lee. Transformers provably learn sparse token selection while fully-connected nets cannot. *International Conference on Machine Learning*, 2024.

[44] Hongru Yang, Bhavya Kailkhura, Zhangyang Wang, and Yingbin Liang. Training dynamics of transformers to recognize word co-occurrence via gradient flow analysis. *NeurIPS*, 2024.

[45] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. One fits all: Power general time series analysis by pretrained lm. *NeurIPS*, 2023.

Additional notation. Throughout the appendices, to simplify the notation, we write

$$a_i(X) := p^\top E_{x_i}, \qquad q_i(X) := \frac{\exp(a_i(X))}{\sum_{j=1}^T \exp(a_j(X))}, \tag{10}$$

so that $f(X; p, \boldsymbol{E}) = \sum_{i=1}^T q_i(X) E_{x_i}^\top v$. We will drop the dependence on X in $a_i(X), q_i(X)$ when there is no confusion. We also denote

$$\gamma_i(X, y) := y E_{x_i}^\top v, \tag{11}$$

dropping again the dependency on X, y when there is no confusion. Finally, we define

$$g(X, y) := \frac{1}{1 + \exp(y f(X; p, \boldsymbol{E}))}. \tag{12}$$

Properties of initialization. By standard concentration inequalities, with probability at least $1 - \delta$, at initialization we have

$$\begin{aligned} \max\left\{ \max_{s \in \mathcal{S}} |E_s^\top v|, \max_{s \in \mathcal{S}} |E_s^\top p|, |p^\top v| \right\} &\le \frac{1}{\sqrt{d}} \sqrt{2 \log \frac{|\mathcal{S}|^2}{\delta}}, \\ \max\left\{ \max_{s \in \mathcal{S}} \|E_s\|_2, \|p\|_2 \right\} &\le 2. \end{aligned} \tag{13}$$

For all results of the paper holding with probability at least $1 - \delta$, we will be implicitly conditioning on (13).

Appendix A. Additional related work

A line of work [23, 24, 27] has explored whether attention-based architectures can extract causal structure from Markovian inputs. The mechanics of next-token prediction when training a single self-attention layer is characterized in [21]. Towards understanding how to utilize structural properties of the data, the behavior of transformers on sparse token selection tasks is considered in [33, 43]. The study [14] provides a theoretical justification to the tendency of modern language models to generate repetitive text by showing that the underlying self-attention mechanism collapses into sampling only a limited subset of tokens. This stands in contrast to the slightly different setup of [38] where the transformer model does not degrade to a "winner-takes-all" strategy. The works [9–11] take a mean-field view to analyze the clustering behavior in transformer representations that emerges after successive applications of the attention block. Under a random feature design, it is shown in [6] that softmax attention exhibits a sensitivity property which allows for a sharp change in attention scores given the perturbation of a single token. The role of the attention mechanism is also studied in [29] for prompt-tuning and in [12] for test-time-training.

Appendix B. Technical lemmas

Lemma 7 *The gradients of the empirical loss are given by*

$$\nabla_{E_s}\mathcal{L}(\boldsymbol{E},p) = -\widehat{\mathbb{E}}\left[yg(X,y)\left(\sum_{i=1}^{T}(\sum_{j\neq i}(\mathbb{1}_{x_i=s}-\mathbb{1}_{x_j=s})q_iq_j)E_{x_i}^\top vp + \sum_{i=1}^{T}\mathbb{1}_{x_i=s}q_iv\right)\right],$$

$$\nabla_{p}\mathcal{L}(\boldsymbol{E},p) = -\widehat{\mathbb{E}}\left[yg(X,y)\left(\sum_{i=1}^{T}(\sum_{j\neq i}q_iq_j(E_{x_i}-E_{x_j}))E_{x_i}^\top v\right)\right],$$

where we have defined $g(X,y)=\frac{1}{1+\exp(yf(X))}$.

Proof We start by taking the gradient of q_i as

$$\begin{aligned}
\nabla_{E_s}q_i(X) &= \frac{\mathbb{1}_{x_i=s}\exp\left(E_{x_i}^\top p\right)p\left(\sum_{j=1}^{T}\exp\Big(E_{x_j}^\top p\Big)\right)-\left(\sum_{j=1}^{T}\mathbb{1}_{x_j=s}\exp\Big(E_{x_j}^\top p\Big)p\right)\exp\big(E_{x_i}^\top p\big)}{\left(\sum_{j=1}^{T}\exp\Big(E_{x_j}^\top p\Big)\right)^2}\\
&= \frac{p\sum_{j=1}^{T}(\mathbb{1}_{x_i=s}-\mathbb{1}_{x_j=s})\exp\Big(E_{x_j}^\top p\Big)\exp\left(E_{x_i}^\top p\right)}{\left(\sum_{j=1}^{T}\exp\Big(E_{x_j}^\top p\Big)\right)^2}\\
&= p\left(\sum_{j=1}^{T}(\mathbb{1}_{x_i=s}-\mathbb{1}_{x_j=s})q_iq_j\right)\\
&= p\left(\sum_{j\neq i}(\mathbb{1}_{x_i=s}-\mathbb{1}_{x_j=s})q_iq_j\right),
\end{aligned}$$

$$\begin{aligned}
\nabla_{p}q_i(X) &= \frac{\big(\exp\big(E_{x_i}^\top p\big)E_{x_i}\big)\left(\sum_{j=1}^{T}\exp\Big(E_{x_j}^\top p\Big)\right)-\sum_{j=1}^{T}\exp\Big(E_{x_j}^\top p\Big)E_{x_j}\exp\big(E_{x_i}^\top p\big)}{\left(\sum_{j=1}^{T}\exp\Big(E_{x_j}^\top p\Big)\right)^2}\\
&= \frac{\sum_{j=1}^{T}\exp\Big(E_{x_j}^\top p\Big)\exp\big(E_{x_i}^\top p\big)(E_{x_i}-E_{x_j})}{\left(\sum_{j=1}^{T}\exp\Big(E_{x_j}^\top p\Big)\right)^2}\\
&= \sum_{j=1}^{T}q_iq_j(E_{x_i}-E_{x_j})\\
&= \sum_{j\neq i}q_iq_j(E_{x_i}-E_{x_j}).
\end{aligned}$$

Next, we look at the gradient of $f(X; p, \boldsymbol{E})$:

$$\begin{aligned}
\nabla_{E_s} f(X; p, \boldsymbol{E}) &= \sum_{i=1}^{T} \left(\nabla_{E_s} q_i\right) E_{x_i}^\top v + \sum_{i=1}^{T} \mathbb{1}_{x_i=s} \cdot q_i v \\
&= \sum_{i=1}^{T} \left(\sum_{j \neq i} (\mathbb{1}_{x_i=s} - \mathbb{1}_{x_j=s}) q_i q_j \right) E_{x_i}^\top v p + \sum_{i=1}^{T} \mathbb{1}_{x_i=s} \cdot q_i v, \\
\nabla_{p} f(X; p, \boldsymbol{E}) &= \sum_{i=1}^{T} \left(\sum_{j \neq i} q_i q_j (E_{x_i} - E_{x_j}) \right) E_{x_i}^\top v.
\end{aligned}$$

This allows us to conclude that

$$\begin{aligned}
\nabla_{E_s} \mathcal{L}(\boldsymbol{E}, p) &= \widehat{\mathbb{E}} \left[\frac{-y}{1 + \exp(y f(X; p, \boldsymbol{E}))} \nabla_{E_s} f(X; p, \boldsymbol{E}) \right] \\
&= \widehat{\mathbb{E}} \left[\frac{-y}{1 + \exp(y f(X; p, \boldsymbol{E}))} \left(\sum_{i=1}^{T} \left(\sum_{j \neq i} (\mathbb{1}_{x_i=s} - \mathbb{1}_{x_j=s}) q_i q_j \right) E_{x_i}^\top v p + \sum_{i=1}^{T} \mathbb{1}_{x_i=s} q_i v \right) \right], \\
\nabla_{p} \mathcal{L}(\boldsymbol{E}, p) &= \widehat{\mathbb{E}} \left[\frac{-y}{1 + \exp(y f(X; p, \boldsymbol{E}))} \nabla_{p} f(X; p, \boldsymbol{E}) \right] \\
&= \widehat{\mathbb{E}} \left[\frac{-y}{1 + \exp(y f(X; p, \boldsymbol{E}))} \left(\sum_{i=1}^{T} \left(\sum_{j \neq i} q_i q_j (E_{x_i} - E_{x_j}) \right) E_{x_i}^\top v \right) \right],
\end{aligned}$$

thus concluding the proof. ■

Lemma 8 *For any vector $\widehat{p}$, we have*

$$-\widehat{p}^\top \nabla_p \mathcal{L}(\boldsymbol{E}, p) = \widehat{\mathbb{E}} \left[g(X, y) \left(\sum_{i=1}^{T} \sum_{j>i} (\widehat{a}_i(X) - \widehat{a}_j(X)) q_i(X) q_j(X) (\gamma_i(X, y) - \gamma_j(X, y)) \right) \right],$$

where $\widehat{a}_i = \widehat{p}^\top E_{x_i}$ for all $i \in \{1, \ldots, T\}$.

Proof From Lemma 7, we have

$$\begin{aligned}
\nabla_p \mathcal{L}(\boldsymbol{E}, p) &= -\widehat{\mathbb{E}} \left[y g(X, y) \left(\sum_{i=1}^{T} \left(\sum_{j \neq i} q_i(X) q_j(X) (E_{x_i} - E_{x_j}) \right) E_{x_i}^\top v \right) \right] \\
&= -\widehat{\mathbb{E}} \left[g(X, y) \left(\sum_{i=1}^{T} \left(\sum_{j \neq i} q_i(X) q_j(X) (E_{x_i} - E_{x_j}) \right) \gamma_i(X, y) \right) \right] \\
&= -\widehat{\mathbb{E}} \left[g(X, y) \boldsymbol{E}_X^\top \left(\mathrm{Diag}(q_X) - q_X q_X^\top \right) \gamma(X, y) \right],
\end{aligned}$$

where $q_X = [q_1(X), \ldots, q_T(X)]^\top, \gamma(X, y) = [\gamma_1(X, y), \ldots \gamma_T(X, y)]^\top$ and $\mathrm{Diag}(q_X)$ denotes the diagonal matrix with $[\mathrm{Diag}(q_X)]_{i,i} = q_i(X)$.

