Towards Understanding the Word Sensitivity of Attention Layers: A Study via Random Features

Simone Bombari [1] **Marco Mondelli** [1]

Abstract

Understanding the reasons behind the exceptional success of transformers requires a better analysis of why attention layers are suitable for NLP tasks. In particular, such tasks require predictive models to capture contextual meaning which often depends on one or few words, even if the sentence is long. Our work studies this key property, dubbed *word sensitivity* (WS), in the prototypical setting of random features. We show that attention layers enjoy high WS, namely, there exists a vector in the space of embeddings that largely perturbs the random attention features map. The argument critically exploits the role of the $\mathrm{softmax}$ in the attention layer, highlighting its benefit compared to other activations (e.g., ReLU). In contrast, the WS of standard random features is of order $1/\sqrt{n}$, n being the number of words in the textual sample, and thus it decays with the length of the context. We then translate these results on the word sensitivity into generalization bounds: due to their low WS, random features provably cannot learn to distinguish between two sentences that differ only in a single word; in contrast, due to their high WS, random attention features have higher generalization capabilities. We validate our theoretical results with experimental evidence over the BERT-Base word embeddings of the imdb review dataset.

1. Introduction

Deep learning theory has provided a quantitative description of phenomena routinely occuring in state-of-the-art models, such as double-descent (Nakkiran et al., 2020; Mei & Montanari, 2022), benign overfitting (Bartlett et al., 2020; Belkin, 2021), and feature learning (Ba et al., 2022; Damian et al., 2022). However, most existing works focus on architectures given by the composition of matrix multiplications and non-linearities, which model e.g. fully connected and convolutional layers. In contrast, the recent impressive results achieved by large language models (Brown et al., 2020; Bubeck et al., 2023) are largely attributed to the introduction of transformers (Vaswani et al., 2017), which are in turn based on attention layers (Bahdanau et al., 2015; Kim et al., 2017). Hence, isolating the unique features of the attention mechanism stands out as a critical challenge to understand the success of transformers, thus paving the way to the principled design of large language models.

Recent work tackling this problem characterizes the sample complexities required by simplified attention models (Edelman et al., 2022; Fu et al., 2023). However, learning is limited to a specific set of targets, such as sparse functions (Edelman et al., 2022) or functions of the correlations between the first query token and key tokens (Fu et al., 2023). This paper takes a different perspective and starts from the simple empirical observation that *one or few words can change the meaning of a sentence*. Think, for example, to the pair of sentences "I love her much" and "I love her smile", where replacing a single word alters the meaning of the text. This is captured by the BERT-Base model (Devlin et al., 2019), which shows a different attention score pattern over these two sentences (see Figure 1, left). Another example can be found in the table on the right of Figure 1, where changing one word in the prompt would require a well-behaved model (in this case, Llama2-7b (Touvron et al., 2023)) to modify its output. In general, language models need to have a high *word sensitivity* to capture semantic changes when just a single word is modified in the context, which motivates the following question:

> *Do attention layers have a larger word sensitivity than fully connected architectures?*

To formalize the problem, we represent the textual data with $X \in \mathbb{R}^{n \times d}$, where the n rows represent the word embeddings in $\mathbb{R}^d$. Then, we define the word sensitivity (WS) in (4), as a measure of how changing a row in X modifies the embedding of a given feature map. We focus our study on the prototypical setting of *random features*, i.e., where the weights of the layers are random. In particular, we consider *(i)* the *random features* (RF) map (Rahimi & Recht, 2007), defined in (1), and *(ii)* the *random attention*

---

[1]Institute of Science and Technology, Austria. Correspondence to: Simone Bombari <simone.bombari@ista.ac.at>.

Prompt	Output
Reply with "Yes" if the review I will provide you is positive, and "No" otherwise. Review: Sorry, gave it a 1, which is the rating I give to movies on which I walk out or fall asleep.	No
Reply with "Yes" if the review I will provide you is negative, and "No" otherwise. Review: Sorry, gave it a 1, which is the rating I give to movies on which I walk out or fall asleep.	Yes

Figure 1. Left. Average attention scores for the word embeddings of the two sentences "I love her much" and "I love her smile". The embeddings are computed with the BERT-Base model, the scores are averaged over the 12 heads and displayed without the `[CLS]` token. *Right.* Output of the Llama2-7b-chat model for two prompts differing only in a single word.

features (RAF) map (Fu et al., 2023), defined in (3). The former models a fully connected architecture, while the latter captures the structure typical of attention layers.