Thus, letting $\widehat{a} = [\widehat{a}_1, \ldots, \widehat{a}_T] \in \mathbb{R}^T$ with $\widehat{a}_i = \hat{p}^\top E_{x_i}$, we have

$$\begin{aligned}
-\widehat{p}^\top \nabla_p \mathcal{L}(\boldsymbol{E}, p) &= \widehat{\mathbb{E}}\left[g(X,y)\widehat{p}^\top \boldsymbol{E}_X^\top (\mathrm{Diag}(q_X) - q_X q_X^\top)\gamma(X,y)\right] \\
&= \widehat{\mathbb{E}}\left[g(X,y)\widehat{a}^\top (\mathrm{Diag}(q_X) - q_X q_X^\top)\gamma(X,y)\right] \\
&= \widehat{\mathbb{E}}\left[g(X,y)\left(\sum_{i=1}^T \widehat{a}_i q_i(1-q_i)\gamma_i - \sum_{i=1}^T \sum_{j\neq i} \widehat{a}_i q_i q_j \gamma_j\right)\right] \\
&= \widehat{\mathbb{E}}\left[g(X,y)\left(\sum_{i=1}^T \sum_{j\neq i} \widehat{a}_i q_i q_j (\gamma_i - \gamma_j)\right)\right] \quad (\text{use } 1 - q_i = \sum_{j\neq i} q_j) \\
&= \widehat{\mathbb{E}}\left[g(X,y)\left(\frac{1}{2}\sum_{i=1}^T \sum_{j\neq i} \widehat{a}_i q_i q_j (\gamma_i - \gamma_j) + \frac{1}{2}\sum_{j=1}^T \sum_{i\neq j} \widehat{a}_j q_i q_j (\gamma_j - \gamma_i)\right)\right] \\
&= \widehat{\mathbb{E}}\left[g(X,y)\left(\frac{1}{2}\sum_{i=1}^T \sum_{j\neq i} (\widehat{a}_i - \widehat{a}_j) q_i q_j (\gamma_i - \gamma_j)\right)\right] \\
&= \widehat{\mathbb{E}}\left[g(X,y)\left(\sum_{i=1}^T \sum_{j>i} (\widehat{a}_i - \widehat{a}_j) q_i q_j (\gamma_i - \gamma_j)\right)\right].
\end{aligned}$$

■

Lemma 9 (Convergence lemma) *Let $\|p_t\|_2 \to \infty$ and suppose there exists $\widehat{p}$ such that, for any $\epsilon > 0$, there is a $\bar{t}(\epsilon)$ ensuring*

$$-\frac{\widehat{p}^\top}{\|\widehat{p}\|_2}\nabla_p \mathcal{L}(\boldsymbol{E}, p_t) \geq -(1-\epsilon)\frac{p_t^\top}{\|p_t\|_2}\nabla_p \mathcal{L}(\boldsymbol{E}, p_t), \qquad \text{for all } t \geq \bar{t}(\epsilon). \tag{14}$$

Then, if $\lim_{t\to\infty} \frac{p_t}{\|p_t\|_2}$ exists, we have

$$\lim_{t\to\infty} \frac{p_t}{\|p_t\|_2} = \frac{\widehat{p}}{\|\widehat{p}\|_2}.$$

Proof By the definition of the gradient flow, (14) is equivalent to

$$\frac{\widehat{p}^\top}{\|\widehat{p}\|_2}\frac{\mathrm{d}p_t}{\mathrm{d}t} \geq (1-\epsilon)\frac{p_t^\top}{\|p_t\|_2}\frac{\mathrm{d}p_t}{\mathrm{d}t}.$$

We note that

$$\frac{p_t^\top}{\|p_t\|_2}\frac{\mathrm{d}p_t}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\|p_t\|_2.$$

Thus, by integrating both sides from $[\bar{t}(\epsilon), t]$, we have:

$$\frac{\widehat{p}^\top}{\|\widehat{p}\|_2}(p_t - p_{\bar{t}(\epsilon)}) \geq (1-\epsilon)(\|p_t\|_2 - \|p_{\bar{t}(\epsilon)}\|_2),$$

which gives

$$\frac{\widehat{p}^\top p_t}{\|\widehat{p}\|_2\|p_t\|_2} \geq (1-\epsilon) - (1-\epsilon)\frac{\|p_{\bar{t}(\epsilon)}\|_2}{\|p_t\|_2} + \frac{\widehat{p}^\top p_{\bar{t}(\epsilon)}}{\|\widehat{p}\|_2\|p_t\|_2}.$$

Since $p_{\bar{t}(\epsilon)}, \widehat{p}$ have finite norm for fixed ϵ, by taking the limit on both sides, we have

$$\liminf_{t\to\infty} \frac{\widehat{p}^\top p_t}{\|\widehat{p}\|_2\|p_t\|_2} \geq 1-\epsilon.$$

As we assume that $\lim_{t\to\infty} \frac{p_t}{\|p_t\|_2}$ exist and the above argument holds for any ϵ, we conclude

$$\lim_{t\to\infty} \frac{p_t}{\|p_t\|_2} = \frac{\hat{p}}{\|\hat{p}\|_2}.$$

■

Lemma 10 *Given a sequence X, model parameters $\boldsymbol{E}, p, v$, and indices i_*, j s.t. $x_{i_*} \in \mathcal{S}_X(p), x_j \in X \setminus \mathcal{S}_X(p)$, the following results hold.*

1. *We have*
$$\frac{1}{T} \leq q_{i_*} \leq 1.$$
2. *If there exist $\tau > 0$ such that $p^\top(E_{x_{i_*}} - E_{x_j}) \geq \tau$ for all $x_{i_*} \in \mathcal{S}_X(p)$, then we have*
$$q_j \leq \frac{1}{1+\exp(\tau)}.$$
3. *If there exist $\tau > 0$ such that $p^\top(E_{x_{i_*}} - E_{x_j}) \leq \tau$ for all $x_{i_*} \in \mathcal{S}_X(p)$, then we have*
$$q_j \geq \frac{1}{T\exp(\tau)}.$$

Proof The upper bound on q_{i_*} is trivial. For the lower bound:

$$\begin{aligned} q_{i_*} &= \frac{\exp\big(p^\top E_{x_{i_*}}\big)}{\exp\big(p^\top E_{x_{i_*}}\big) + \sum_{j\neq i_*}\exp\big(p^\top E_{x_j}\big)} \\ &\geq \frac{\exp\big(p^\top E_{x_{i_*}}\big)}{T\exp\big(p^\top E_{x_{i_*}}\big)} = \frac{1}{T}. \end{aligned}$$

If there exists $\tau > 0$ such that $p^\top(E_{x_{i_*}} - E_{x_j}) \geq \tau$ for all $x_i \in \mathcal{S}_X(p)$, then we have

$$\begin{aligned} q_j &= \frac{1}{1+\sum_{i\neq j}\exp\big(p^\top(E_{x_i} - E_{x_j})\big)} \\ &\leq \frac{1}{1+\exp\big(p^\top(E_{x_{i_*}} - E_{x_j})\big)} \\ &\leq \frac{1}{1+\exp(\tau)}. \end{aligned}$$

If there exists $\tau > 0$ such that $p^\top (E_{x_{i_*}} - E_{x_j}) \leq \tau$ for all $x_{i_*} \in \mathcal{S}_X(p)$, then we have

$$\begin{aligned} q_{i_*} &= \frac{1}{1 + \sum_{i \neq j} \exp\big(p^\top (E_{x_i} - E_{x_j})\big)} \\ &\geq \frac{1}{1 + (T-1)\exp\big(p^\top (E_{x_{i_*}} - E_{x_j})\big)} \quad \text{(by definition of } \mathcal{S}_X(p)\text{)} \\ &\geq \frac{1}{T\exp(\tau)}. \end{aligned}$$

■

Appendix C. Properties after the first gradient step

Lemma 11 (Boundedness of the embeddings) *For any $\delta > 0$, let*

$$d \geq \max\left\{256, \left(2\log\frac{|\mathcal{S}|^2}{\delta}\right)^2\right\},$$

then with probability at least $1 - \delta$,

$$\max_{s \in \mathcal{S}} \|E_s^1\|_2 \leq 2(1 + 2\eta_0), \qquad \|p^1\|_2 \leq 2 + 11\eta_0 d^{-\frac{1}{4}}.$$

Proof By using (13), we have that

$$\begin{aligned} \max_{s \in \mathcal{S}} \|E_s^1\|_2 &\leq \max_s \left(\|E_s^0\|_2 + \frac{\eta_0}{2}\|v\|_2 + \|err_s\|_2\right) \\ &\leq \max_{s \in \mathcal{S}} \left(2 + \frac{\eta_0}{2} + 11\eta_0 d^{-\frac{1}{4}}\right) \\ &\leq 2 + 4\eta_0, \end{aligned}$$

and that

$$\|p^1\|_2 \leq \|p^0\|_2 + \|\text{err}_p\|_2 \leq 2 + 11\eta_0 d^{-\frac{1}{4}}. \tag{15}$$

■

Lemma 12 (Upper bound on the loss) *For any $\delta > 0$, let*

$$d \geq \max\left\{256, \left(2\log\frac{|\mathcal{S}|^2}{\delta}\right)^2, (88\eta_0^2 + 111\eta_0 + 2)^8\right\},$$

then with probability at least $1 - \delta$,

$$\mathcal{L}(\boldsymbol{E}^1, p^1) \leq \widehat{\mathbb{E}}\left[\log\left(1 + \exp\left(-\frac{1}{T}\sum_{i=1}^{T}\frac{\eta_0}{2} y \alpha_{x_i} + \frac{1}{22\eta_0}\right)\right)\right].$$

Proof We first lower bound $yf(X;p,\boldsymbol{E})$ for each pair X, y. After the first step, we have

$$\max_{s,s'} |(p^1)^\top (E^1_s - E^1_{s'})| = \max_{s,s'} \Big| (p^0)^\top (E^0_s - E^0_{s'}) + \frac{\eta_0}{2}(\alpha_s - \alpha_{s'})(p^0)^\top v + \mathrm{err}_p^\top (E^1_s - E^1_{s'}) + (\mathrm{err}_s - \mathrm{err}_{s'})^\top p^1 \Big|.$$

We bound each term separately:

$$\begin{aligned}
\max_{s,s'} |(p^0)^\top (E^0_s - E^0_{s'})| &\le 2 \max_s |(p^0)^\top E^0_s| \le 2d^{-\frac{1}{4}}, \\
\frac{\eta_0}{2}(\alpha_s - \alpha_{s'})|(p^0)^\top v| &\le \eta_0 |(p^0)^\top v| \le \eta_0 d^{-\frac{1}{4}}, \\
|\mathrm{err}_p^\top (E^1_s - E^1_{s'})| &\le \|\mathrm{err}_p^\top\|_2 \|E^1_s - E^1_{s'}\|_2 \le 44\eta_0 d^{-\frac{1}{4}}(1 + 2\eta_0), \\
|(\mathrm{err}_s - \mathrm{err}_{s'})^\top p^1| &\le 2\|p^1\|_2 \max_s \|\mathrm{err}_s\|_2 \le 22\eta_0 d^{-\frac{1}{4}} \left(2 + 11\eta_0 d^{-\frac{1}{4}}\right),
\end{aligned}$$

where we have used (13). By picking $d \ge (88\eta_0^2 + 111\eta_0 + 2)^8$, we get $\max_{s,s'} |(p^1)^\top (E^1_s - E^1_{s'})| \le d^{-\frac{1}{8}}$, which implies that, for any X and any $i \in \{1, \dots, T\}$,

$$\frac{1}{T} - \frac{2d^{-\frac{1}{8}}}{T} \le q_i(X) \le \frac{1}{T} + \frac{2d^{-\frac{1}{8}}}{T}.$$

Thus, we lower bound $yf(X;p,\boldsymbol{E})$ for each pair (X, y) as

$$\begin{aligned}
yf(X;p,\boldsymbol{E}) &= \sum_{i=1}^T q_i(X)\gamma_i(X) \\
&\ge \frac{1}{T}\sum_{i=1}^T \frac{\eta_0}{2} y\alpha_{x_i} - \sum_{i=1}^T \frac{2d^{-\frac{1}{8}}}{T}\frac{\eta_0}{2}\alpha_{x_i} + \sum_{i=1}^T y q_i(X) v^\top (E^0_{x_i} + \mathrm{err}_{x_i}) \\
&\ge \frac{1}{T}\sum_{i=1}^T \frac{\eta_0}{2} y\alpha_{x_i} - d^{-\frac{1}{8}}\eta_0 - (1 + 2d^{-\frac{1}{8}}) v^\top (E^0_{x_i} + \mathrm{err}_{x_i}) \\
&\ge \frac{1}{T}\sum_{i=1}^T \frac{\eta_0}{2} y\alpha_{x_i} - d^{-\frac{1}{8}}\eta_0 - 3(1 + 11\eta_0) d^{-\frac{1}{4}} \\
&\ge \frac{1}{T}\sum_{i=1}^T \frac{\eta_0}{2} y\alpha_{x_i} - \frac{1}{22\eta_0},
\end{aligned}$$

which allows us to conclude that

$$\begin{aligned}
\mathcal{L}(\boldsymbol{E}^1, p^1) &= \widehat{\mathbb{E}}\left[\log(1 + \exp(-yf(X;p,\boldsymbol{E})))\right] \\
&\le \widehat{\mathbb{E}}\left[\log\left(1 + \exp\left(-\frac{1}{T}\sum_{i=1}^T \frac{\eta_0}{2} y\alpha_{x_i} + \frac{1}{22\eta_0}\right)\right)\right].
\end{aligned} \tag{16}$$

■

Appendix D. Proofs for Section 3

D.1. Proof of Lemma 2

For simplicity, in the proof we drop the time dependency in all the variables. By picking

$$d \geq \left(2 \log \frac{|\mathcal{S}|^2}{\delta}\right)^2,$$

from (13) we have

$$\max\left\{\max_{s\in\mathcal{S}} |E_s^\top v|, \max_{s\in\mathcal{S}} |E_s^\top p|, |p^\top v|\right\} \leq d^{-\frac{1}{4}},$$
$$\max\left\{\max_{s\in\mathcal{S}} \|E_s\|_2, \|p\|_2\right\} \leq 2.$$

Thus, at initialization, we have that, for all s,

$$\exp\left(-d^{-\frac{1}{4}}\right) \leq \exp\left(p^\top E_s\right) \leq \exp\left(d^{-\frac{1}{4}}\right),$$

which implies that, for any sequence X and any position i,

$$\frac{1}{T+2T\left(d^{-\frac{1}{4}}\right)} \leq \frac{1}{1+(T-1)\exp\left(2d^{-\frac{1}{4}}\right)} \leq q_i(X) \leq \frac{1}{1+(T-1)\exp\left(-2d^{-\frac{1}{4}}\right)} \leq \frac{1}{T-2T\left(d^{-\frac{1}{4}}\right)},$$

where we use the fact that for $z \in [-1,1], 1-|z| \leq \exp(z) \leq 1+|z|$.

Furthermore, for $d > 256$ and for any sequence (X, y), we have

$$\frac{1}{T} - \frac{4d^{-\frac{1}{4}}}{T} \leq q_i(X) \leq \frac{1}{T} + \frac{4d^{-\frac{1}{4}}}{T},$$

and

$$-2d^{-\frac{1}{4}} \leq \frac{-Td^{-\frac{1}{4}}}{T-2Td^{-\frac{1}{4}}} \leq yf(X;p,\boldsymbol{E}) \leq \frac{Td^{-\frac{1}{4}}}{T-2Td^{-\frac{1}{4}}} \leq 2d^{-\frac{1}{4}}.$$

Then,

$$g(X,y) \leq \frac{1}{1+\exp\left(-2d^{-\frac{1}{4}}\right)} \leq \frac{1}{2-2d^{-\frac{1}{4}}} \leq \frac{1}{2} + d^{-\frac{1}{4}},$$

and similarly

$$g(X,y) \geq \frac{1}{2} - d^{-\frac{1}{4}}.$$

Now we look at the gradient update of the first step. By Lemma 7, we have

$$\begin{aligned}
-\nabla_{E_s}\mathcal{L}(\boldsymbol{E},p) &= \widehat{\mathbb{E}}\left[yg(X,y)\left(\sum_{i=1}^{T}\left(\sum_{j\neq i}(\mathbb{1}_{x_i=s}-\mathbb{1}_{x_j=s})q_iq_j\right)E_{x_i}^{\top}vp+\sum_{i=1}^{T}\mathbb{1}_{x_i=s}q_iv\right)\right]\\
&= \frac{1}{2T}\widehat{\mathbb{E}}\left[y\sum_{i=1}^{T}\mathbb{1}_{x_i=s}\right]v\\
&\quad+\frac{1}{2}\widehat{\mathbb{E}}\left[y\sum_{i=1}^{T}\mathbb{1}_{x_i=s}\left(q_i-\frac{1}{T}\right)\right]v\\
&\quad+\widehat{\mathbb{E}}\left[yg(X,y)\left(\sum_{i=1}^{T}\left(\sum_{j\neq i}(\mathbb{1}_{x_i=s}-\mathbb{1}_{x_j=s})q_iq_j\right)E_{x_i}^{\top}vp\right)\right]\\
&\quad+\widehat{\mathbb{E}}\left[y\left(g(X,y)-\frac{1}{2}\right)\sum_{i=1}^{T}\mathbb{1}_{x_i=s}q_iv\right],\\
-\nabla_{p}\mathcal{L}(\boldsymbol{E},p) &= \widehat{\mathbb{E}}\left[yg(X,y)\left(\sum_{i=1}^{T}\left(\sum_{j\neq i}q_iq_j(E_{x_i}-E_{x_j})E_{x_i}^{\top}v\right)\right)\right].
\end{aligned}$$

We note that

$$\frac{1}{2T}\widehat{\mathbb{E}}\left[y\sum_{i=1}^{T}\mathbb{1}_{x_i=s}\right]v=\frac{1}{2}\alpha_sv,$$

and we bound the remaining error terms.

We have that

$$\left\|\frac{1}{2}\widehat{\mathbb{E}}\left[y\sum_{i=1}^{T}\mathbb{1}_{x_i=s}\left(q_i-\frac{1}{T}\right)\right]v\right\|_2\leq d^{-\frac{1}{4}},$$

and

$$\left\|\widehat{\mathbb{E}}\left[yg(X,y)\left(\sum_{i=1}^{T}\left(\sum_{j\neq i}(\mathbb{1}_{x_i=s}-\mathbb{1}_{x_j=s})q_iq_j\right)E_{x_i}^{\top}vp\right)+y\left(g(X,y)-\frac{1}{2}\right)\sum_{i=1}^{T}\mathbb{1}_{x_i=s}q_iv\right]\right\|_2\leq 10d^{-\frac{1}{4}}.$$

Furthermore, we also have that

$$\|\nabla_p\mathcal{L}(\boldsymbol{E},p)\|_2\leq 8d^{-\frac{1}{4}}.$$

Thus, the desired claim follows.

D.2. Proof of Lemma 3

Proof We first show that, if (9) is feasible, then the solution is unique. Indeed, assume by contradiction that p_1, p_2 are two different solutions of (9). Clearly, p_1 and p_2 have the same norm, so

$\frac{p_1^\top p_2}{\|p_1\|_2\|p_2\|_2} \neq 1$. Then, any convex combination of p_1, p_2 gives a feasible solution with a strictly smaller norm, which is a contradiction.

Next, we show that (9) is always feasible. To see this, by definition, there exists some τ such that

$$p_\circ^\top (E_s - E_{s'}) \geq \tau, \qquad \forall s \in \mathcal{S}_X(p_\circ),\ \forall s' \in \overline{\mathcal{S}_X(p_\circ)},\ \forall X \in \mathcal{X}_n.$$

Then, $\frac{p_\circ}{\tau}$ is a feasible solution of (9) which concludes the proof of uniqueness.

To characterize $p_*(p_\circ)$, we first note that (9) can be equivalently written as:

$$\begin{aligned} &\arg\min_p \frac{1}{2}\|p\|_2^2 \\ &s.t. \quad p^\top (E_s - E_{s'}) \geq 1, \qquad \forall s \in \mathcal{S}_X(\mathcal{P}_{p_\circ}),\ \forall s' \in \overline{\mathcal{S}_X(\mathcal{P}_{p_\circ})},\ \forall X \in \mathcal{X}_n. \end{aligned} \tag{17}$$

Now we characterize the solution of (17) explicitly. First of all, we can rewrite the constraints as

$$\mathbf{1}_N - \boldsymbol{M} p \leq 0.$$

Then we can write the Lagrangian of (17) as

$$\mathtt{L}(p, \lambda) = \frac{1}{2}\|p\|_2^2 + \lambda^\top (\mathbf{1}_N - \boldsymbol{M} p),$$

where $\lambda \in \mathbb{R}^N$ and p is a KKT point if

$$\begin{aligned} \nabla_p \mathtt{L}(p, \lambda) &= p - \boldsymbol{M}^\top \lambda = 0, \\ \nabla_\lambda \mathtt{L}(p, \lambda) &= \mathbf{1}_N - \boldsymbol{M} p = 0. \end{aligned}$$

Since the objective function is convex and the constraints are affine, the global optimum is achieved at the KKT point, which satisfies $\boldsymbol{M} p = \mathbf{1}_N$. Thus, if there exists a p satisfying this condition, we can rewrite (17) as

$$\begin{aligned} &\arg\min_p \frac{1}{2}\|p\|_2^2 \\ &s.t. \quad \boldsymbol{M} p = \mathbf{1}_N, \end{aligned}$$

whose solution is

$$\hat{p} = \boldsymbol{M}^\dagger \mathbf{1}_N.$$

It remains to show that there exists a feasible p. Since $d > |\mathcal{S}| + 2$, we have that, with high-probability, $\boldsymbol{E}^0$ is full rank. Furthermore, $\boldsymbol{E}^1 = \boldsymbol{E}^0 + \boldsymbol{\Delta}$ and each row of $\boldsymbol{\Delta}$ is in the subspace generated by v and p_0. Thus, we can pick $\widehat{p} \perp v, p_0$, so that

$$\boldsymbol{E}^1 \widehat{p} = \boldsymbol{E}^0 \widehat{p}.$$

Then, we define $a \in \mathbb{R}^{|\mathcal{S}|}$ such that $a_i = 1$ for all $i \in \bigcup_X \mathcal{S}_X(p_\circ)$, and $a_i = 0$ otherwise. Let

$$\boldsymbol{E}^0 \widehat{p} = a.$$

Since $d > |\mathcal{S}|$ and $\boldsymbol{E}^0$ has full row rank, there exists a non-zero $\widehat{p}$ that solves the above equation, which finishes the proof.

■

D.3. Proof of Theorem 5

We prove each part separately. We first show that $\lim_{t\to\infty} \|p_t\|_2 = \infty$.

Lemma 13 *Under Assumption 4, for any $\delta > 0$, by picking*

$$d \geq \max\left\{256, \left(2\log\frac{|\mathcal{S}|^2}{\delta}\right)^2, |\mathcal{S}| + 3\right\},$$

with probability at least $1-\delta$, we have $\lim_{t\to\infty} \|p_t\|_2 = \infty$.

Proof It is sufficient to show that there exists a non-zero finite-norm $\widehat{p}$, such that for any finite norm p,

$$\widehat{p}^\top \nabla_p \mathcal{L}(\boldsymbol{E}^1, p) \neq 0.$$

Indeed, the above condition means that there is no stationary point for any finite-norm p. For gradient flow, we have that

$$\lim_{t\to\infty} \nabla_p \mathcal{L}(\boldsymbol{E}^1, p_t) = 0,$$

which by contradiction implies the desired result.

Now we construct such $\widehat{p}$. Since $d > |\mathcal{S}| + 2$, we have that with high-probability $\boldsymbol{E}^0$ is full rank. Furthermore, $\boldsymbol{E}^1 = \boldsymbol{E}^0 + \boldsymbol{\Delta}$ and each row of $\boldsymbol{\Delta}$ is in the subspace generated by v and p_0. Thus, we can pick $\widehat{p} \perp v, p_0$, so that

$$\boldsymbol{E}^1\widehat{p} = \boldsymbol{E}^0\widehat{p}.$$

Without loss of generality, let x_1 be an important token in a positive sequence X_k, i.e., $\gamma_1(X_k) \geq \frac{\eta_0}{4nT}$. Then, we define $a \in \mathbb{R}^{|\mathcal{S}|}$ such that $a_1 = 1$ and $a_i = 0$ for all $i \neq 1$. Let

$$\boldsymbol{E}^0\widehat{p} = a.$$

Since $d > |\mathcal{S}|$ and $\boldsymbol{E}^0$ has full row rank, there exists a non-zero $\widehat{p}$ that solves the above equation. By Lemma 8, we have that, for any p,

$$\begin{aligned}
-\widehat{p}^\top \nabla_p \mathcal{L}(\boldsymbol{E}^1, p) &= \widehat{\mathbb{E}}\left[g(X,y)\left(\sum_{i=1}^{T}\sum_{j>i}(a_i - a_j)q_i q_j(\gamma_i - \gamma_j)\right)\right] \\
&= g(X_k, y_k)\sum_{j>1} q_1(X_k)q_j(X_k)\frac{\eta_0}{4nT} > 0,
\end{aligned}$$

which concludes the proof. ∎

Next, we show that, if the directional limit exists, then it must select all completely positive/negative tokens.

Lemma 14 *Under Assumption 4, for any $\delta > 0$, by picking*

$$\eta_0 \geq 4n^2T^2, \quad d \geq \max\left\{256, \left(2\log\frac{|\mathcal{S}|^2}{\delta}\right)^2, (88\eta_0^2 + 111\eta_0 + 2)^8\right\},$$

with probability at least $1-\delta$, if $p_\infty = \lim_{t\to\infty} \frac{p_t}{\|p_t\|_2}$ exists, then p_∞ satisfies

$$s_*^X \in \mathcal{S}_X(p_\infty), \qquad \textit{for all } X \in \mathcal{X}_n,$$

where s_*^X denotes the unique completely positive/negative token in the sequence X.

Proof We prove the lemma by contradiction. W.l.o.g., assume by contradiction that there exists $X \in \mathcal{X}_n$ cointaining the important token x_1 s.t. $x_1 \notin \mathcal{S}_X(p_\infty)$. We show that there exists $\bar{t}$ such that, for all $t \geq \bar{t}$,

$$\mathcal{L}(\boldsymbol{E}^1, p^t) > \mathcal{L}(\boldsymbol{E}^1, p^1),$$

which contradicts the fact that the gradient flow always decreases the loss.

To see this, we first note that by the definition of $\mathcal{S}_X(p_\infty)$, there exists some $\tau > 0$ independent of t such that

$$\min_{j \neq 1} p_\infty^\top (E_{x_1} - E_{x_j}) = -\tau.$$

W.l.o.g, we assume that x_2 is the token that achieves the minimum.

As $\lim_{t\to\infty} \|p_t\|_2 = \infty$ and $\lim_{t\to\infty} \frac{p_t}{\|p_t\|_2} = p_\infty$, we have that, for any $\mu > 0, R > 0$, there exists a large enough $\bar{t}$ such that

$$\|p_t\|_2 \geq 2R, \quad \left\| \frac{p_t}{\|p_t\|_2} - p_\infty \right\|_2 \leq \mu, \qquad \text{for all } t \geq \bar{t}.$$

Thus, we have:

$$\begin{aligned}
\frac{p_t^\top}{\|p_t\|_2}(E_{x_1} - E_{x_2}) &= p_\infty^\top (E_{x_1} - E_{x_2}) + \left(\frac{p_t}{\|p_t\|_2} - p_\infty \right)^\top (E_{x_1} - E_{x_2}) \\
&\leq -\tau + 2\mu(4\eta_0 + 2)^2,
\end{aligned}$$

where we have used the result of Lemma 11. Thus, by picking $\mu = \frac{\tau}{4(4\eta_0+2)^2}$, we have

$$\frac{p_t^\top}{\|p_t\|_2}(E_{x_1} - E_{x_2}) \leq -\frac{\tau}{2},$$

which implies that

$$p_t^\top (E_{x_1} - E_{x_2}) \leq -\tau R.$$

Next, we upper bound $yf(X; p_t, \boldsymbol{E}^1)$. We first note that

$$\frac{q_1}{q_2} = \exp\Big(p_t^\top (E_{x_1} - E_{x_2}) \Big) \leq \exp(-\tau R),$$

which gives

$$q_1 \leq \exp(-\tau R).$$

Note that

$$\begin{aligned}
yf(X; p_t, \boldsymbol{E}^1) &= \sum_{i=1}^{T} q_i \gamma_i \\
&\leq \exp(-\tau R)\gamma_1 + \max_{j \neq 1} \gamma_j \\
&\leq \exp(-\tau R) \left(\frac{\eta_0}{2} + (1 + 11\eta_0) d^{-\frac{1}{4}} \right) + (1 + 11\eta_0) d^{-\frac{1}{4}}.
\end{aligned}$$

Thus, by picking $R \geq \frac{\log d}{4\tau}$, we have

$$yf(X; p_t, \boldsymbol{E}^1) \leq \left(\frac{3}{2} + \frac{23}{2}\eta_0 \right) d^{-\frac{1}{4}} \leq \frac{3}{4} d^{-\frac{1}{8}},$$

which implies a lower bound on the loss:

$$\mathcal{L}(\boldsymbol{E}^1, p_t) \geq \frac{1}{n}\log\big(1+\exp\big(-yf(X;p_t,\boldsymbol{E}^1)\big)\big) \geq \frac{1}{n}\log\left(1+\exp\left(-\frac{3}{4}d^{-\frac{1}{8}}\right)\right) \geq \frac{1}{2n}, \quad (18)$$

where we used that $d \geq 256$ in the last passage. Under Assumption 4, by Lemma 2, we have that $y\alpha_{x_i} \geq 1/(nT)$ if x_i is either the completely positive or the completely negative token in X, and otherwise $y\alpha_{x_i} = 0$. Hence, given that each sequence X contains a completely positive or negative token, we have that

$$\frac{1}{T}\sum_{i=1}^{T} y\alpha_{x_i} \geq \frac{1}{nT^2}.$$

As $\eta_0 > 4n^2T^2 > \sqrt{2nT^2/11}$, by applying Lemma 12, we obtain

$$\mathcal{L}(\boldsymbol{E}^1, p_1) \leq \log\left(1+\exp\left(-\frac{\eta_0}{4nT^2}\right)\right) \leq \log(1+\exp(-n)) \leq \exp(-n) < \frac{1}{2n},$$

which gives a contradiction and concludes the proof. ■

Finally, we show that for each possible selection, if p_t converges in direction, it must converge to the max-margin solution. In particular, we first prove the following lemma which gives an approximation to the directional gradient of the locally optimal selection. To do so, we define the secondary selection set and the locally optimal selection as follows:

Definition 15 *Given a vector p, for each sequence X, denote by $\mathcal{S}_X^2(p)$ the secondary selection set given by*

$$\mathcal{S}_X^2(p) = \arg\max\{s : p^\top E_s, s \notin \mathcal{S}_X(p)\}. \quad (19)$$

We also denote by $\mathcal{S}_X^<(p)$ the set of tokens that are not chosen in the first and in the second place, i.e.,

$$\mathcal{S}_X^<(p) = X \setminus (\mathcal{S}_X(p) \bigcup \mathcal{S}_X^2(p)). \quad (20)$$

Definition 16 *Given a vector p, we say that p is locally optimal if for every (X, y) pair, we have*

$$\sum_{i\in\mathcal{S}_X(p)} (\gamma_i(X,y) - \gamma_j(X,y)) \geq \mu > 0, \qquad \text{for all } j \in \mathcal{S}_X^2(p),$$

for some constant μ that does not depends on p.

In the definition above and for the rest of this appendix, to help readability, we will abuse notation by letting indices (e.g., i, j above) also denote the corresponding tokens (e.g., x_i, x_j above).

Lemma 17 *Let $\overline{p}$ be a unit-norm vector and $p = R\overline{p}$ for some positive constant R. Suppose $\overline{p}$ is a locally optimal direction as defined in Definition 16 with some μ that does not depends on R. Moreover, suppose there exists a constant τ_1 that may depend on $\overline{p}, \eta_0, n, T, d$ but not on R, such that:*

$$\begin{aligned} &\min_X\{\overline{p}^\top(E_s - E_{s'}), \forall s \in \mathcal{S}_X(p), \forall s' \in \mathcal{S}_X^2(p)\} \geq \tau_1, \\ &\min_X\{\overline{p}^\top(E_s - E_{s'}), \forall s \in \mathcal{S}_X^2(p), \forall s' \in \mathcal{S}_X^<(p)\} \geq \tau_1. \end{aligned} \quad (21)$$

Then, for any $\epsilon > 0$, for any $\hat{p} \cong p$ such that $\|\hat{p}\|_2$ does not depend on R and

$$\min_X \{\hat{p}^\top (E_s - E_{s'}), \forall s \in \mathcal{S}_X(\hat{p}), \forall s' \in X \setminus \mathcal{S}_X(\hat{p})\} \geq \tau_2,$$

there exists R large enough such that:

$$-\hat{p}^\top \nabla \mathcal{L}(\boldsymbol{E}^1, p) \leq (1+\epsilon)\widehat{\mathbb{E}}\left[\sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\widehat{a}_i(X) - \widehat{a}_j(X)) h_{i,j}(X, y, p)\right],$$

$$-\hat{p}^\top \nabla \mathcal{L}(\boldsymbol{E}^1, p) \geq (1-\epsilon)\widehat{\mathbb{E}}\left[\sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\widehat{a}_i(X) - \widehat{a}_j(X)) h_{i,j}(X, y, p)\right],$$

where $\widehat{a}_i(X) = \hat{p}^\top E_{x_i}, \widehat{a}_j(X) = \hat{p}^\top E_{x_j}$ and

$$h_{i,j}(X, y, p) = g(X, y) q_i(X) q_j(X) (\gamma_i(X, y) - \gamma_j(X, y)).$$

Proof By Lemma 8, we can write the directional gradient as follows:

$$\begin{aligned}
-\widehat{p}^\top \nabla_p \mathcal{L}(\boldsymbol{E}^1, p) &= \widehat{\mathbb{E}}\left[\sum_{i=1}^{T} \sum_{j>i} (\widehat{a}_i(X) - \widehat{a}_j(X)) h_{i,j}(X, y, p))\right] \\
&= \widehat{\mathbb{E}}\left[\sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\widehat{a}_i(X) - \widehat{a}_j(X)) h_{i,j}(X, y, p)\right] \qquad \text{(B0)} \\
&\quad + \widehat{\mathbb{E}}\left[\sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^<(p)} (\widehat{a}_i(X) - \widehat{a}_j(X)) h_{i,j}(X, y, p))\right] \qquad \text{(B1)} \\
&\quad + \widehat{\mathbb{E}}\left[\sum_{i \in X \setminus \mathcal{S}_X(p)} \sum_{j > i : j \in X \setminus \mathcal{S}_X(p)} (\widehat{a}_i(X) - \widehat{a}_j(X)) h_{i,j}(X, y, p))\right]. \qquad \text{(B2)}
\end{aligned}$$

The rest of the proof is to show that

$$\begin{aligned}
&- C_1 \exp(-\tau_1 R)\text{(B0)} \leq \text{(B1)} \leq C_1 \exp(-\tau_1 R)\text{(B0)}, \\
&- C_2 \exp(-\tau_1 R)\text{(B0)} \leq \text{(B2)} \leq C_2 \exp(-\tau_1 R)\text{(B0)},
\end{aligned}$$

for some $C_1, C_2 > 0$ that do not depend on R. Then, by taking R large enough, we obtain the desired result.

First, we simplify (B0). Note that, for all $i, i_0 \in \mathcal{S}_X(p)$, we have that $\widehat{a}_i(X) = \widehat{a}_{i_0}(X)$. Hence, by switching the order of i, j, we obtain

$$\begin{aligned}
\sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\widehat{a}_i(X) - \widehat{a}_j(X)) h_{i,j}(X, y, p) &= \sum_{j \in \mathcal{S}_X^2(p)} (\widehat{a}_{i_0}(X) - \widehat{a}_j(X)) \sum_{i \in \mathcal{S}_X(p)} h_{i,j}(X, y, p) \\
&= g(X, y) \sum_{j \in \mathcal{S}_X^2(p)} (\widehat{a}_{i_0}(X) - \widehat{a}_j(X)) q_{i_0}(X) q_j(X) \sum_{i \in \mathcal{S}_X(p)} (\gamma_i(X, y) - \gamma_j(X, y)),
\end{aligned}$$

for any $i_0 \in \mathcal{S}_X(p)$. Since p is a locally optimal direction, we have

$$\sum_{i\in\mathcal{S}_X(p)} (\gamma_i(X,y) - \gamma_j(X,y)) \geq \mu, \qquad \text{for all } j \in \mathcal{S}_X^2(p).$$

Now, we compare (B1) and (B0). By the exact same reason above, we can rewrite

$$\sum_{i\in\mathcal{S}_X(p)} \sum_{j\in\mathcal{S}_X^<(p)} (\widehat{a}_i(X) - \widehat{a}_j(X))h_{i,j}(X,y,p)$$
$$= g(X,y) \sum_{j\in\mathcal{S}_X^<(p)} (\widehat{a}_{i_0}(X) - \widehat{a}_j(X))q_{i_0}(X)q_j(X) \sum_{i\in\mathcal{S}_X(p)} (\gamma_i(X,y) - \gamma_j(X,y)),$$

for any $i_0 \in \mathcal{S}_X(p)$, and we compare to (B0) term-by-term. Namely, for any X, $j \in \mathcal{S}_X^2(p)$ and $k \in \mathcal{S}_X^<(p)$, we have:

$$\frac{|\widehat{a}_{i_0}(X) - \widehat{a}_k(X)|}{\widehat{a}_{i_0}(X) - \widehat{a}_j(X)} \leq \frac{\|\hat{p}\|_2 \|E_{x_{i_0}} - E_{x_j}\|_2}{\tau_2} \leq \frac{2\|\hat{p}\|_2 \max_s \|E_s\|_2}{\tau_2} := C_3, \tag{22}$$

$$\frac{q_k(X)}{q_j(X)} = \exp(a_k(X) - a_j(X)) \leq \exp(-\tau_1 R), \tag{23}$$

$$\frac{\sum_{i\in\mathcal{S}_X(p)} |\gamma_i(X,y) - \gamma_k(X,y)|}{\sum_{i\in\mathcal{S}_X(p)} (\gamma_i(X,y) - \gamma_j(X,y))} \leq \frac{2T \max_s |\gamma_s|}{\mu} \leq \frac{2T \max_s \|E_s\|_2}{\mu} := C_4, \tag{24}$$

which implies that, for any X, $j \in \mathcal{S}_X^2(p)$ and $k \in \mathcal{S}_X^<(p)$,

$$|\widehat{a}_{i_0}(X) - \widehat{a}_k(X)|q_{i_0}(X)q_k(X) \sum_{i\in\mathcal{S}_X(p)} |\gamma_i(X,y) - \gamma_k(X,y)|$$
$$\leq \exp(-\tau_1 R) C_3 C_4 (\widehat{a}_{i_0}(X) - \widehat{a}_j(X))q_{i_0}(X)q_j(X) \sum_{i\in\mathcal{S}_X(p)} (\gamma_i(X,y) - \gamma_j(X,y)).$$

Thus, we get that:

$$|(\text{B1})| \leq \exp(-\tau_1 R) T C_3 C_4 |(\text{B0})|.$$

Next, we compare (B2) and (B0). Take any $i' \in X \setminus \mathcal{S}_X(p), k > i' \in X \setminus \mathcal{S}_X(p), i_0 \in \mathcal{S}_X(p), j \in \mathcal{S}_X^2(p)$. We compare

$$(\widehat{a}_{i'}(X) - \widehat{a}_k(X))h_{i',k}(X,y,p)$$

with each term in (B1). We note that the bounds on $\frac{\widehat{a}_{i'}(X) - \widehat{a}_k(X)}{\widehat{a}_{i_0}(X) - \widehat{a}_j(X)}$ and $\frac{|\gamma_{i'}(X) - \gamma_k(X)|}{\sum_{i\in\mathcal{S}_X(p)} (\gamma_i(X,y) - \gamma_j(X,y))}$ are the same as those in (22) and (24). Furthermore,

$$\frac{q_{i'}q_k}{q_{i_0}q_j} \leq \exp(-\tau_1 R),$$

which gives that

$$|(\text{B2})| \leq T^2 \exp(-\tau_1 R) C_3 C_4 |(\text{B0})|,$$

thus concluding the proof.

■

Lemma 18 *Under Assumption 4, for any $\delta > 0$, by picking*

$$\eta_0 \geq 4n^2T^2, \quad d \geq \max\left\{256, \left(2\log\frac{|\mathcal{S}|^2}{\delta}\right)^2, (88\eta_0^2 + 111\eta_0 + 2)^8\right\},$$

with probability $\geq 1-\delta$ over the initialization, if $p_\infty = \lim_{t\to\infty}\frac{p_t}{\|p_t\|_2}$ exists, then $p_\infty \in \mathcal{P}_*(p_\infty)$.

Proof We prove the lemma by contradiction. We first assume that there exists p_∞ such that $p_\infty \notin \mathcal{P}_*(p_\infty)$ and $p_\infty = \lim_{t\to\infty}\frac{p_t}{\|p_t\|_2}$. Then, we show that there exists $\widehat{p} \in \mathcal{P}_*(p_\infty)$ such that, for any $\epsilon > 0$, there is $\bar{t}(\epsilon)$ ensuring

$$-\frac{\widehat{p}^\top}{\|\widehat{p}\|_2}\nabla_p\mathcal{L}(\boldsymbol{E}^1, p_t) \geq -(1-\epsilon)\frac{p_t^\top}{\|p_t\|_2}\nabla_p\mathcal{L}(\boldsymbol{E}^1, p_t), \qquad \text{for all } t \geq \bar{t}(\epsilon).$$

As a consequence, by Lemma 9, we have that $p_\infty = \frac{\widehat{p}}{\|\widehat{p}\|_2}$, which gives a contradiction.

For the rest of the proof, we fix any $\epsilon > 0$ and denote $R = \|p_t\|_2$. We define $\overline{p_t} = \frac{p_t\|\widehat{p}\|_2}{\|p_t\|_2}$, and we equivalently show that:

$$-\widehat{p}^\top\nabla_p\mathcal{L}(\boldsymbol{E}^1, p_t) \geq -(1-\epsilon)\overline{p_t}\nabla_p\mathcal{L}(\boldsymbol{E}^1, p_t). \tag{25}$$

To prove this, we first note that since $p_\infty \notin \mathcal{P}_*(p_\infty)$, for all $\frac{\widehat{p}}{\|\widehat{p}\|_2} \in \mathcal{P}_*(p_\infty)$, there exists τ_0 independent of R such that

$$\|\widehat{p} - p_\infty\|\widehat{p}\|_2\|_2 \geq \tau_0.$$

Thus, by the definition of $\mathcal{P}_*(p_\infty)$, there exists $\mathcal{X}_0 \subseteq \mathcal{X}_n$ such that for each sequence $X \in \mathcal{X}_0$, we can find a pair of indices (i,j) with $i \in \mathcal{S}_X(p_\infty), j \in X \setminus \mathcal{S}_X(p_\infty)$ violating the margin, i.e.,

$$(\|\widehat{p}\|_2 p_\infty)^\top(E_{x_i} - E_{x_j}) \leq 1 - 3\tau,$$

for some $\tau < \frac{1}{6}$ that does not depend on R. With a slight abuse of notation, we define τ as

$$\begin{aligned}\tau = \frac{1}{3}\min\{&\min_{X\in\mathcal{X}_0}\{1 - (\|\widehat{p}\|_2 p_\infty)^\top(E_{x_i} - E_{x_j}), i \in \mathcal{S}_X(p_\infty), j \in \mathcal{S}_X^2(p_\infty)\},\\ &\min_{X\in\mathcal{X}_n}\{(\|\widehat{p}\|_2 p_\infty)^\top(E_{x_i} - E_{x_j}), i \in \mathcal{S}_X(p_\infty), j \in \mathcal{S}_X^2(p_\infty)\},\\ &\min_{X\in\mathcal{X}_n}\{(\|\widehat{p}\|_2 p_\infty)^\top(E_{x_i} - E_{x_j}), i \in \mathcal{S}_X^2(p_\infty), j \in \mathcal{S}_X^<(p_\infty)\}\}.\end{aligned}$$

This means that, for all $X \in \mathcal{X}_n$ and for all (i,j) pairs such that $i \in \mathcal{S}_X(p_\infty), j \in \mathcal{S}_X^2(p_\infty)$, we have

$$(\|\widehat{p}\|_2 p_\infty)^\top(E_{x_i} - E_{x_j}) \geq 3\tau;$$

for all pairs (i,j) such that $i \in \mathcal{S}_X^2(p_\infty), j \in \mathcal{S}_X^<(p_\infty)$, we have

$$(\|\widehat{p}\|_2 p_\infty)^\top(E_{x_i} - E_{x_j}) \geq 3\tau;$$

and for all $X \in \mathcal{X}_0, i \in \mathcal{S}_X(p_\infty), j \in \mathcal{S}_X^2(p_\infty)$, we have

$$(\|\widehat{p}\|_2 p_\infty)^\top(E_{x_i} - E_{x_j}) \leq 1 - 3\tau,$$

with some τ that does not depend on R.

Now, we compute the overlap with $\overline{p_t}$. For all X and (i,j), we have

$$\overline{p_t}^\top (E_{x_i} - E_{x_j}) = (\|\widehat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j}) + (\overline{p_t} - \|\widehat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j}).$$

We upper bound

$$|(\overline{p_t} - \|\widehat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j})| \leq \|\widehat{p}\|_2 \left\| \frac{p_t}{\|p_t\|_2} - p_\infty \right\|_2 \|E_{x_1} - E_{x_2}\|_2,$$

and since $\|\widehat{p}\|_2, \|E_{x_1} - E_{x_2}\|_2$ are finite, we have

$$\lim_{t\to\infty} |(\overline{p_t} - \|\widehat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j})| = 0.$$

Thus, we can pick t_1, such that for $t \geq t_1$, we have

$$|(\overline{p_t} - \|\widehat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j})| \leq \tau,$$

which implies that, for all $X \in \mathcal{X}_n$ and for all (i,j) pairs such that $i \in \mathcal{S}_X(p_\infty), j \in \mathcal{S}_X^2(p_\infty)$, we have

$$\overline{p_t}^\top (E_{x_i} - E_{x_j}) \geq \tau;$$

for all (i,j) pairs such that $i \in \mathcal{S}_X^2(p_\infty), j \in \mathcal{S}_X^<(p_\infty)$, we have

$$\overline{p_t}^\top (E_{x_i} - E_{x_j}) \geq \tau;$$

and for all $X \in \mathcal{X}_0, i \in \mathcal{S}_X(p_\infty), j \in \mathcal{S}_X^2(p_\infty)$, we have:

$$\overline{p_t}^\top (E_{x_i} - E_{x_j}) \leq 1 - \tau,$$

for some τ that does not depend on R.

Next, we show that $\overline{p_t}$ is a locally optimal solution as per Definition 16. By Lemma 13, p_∞ selects all the completely positive/negative tokens. Thus, as $\overline{p_t} \approxeq p_\infty$, $\overline{p_t}$ also selects such tokens, the rest being irrelevant by Assumption 4. Hence, for any pair (X,y) and for any $j \in X \setminus \mathcal{S}_X(p_t)$, we have:

$$\sum_{i \in \mathcal{S}_X(p_t)} (\gamma_i(X,y) - \gamma_j(X,y)) \geq \frac{\eta_0}{4nT},$$

by picking d large enough (as per the hypothesis of the lemma). By construction, $\hat{p} \approxeq p_t$, $\|\hat{p}\|_2$ does not depends on R and, moreover, for any X,

$$\hat{p}^\top (E_{x_i} - E_{x_j}) \geq 1, \qquad \text{for all } i \in \mathcal{S}_X(\hat{p}),\ j \in X \setminus \mathcal{S}_X(\hat{p}).$$

By applying Lemma 17 on both $\hat{p}$ and $\overline{p_t}$, we have that for any $\epsilon_1 > 0$ there exist t_2 s.t. for all $t \geq \max\{t_1, t_2\}$, we have

$$-\hat{p}^\top \nabla_p \mathcal{L}(\boldsymbol{E}^1, p_t) \geq (1-\epsilon_1)\widehat{\mathbb{E}} \left[\sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\widehat{a}_i(X) - \widehat{a}_j(X)) h_{i,j}(X,y,p_t) \right],$$

$$-\overline{p_t}^\top \nabla_p \mathcal{L}(\boldsymbol{E}^1, p_t) \leq (1+\epsilon_1)\widehat{\mathbb{E}} \left[\sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\overline{a_i}(X) - \overline{a_j}(X)) h_{i,j}(X,y,p_t) \right],$$

where $\overline{a_i}(X), \overline{a_j}(X)$ are defined analogously to $\widehat{a}_i(X), \widehat{a}_j(X)$ by replacing $\hat{p}$ with $\overline{p_t}$.

Now, we further show that, for any $\epsilon_2 > 0$, there exist t_3 such that for all $t \geq t_3$,

$$\widehat{\mathbb{E}}\left[\sum_{i\in\mathcal{S}_X(p)}\sum_{j\in\mathcal{S}_X^2(p)}(\overline{a_i}(X)-\overline{a_j}(X))h_{i,j}(X,y,p_t)\right] \leq (1+\epsilon_2)\widehat{\mathbb{E}}_{X\in\mathcal{X}_0}\left[\sum_{i\in\mathcal{S}_X(p)}\sum_{j\in\mathcal{S}_X^2(p)}(\overline{a_i}(X)-\overline{a_j}(X))h_{i,j}(X,y,p_t)\right].$$

To see this, we use the same idea as in the proof of Lemma 17. We can write

$$\widehat{\mathbb{E}}\left[\sum_{i\in\mathcal{S}_X(p)}\sum_{j\in\mathcal{S}_X^2(p)}(\overline{a_i}(X)-\overline{a_j}(X))h_{i,j}(X,y,p_t)\right]$$

$$= \widehat{\mathbb{E}}_{X\in\mathcal{X}_0}\left[\sum_{i\in\mathcal{S}_X(p)}\sum_{j\in\mathcal{S}_X^2(p)}(\overline{a_i}(X)-\overline{a_j}(X))h_{i,j}(X,y,p_t)\right] \tag{A0}$$

$$+ \widehat{\mathbb{E}}_{X'\in\mathcal{X}_n\setminus\mathcal{X}_0}\left[\sum_{i\in\mathcal{S}_{X'}(p)}\sum_{j\in\mathcal{S}_{X'}^2(p)}(\overline{a_i}(X')-\overline{a_j}(X'))h_{i,j}(X',y',p_t)\right], \tag{A1}$$

and it is sufficient to show that

$$\text{(A1)} \leq \epsilon_2\text{(A0)}.$$

To prove this, we compare term-by-term. Let $X \in \mathcal{X}_0, X' \in \mathcal{X}_n \setminus \mathcal{X}_0, j \in \mathcal{S}_X^2(p_t), j' \in \mathcal{S}_X^2(p_t)$, and recall that:

$$\sum_{i\in\mathcal{S}_X(p_t)}(\overline{a_i}(X)-\overline{a_j}(X))h_{i,j}(X,y,p_t) = g(X,y)(\overline{a_{i_0}}(X)-\overline{a_j}(X))q_{i_0}(X)q_j(X)\sum_{i\in\mathcal{S}_X(p_t)}(\gamma_i(X,y)-\gamma_j(X,y)),$$

$$\sum_{i\in\mathcal{S}_{X'}(p_t)}(\overline{a_i}(X')-\overline{a_{j'}}(X'))h_{i,j'}(X',y',p_t) = g(X',y')(\overline{a_{i_1}}(X')-\overline{a_{j'}}(X'))q_{i_1}(X')q_{j'}(X')\sum_{i\in\mathcal{S}_{X'}(p_t)}(\gamma_i(X',y')-\gamma_{j'}(X',y')),$$

for any $i_0 \in \mathcal{S}_X(p_t), i_1 \in \mathcal{S}_{X'}(p_t)$. Note that

$$\frac{g(X',y')}{g(X,y)} \leq \frac{\max_{X,y} g(X,y)}{\min_{X,y} g(X,y)} \leq \max_{X,y}(1+\exp(yf(X))) \leq (1+\exp(\eta_0)) := C_5.$$

By using the same argument as in (22) and (24), we have

$$\frac{\overline{a_{i_1}}(X')-\overline{a_{j'}}(X')}{\overline{a_{i_0}}(X)-\overline{a_j}(X)} \leq C_3,$$
$$\frac{\sum_{i\in\mathcal{S}_{X'}(p_t)}(\gamma_i(X',y')-\gamma_{j'}(X',y'))}{\sum_{i\in\mathcal{S}_X(p_t)}(\gamma_i(X,y)-\gamma_j(X,y))} \leq C_4.$$

Finally, we need to upper bound:

$$\frac{q_{i_1}(X')q_{j'}(X')}{q_{i_0}(X)q_j(X)}.$$

We note that

$$\begin{aligned} a_{i_1}(X') - a_{j'}(X') &\geq R/\|\hat{p}\|_2, \\ a_{i_0}(X) - a_j(X) &\leq (1-\tau)R/\|\hat{p}\|_2, \end{aligned}$$

where $a_i(X) = p_t^\top E_{x_i}$. Thus by Lemma 10, we have:

$$q_{i_0}(X) \geq \frac{1}{T}, \quad q_j(X) \geq \frac{1}{T\exp((1-\tau)R/\|\hat{p}\|_2)}, \quad q_{i_1}(X') \leq 1, \quad q_{j'}(X') \leq \frac{1}{\exp(R/\|\hat{p}\|_2)},$$

which implies that

$$\frac{q_{i_1}(X')q_{j'}(X')}{q_{i_0}(X)q_j(X)} \leq T^2\exp(-\tau R/\|\hat{p}\|_2).$$

Thus, for each $X \in \mathcal{X}_0, X' \in \mathcal{X}_n \setminus \mathcal{X}_0, j \in \mathcal{S}_X^2(p_t), j' \in \mathcal{S}_X^2(p_t)$, we have

$$\sum_{i\in\mathcal{S}_{X'}(p)} (\overline{a_i}(X') - \overline{a_{j'}}(X'))h_{i,j'}(X', y', p_t) \leq C_6\exp(-\tau R/\|\hat{p}\|_2) \sum_{i\in\mathcal{S}_X(p)} (\overline{a_i}(X) - \overline{a_j}(X))h_{i,j}(X, y, p_t).$$

Thus by picking large enough t_3 which gives large enough R, we have:

$$\text{(A1)} \leq \epsilon_2\text{(A0)}.$$

This allows us to conclude that

$$\begin{aligned} -\hat{p}^\top\nabla_p\mathcal{L}(\boldsymbol{E}^1, p_t) &\geq (1-\epsilon_1)\widehat{\mathbb{E}}\left[\sum_{i\in\mathcal{S}_X(p)}\sum_{j\in\mathcal{S}_X^2(p)} (\widehat{a}_i(X) - \widehat{a}_j(X))h_{i,j}(X, y, p_t)\right] \\ &\geq (1-\epsilon_1)\widehat{\mathbb{E}}_{X\in\mathcal{X}_0}\left[\sum_{i\in\mathcal{S}_X(p)}\sum_{j\in\mathcal{S}_X^2(p)} (\widehat{a}_i(X) - \widehat{a}_j(X))h_{i,j}(X, y, p_t)\right], \\ -\overline{p_t}^\top\nabla_p\mathcal{L}(\boldsymbol{E}^1, p_t) &\leq (1+\epsilon_1)(1+\epsilon_2)\widehat{\mathbb{E}}_{X\in\mathcal{X}_0}\left[\sum_{i\in\mathcal{S}_X(p)}\sum_{j\in\mathcal{S}_X^2(p)} (\overline{a_i}(X) - \overline{a_j}(X))h_{i,j}(X, y, p_t)\right]. \end{aligned}$$

Note that, for each $X \in \mathcal{X}_0$,

$$\widehat{a}_i(X) - \widehat{a}_j(X) \geq 1, \qquad \overline{a_i}(X) - \overline{a_j}(X) \leq 1-\tau,$$

which gives that

$$-\hat{p}^\top\nabla_p\mathcal{L}(\boldsymbol{E}^1, p_t) \geq -\frac{1-\epsilon_1}{(1+\epsilon_1)(1+\epsilon_2)(1-\tau)}\overline{p_t}^\top\nabla_p\mathcal{L}(\boldsymbol{E}^1, p_t).$$

Since ϵ_1, ϵ_2 can be arbitrarily small, the proof is complete. ■

D.4. Proof of Lemma 6

Proof If $\hat{p}$ selects all the tokens, then $\hat{p}^\top E_{x_i} = \hat{p}^\top E_{x_j}$ for all $x_i, x_j \in X$ and for all $X \in \mathcal{X}_n$. Thus, by Lemma 8, $\hat{p}^\top \nabla_p \mathcal{L}(\boldsymbol{E}, p) = 0$ for any p, which gives the desired result. ■

D.5. Statement of Lemma 19

Lemma 19 *Under Assumption 4, for any $\delta > 0$, assume that*

$$\eta_0 \geq 4n^2T^2, \quad d \geq \max\left\{256, \left(2\log\frac{|\mathcal{S}|^2}{\delta}\right)^2, (88\eta_0^2 + 111\eta_0 + 2)^8, |\mathcal{S}| + 3\right\}. \tag{26}$$

holds. Let $\hat{p}$ be the solution of the max-margin problem (9) that only selects the completely positive/negative tokens, i.e.,

$$\hat{p} = \arg\min_p \|p\|_2, \qquad \text{s.t.} \quad p^\top (E_{s_*^X} - E_s) \geq 1, \qquad \forall s \in X \setminus \{s_*^X\},\ \forall X \in \mathcal{X}_n,$$

where s_*^X denotes the unique completely positive/negative token in the sequence X. Assume that $p_\infty := \lim_{t\to\infty} \frac{p_t}{\|p_t\|_2}$ exists and that, for any $\hat{p}'$ solving (9) with a different selection, $\|\hat{p}\|_2 < (1-\mu)\|\hat{p}'\|_2$ for some constant μ that does not depend d. Then, by taking $d \geq \left(\frac{8T(1-\mu)}{\mu\eta_0}\right)^4$, we have that $p_\infty = \frac{\hat{p}}{\|\hat{p}\|_2}$ with probability at least $1-\delta$.

The sufficient condition of the result above requires the max-margin direction that does not select irrelevant tokens to have a larger margin than any other max-margin solution associated to a different token selection. We expect this to be the case e.g. for datasets where all the completely positive/negative tokens have the same α_s. In fact, given the structure of the context embeddings in (5), the max-margin solution $\hat{p}$ is expected to satisfy $\hat{p}^\top v \approx 0, \hat{p}^\top E_s \approx 1, \hat{p}^\top E_{s'} \approx 0$ for all $s \in \mathcal{S}_X(\hat{p})$ and $s' \in \overline{\mathcal{S}_X(\hat{p})}$. Since the token embeddings at initialization are approximately orthogonal to each other, $\hat{p} \approx \sum_{s\in\mathcal{S}_X(p)} E_s^0$, meaning that $\|\hat{p}\|_2 \approx \sqrt{|\mathcal{S}_X(p)|}$, which implies that the sufficient condition holds.

D.6. Proof of Lemma 19

Proof Let $\hat{p}'$ be the max-margin solution of (9) with a different selection. By Theorem 5, we have that, for all $X, s_*^X \in \mathcal{S}_X(\hat{p}')$. We denote by i_*^X the index of s_*^X. Assume by contradiction $p_\infty = \frac{\hat{p}'}{\|\hat{p}'\|_2}$. We will now show that this implies the following statement: for any $\epsilon > 0$, there is a $t(\epsilon)$ ensuring

$$-\frac{\hat{p}^\top}{\|\hat{p}\|_2}\nabla_p\mathcal{L}(\boldsymbol{E}, p_t) \geq -(1-\epsilon)\frac{p_t^\top}{\|p_t\|_2}\nabla_p\mathcal{L}(\boldsymbol{E}, p_t), \qquad \text{for all } t \geq t(\epsilon). \tag{27}$$

Then, by Lemma 9, we have that $p_\infty = \frac{\hat{p}}{\|\hat{p}\|_2}$, which gives a contradiction.

As in the proof of Lemma 18, we define $\overline{p_t} = \frac{p_t}{\|p_t\|_2}\|\hat{p}\|_2$. Thus, (27) is equivalent to

$$-\hat{p}^\top\nabla_p\mathcal{L}(\boldsymbol{E}, p_t) \geq -(1-\epsilon)\overline{p_t}^\top\nabla_p\mathcal{L}(\boldsymbol{E}, p_t).$$

First of all, since $\hat{p}, \hat{p}'$ are two max-margin solutions, by Lemma 3 we have that all the constraints are tight, that is:

$$\hat{p}^\top (E_{s_*^X} - E_s) = 1, \qquad \forall s \in X \setminus s_*^X,\ \forall X \in \mathcal{X}_n,$$
$$(\hat{p}')^\top (E_s - E_{s'}) = 1, \qquad \forall s \in \mathcal{S}_X(\hat{p}'),\ \forall s' \in X \setminus \mathcal{S}_X(\hat{p}'),\ \forall X \in \mathcal{X}_n,$$

which implies that

$$\frac{\hat{p}'^\top \|\hat{p}\|_2}{\|\hat{p}'\|_2}(E_s - E_{s'}) = \frac{\|\hat{p}\|_2}{\|\hat{p}'\|_2} = 1 - \mu < 1, \qquad \forall s \in \mathcal{S}_X(\hat{p}'),\ \forall s' \in X \setminus \mathcal{S}_X(\hat{p}'),\ \forall X \in \mathcal{X}_n.$$

As $\|\overline{p_t}\|_2 = \|\hat{p}\|_2 < \|\hat{p}'\|_2$, $\overline{p_t}$ violates the max-margin condition. Moreover, as $\lim_{t\to\infty} \overline{p_t} = \frac{\hat{p}'^\top \|\hat{p}\|_2}{\|\hat{p}'\|_2}$, for any $\epsilon_1 \in (0, \mu)$, there exists a t_1 ensuring the following for all $t \geq t_1$: for all (i, j) pairs such that $i \in \mathcal{S}_X(p_\infty), j \in \mathcal{S}_X^2(p_\infty)$, we have

$$\overline{p_t}^\top (E_{x_i} - E_{x_j}) \leq 1 - \mu + \epsilon_1 \leq 1.$$

By applying Lemma 17 to $\overline{p_t}$, we obtain that, for any $\epsilon_2 > 0$, there exists a t_2 ensuring that, for all $t \geq t_2$,

$$\begin{aligned}
-\overline{p_t}^\top \nabla_p \mathcal{L}(\boldsymbol{E}^1, p_t) &\leq (1+\epsilon_2)\widehat{\mathbb{E}}\left[g(X,y) \sum_{i \in \mathcal{S}_X(p_t)} \sum_{j \in \mathcal{S}_X^2(p_t)} (\overline{a_i}(X) - \overline{a_j}(X) q_i(X) q_j(X)(\gamma_i(X) - \gamma_j(X)))\right] \\
&= (1+\epsilon_2)\widehat{\mathbb{E}}\left[g(X,y) \sum_{j \in \mathcal{S}_X^2(p_t)} (\overline{a_{i_*^X}}(X) - \overline{a_j}(X) q_{i_*^X}(X) q_j(X)(\gamma_{i_*^X}(X) - \gamma_j(X)))\right] \\
&\quad + (1+\epsilon_2)\widehat{\mathbb{E}}\left[g(X,y) \sum_{i \in \mathcal{S}_X(p_t), i \neq i_*^X} \sum_{j \in \mathcal{S}_X^2(p_t)} (\overline{a_i}(X) - \overline{a_j}(X) q_i(X) q_j(X)(\gamma_i(X) - \gamma_j(X)))\right].
\end{aligned}$$

We then compute by Lemma 8 that

$$\begin{aligned}
-\hat{p}^\top \nabla_p \mathcal{L}(\boldsymbol{E}^1, p_t) &= \widehat{\mathbb{E}}\left[g(X,y) \sum_{j \in X \setminus \mathcal{S}_X(p_t)} (\widehat{a_{i_*^X}}(X) - \widehat{a_j}(X)) q_{i_*^X}(X) q_j(X)(\gamma_{i_*^X}(X) - \gamma_j(X)))\right] \\
&\geq \widehat{\mathbb{E}}\left[g(X,y) \sum_{j \in \mathcal{S}_X^2(p_t)} (\widehat{a_{i_*^X}}(X) - \widehat{a_j}(X)) q_{i_*^X}(X) q_j(X)(\gamma_{i_*^X}(X) - \gamma_j(X)))\right] \\
&= (1-\mu)\widehat{\mathbb{E}}\left[g(X,y) \sum_{j \in \mathcal{S}_X^2(p_t)} (\widehat{a_{i_*^X}}(X) - \widehat{a_j}(X)) q_{i_*^X}(X) q_j(X)(\gamma_{i_*^X}(X) - \gamma_j(X)))\right] \\
&\quad + \mu\widehat{\mathbb{E}}\left[g(X,y) \sum_{j \in \mathcal{S}_X^2(p_t)} (\widehat{a_{i_*^X}}(X) - \widehat{a_j}(X)) q_{i_*^X}(X) q_j(X)(\gamma_{i_*^X}(X) - \gamma_j(X)))\right],
\end{aligned}$$

where in the first equality we use the fact that, for all $j, j' \neq i_*^X$, $\widehat{a_{j'}}(X) - \widehat{a_j}(X) = 0$, and in the second inequality we use the fact that all the terms in the summand are positive.

We note that:

$$(1+\epsilon_2)\widehat{\mathbb{E}}\left[g(X,y)\sum_{j\in\mathcal{S}_X^2(p_t)}(\overline{a_{i_*^X}}(X)-\overline{a_j}(X)q_{i_*^X}(X)q_j(X)(\gamma_{i_*^X}(X)-\gamma_j(X)))\right]$$
$$<(1-\mu)\widehat{\mathbb{E}}\left[g(X,y)\sum_{j\in\mathcal{S}_X^2(p_t)}(\widehat{a_{i_*^X}}(X)-\widehat{a_j}(X))q_{i_*^X}(X)q_j(X)(\gamma_{i_*^X}(X)-\gamma_j(X)))\right],$$

as

$$\widehat{a_{i_*^X}}(X)-\widehat{a_j}(X)=1,\qquad \overline{a_{i_*^X}}(X)-\overline{a_j}(X)\le 1-\mu-\epsilon_1.$$

It remains to show that

$$\mu\widehat{\mathbb{E}}\left[g(X,y)\sum_{j\in\mathcal{S}_X^2(p_t)}(\widehat{a_{i_*^X}}(X)-\widehat{a_j}(X))q_{i_*^X}(X)q_j(X)(\gamma_{i_*^X}(X)-\gamma_j(X)))\right]$$
$$\ge(1+\epsilon_2)\widehat{\mathbb{E}}\left[g(X,y)\sum_{i\in\mathcal{S}_X(p_t):i\neq i_*^X}\sum_{j\in\mathcal{S}_X^2(p_t)}(\overline{a_i}(X)-\overline{a_j}(X))q_i(X)q_j(X)(\gamma_i(X)-\gamma_j(X)))\right]. \tag{28}$$

We have that, for each $i\in\mathcal{S}_X(p_t):i\neq i_*^X, j\in\mathcal{S}_X^2(p_t)$,

$$|\gamma_i(X)-\gamma_j(X)|\le 2d^{-1/4},\qquad \gamma_{i_*^X}(X)-\gamma_j(X)\ge\frac{\eta_0}{4}.$$

As $d\ge\left(\frac{8T(1-\mu)}{\mu\eta_0}\right)^4$, (28) holds and the proof is complete. ■

Appendix E. Numerical experiments

To support our theoretical findings, we showcase the correlation of the embeddings with the $\langle\text{cls}\rangle$ embedding p and the output vector v, having trained *all* the parameters with gradient descent until convergence. We consider different datasets (synthetic data in Figure 2; IMDB/Yelp datasets in Figures 1 and 3) and different architectures (one-layer model (1) in Figures 2 and 3; two-layer model (32) in Figure 1). Taken together, the experiments display an excellent agreement with our theory going beyond the one-layer architecture (1) and also beyond the requirements on the data-generating process. Specifically, the trained embeddings capture the importance of the corresponding tokens: the dot-product with v is proportional to how frequently the token appears in positive sequences rather than in negative ones, and the dot-product with p is proportional to the modulus of such frequency. We detail below the experimental design.

Synthetic data. Let us define the data-generating process for the synthetic experiments in Figure 2. The data is generated according to a K-level model. Namely, the vocabulary set $\mathcal{S}$ is partitioned as

$$\mathcal{S}=\tilde{\mathcal{S}}\cup\left\{\mathcal{S}_k^{-1}\right\}_{k=1}^K\cup\left\{\mathcal{S}_k^{+1}\right\}_{k=1}^K. \tag{29}$$

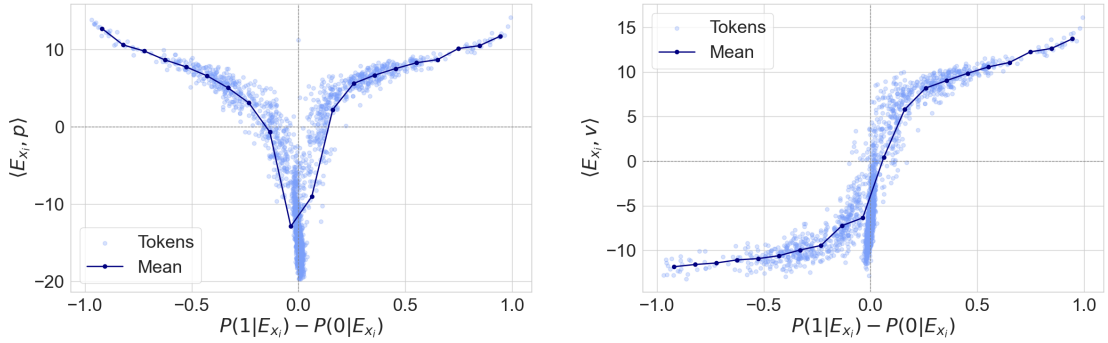

Figure 2: Dot-product of token embeddings with $\langle \text{cls} \rangle$ embedding p (left) and regression coefficients v (right), as a function of the token-wise difference in posterior probabilities for synthetic data sampled according to (30). We consider the one-layer attention model in (1) with all parameters trained until convergence. The point cloud around zero corresponds to the tokens in the irrelevant set.

Here, $\tilde{\mathcal{S}}$ contains *irrelevant* tokens appearing in both positive and negative contexts with equal probability, while $\mathcal{S}_k^{+1}$ and $\mathcal{S}_k^{-1}$ (for $k \in \{1, \dots, K\}$) contain tokens appearing *mostly* in positive and negative contexts, respectively. Formally, define the importance levels $\tilde{\delta}, \delta_1, \dots, \delta_K > 0$. Then, given the sequence label $y \in \{-1, +1\}$ and $s \in \mathcal{S}$, we sample the tokens from the vocabulary as

$$p(s|y) = \begin{cases} \frac{1-\tilde{\delta}}{|\tilde{\mathcal{S}}|}, & s \in \tilde{\mathcal{S}}, \\ \frac{\tilde{\delta}(1-\delta_k)}{\sum_{k=1}^K |\mathcal{S}_k^y|}, & s \in \mathcal{S}_k^y, \\ \frac{\tilde{\delta}\delta_k}{\sum_{k=1}^K |\mathcal{S}_k^{\neg y}|}, & s \in \mathcal{S}_k^{\neg y}, \end{cases} \tag{30}$$

where $\neg$ denotes the binary inversion, i.e., $\neg(+1) = -1$ and $\neg(-1) = +1$. The law (30) implies the following posterior distribution:

$$p(y|s) = \begin{cases} 1/2, & s \in \tilde{\mathcal{S}}, \\ 1 - \delta_k, & s \in \mathcal{S}_k^y, \\ \delta_k, & s \in \mathcal{S}_k^{\neg y}. \end{cases} \tag{31}$$

From (31), it is clear that *(i)* $\tilde{\mathcal{S}}$ contains *irrelevant* tokens as the posterior is uniform, and *(ii)* δ_k quantifies the importance of the tokens in $\mathcal{S}_k^{\pm 1}$ by skewing the posterior to be $(\delta_k, 1 - \delta_k)$. For the experiments in Figure 2, we select the following hyper-parameters: $|\mathcal{S}| = 2048$, $K = 8$ and sequence length $T = 256$; $|\mathcal{S}_k^{+1}| = |\mathcal{S}_k^{-1}|$ with $|\mathcal{S}_k^{+1}| = 4 + 2^{k-1}$, and $|\tilde{\mathcal{S}}| = 964$; $\tilde{\delta} = 0.05$ and $\{\delta_k\}_{k=1}^K = \{0.45, 0.35, 0.3, 0.25, 0.2, 0.1, 0.05, 0.02\}$.

Figure 2 shows a clear separation between positive and negative tokens (right plot with the dot-product $\langle E_{x_i}, v \rangle$), and the selection mechanism ($\langle \text{cls} \rangle$ token) assigns high weights to tokens that have larger importance δ_k (left plot with the dot-product $\langle E_{x_i}, p \rangle$).

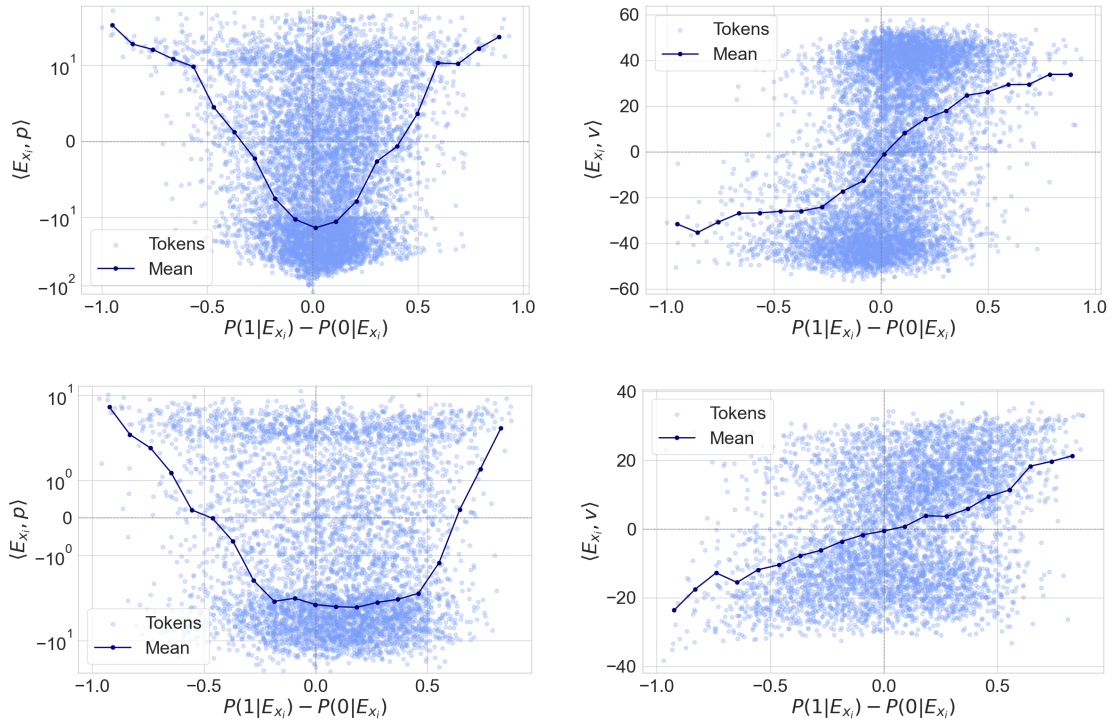

Figure 3: Dot-product of token embeddings with $\langle \text{cls} \rangle$ embedding p and regression coefficients v, as a function of the token-wise difference in posterior for IMDB dataset (top row) and Yelp dataset (bottom row). We consider the one-layer attention model in (1) with all parameters trained until convergence.

IMDB and Yelp datasets. The IMDB dataset[1] consists of 50000 reviews of average length 239 words per review, associated to either a positive or a negative sentiment. Yelp reviews[2] provide a much larger selection. To align the data size and sequence length with the IMDB dataset, we randomly subsample a portion of the Yelp dataset constrained on the sequence length, i.e., we select reviews which have at least 1000 and not more than 1500 characters. In addition, Yelp reviews provide a five-star ranking, which we convert to the binary sentiment based on the following rule: 1/2 stars reviews are assigned label -1; 4/5 star reviews are assigned label $+1$; neutral reviews, i.e., 3-star score, are removed. We adhere to a typical preprocessing pipeline for both datasets: we start by cleaning the data from punctuation symbols and omitting the stop-words, followed by an application of stemming; and we use the Bert tokenizer from Hugging Face[3] to tokenize sequences. Tokens that appear less than 50 times are purged.

The numerical simulations for both datasets are reported in Figure 3, which displays a phenomenology similar to that obtained for synthetic data in Figure 2, thus providing additional grounding for our theoretical claims.

1. https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
2. https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset
3. https://huggingface.co/google-bert/bert-base-uncased

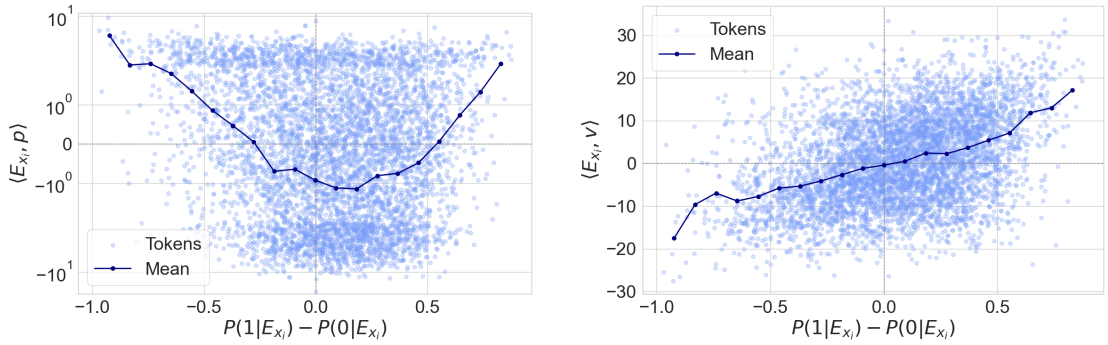

Figure 4: Dot-product of token embeddings with $\langle \text{cls} \rangle$ embedding p (left) and regression coefficients v (right), as a function of the token-wise difference in posterior probabilities, for the two-layer attention model (32) trained on the Yelp dataset.

Two-layer model. We also consider the following two-layer model:

$$\boldsymbol{E}'_X = \text{LayerNorm}(\texttt{Softmax}(\boldsymbol{E}_X \boldsymbol{E}_X^\top)\boldsymbol{E}_X + \boldsymbol{E}_X), \quad f(X; p, \boldsymbol{E}) = \texttt{Softmax}(p^\top (\boldsymbol{E}'_X)^\top)\boldsymbol{E}'_X v, \tag{32}$$

which includes both a skip connection and the layer-norm. We note that, for both IMDB and Yelp data, the model in (32) achieves significantly smaller loss values at convergence (of the order of 10^{-5}, in contrast to the order of 10^{-1} achieved by the model in (1)). However, even if this model is more complex than the one analyzed in Section 3, the results in Figure 1 are still remarkably similar to those in Figures 2 and 3.

We note that all plots consider on the x-axis the difference in posterior probabilities

$$p(1|\boldsymbol{E}_{x_i}) - p(0|\boldsymbol{E}_{x_i}) = \frac{\sum_{(X,y)\in\mathcal{D}} y \sum_{i=1}^T \mathbb{1}_{x_i=s}}{\sum_{(X,y)\in\mathcal{D}} \sum_{i=1}^T \mathbb{1}_{x_i=s}} \tag{33}$$

in place of the quantity α_s defined in (3). In fact, while the quantity in (3) appears naturally from the analysis of gradient descent, the difference in posterior probabilities provides better visuals for real data (IMDB and Yelp). The difference between (3) and (33) lies in the normalization used: the posterior difference in (33) is the discrepancy between counts of the token x_i in positive and negative sentences normalized by the total number of occurrences of x_i, while the quantity in (33) normalizes the discrepancy by the total number of tokens nT in the datasets. For synthetic data sampled according to (30), due to the uniform nature of the sampling procedure, all tokens appear the same number of times. Thus, both quantities are the same up to a fixed scaling and, thus, they are equivalent.

Additional details on hyperparameters. For all numerical simulations, we use the AdamW optimizer from `torch.optim`, and we reduce the learning rate in a multiplicative fashion by a factor $\gamma = 0.1$ at epochs 100 and 200, i.e.,

$$\text{LR}_{\text{new}} = \text{LR}_{\text{old}} \cdot \gamma.$$

We adhere to the batch size of 128 and fix the embedding dimension to 2048. In the experiments on IMDB and Yelp datasets, the hyperparameters *do not* differ between the two-layer model and the one-layer model. We set the number of training epochs to 500, the learning rate to 0.01, and the weight decay to 10^{-8}. In the experiments on synthetic data, we set the number of training epochs to 196, the learning rate to 10^{-4}, and the weight decay to 10^{-4}.