Our contributions can be summarized as follows:

- Theorem 4.1 shows that an RF map has *low* WS, specifically of order $1/\sqrt{n}$, where n is the context length. This means that changing a single word has a negligible effect on the output of the map. In fact, to have a significant effect, one needs to change a constant fraction of the words, see Remark 4.2. Furthermore, increasing the depth of the architecture does not help with the word sensitivity, as shown by Theorem 4.3.

- Our main result, Theorem 5.1, shows that a RAF map has *high* WS, specifically of constant order which does not depend on the length of the context. This means that changing even a single word can have a significant effect on the output of the map, regardless of the context length. The argument critically exploits the role of the $\mathrm{softmax}$ in the attention layer, and numerical simulations show its advantages compared to other activations (e.g., ReLU).

- Section 6 exploits the bounds on the word sensitivity to characterize the generalization error when the data changes meaning after modifying a single word in the context. In particular, we consider generalized linear models trained on RF and RAF embeddings, and we establish whether a fine-tuned or retrained model can learn to distinguish between two samples that differ only in one row. While the answer is provably negative for random features (Theorems 6.2 and 6.3), random attention features are capable of generalizing.

Most of our technical contributions require no distributional assumptions on the data, and the generality of our findings is confirmed by numerical results on the

BERT-Base embeddings of the imdb dataset (Maas et al., 2011), and on the pre-trained BERT-Base model itself. Our code is publicly available at the GitHub repository `simone-bombari/attention-sensitivity`.

2. Related work

Fully connected layers. Several mathematical models have been proposed to understand phenomena occurring for fully connected architectures. A prototypical example is the RF model (Rahimi & Recht, 2007; Pennington & Worah, 2017; Louart et al., 2018; Mei & Montanari, 2022), which can be thought of as a two-layer neural network with random hidden weights. Its feature learning capabilities have been recently studied in settings where one gradient step on the hidden weights is performed before the final training of the outer layer (Ba et al., 2022; 2023; Damian et al., 2022; Moniri et al., 2024). Other popular approaches involve the neural tangent kernel (Jacot et al., 2018; Lee et al., 2019; Ghorbani et al., 2020; 2021) and a mean-field analysis (Mei et al., 2018; Sirignano & Spiliopoulos, 2020; Chizat & Bach, 2018; Rotskoff & Vanden-Eijnden, 2018; Javanmard et al., 2020; Shevchenko et al., 2021). Deep random models have also been considered by Hanin (2019); Bosch et al. (2023); Schröder et al. (2023).

Attention layers. Attention layers (Bahdanau et al., 2015; Kim et al., 2017) and transformer architectures (Vaswani et al., 2017) have attracted significant interest from the theoretical community: Yun et al. (2020); Ben-Shaul et al. (2023) study their approximation capabilities; Edelman et al. (2022); Trauger & Tewari (2023) provide norm-based generalization bounds; Tian et al. (2023) analyze the training dynamics; Wu et al. (2023) provide optimization guarantees; Jelassi et al. (2022); Li et al. (2023) focus on computer vision tasks and Oymak et al. (2023) on prompt-tuning. The study of the attention mechanism is approached through

the lens of associative memories by Bietti et al. (2023); Cabannes et al. (2024). More closely related to our setting is the recent work by Fu et al. (2023), which compares the sample complexity of random attention features with that of random features. We highlight that Fu et al. (2023) focus on an attention layer with ReLU activation, while we unveil the critical role of the $\mathrm{softmax}$.

Sensitivity of neural networks. We informally use the term sensitivity to express how a perturbation of the input changes the output of the model. Previous work explored various mathematical formulations of this concept (e.g., the input-output Lipschitz constant) in the context of both robustness (Weng et al., 2018; Bubeck & Sellke, 2021) and generalization (Bartlett et al., 2017). Sensitivity is generally referred to as an *undesirable* property, motivating research on models that reduce it (Miyato et al., 2018; Prach & Lampert, 2022). In this work, however, high sensitivity represents a *desirable* attribute, as it reflects the ability of the model to capture the role of individual words in a long context. This is a stronger requirement than having large Lipschitz constant, hence earlier results on the matter (Kim et al., 2021) cannot be applied.

3. Preliminaries

We consider a sequence of n tokens $\{x_i\}_{i=1}^n$, with $x_i \in \mathbb{R}^d$ for every i, where d denotes the token embedding dimension, and n the context length. These tokens altogether represent the textual sample $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times d}$. We denote by $\mathrm{flat}(X) \in \mathbb{R}^D$, with $D = nd$, the flattened (or vectorized) version of X. Given a vector x, $\|x\|_2$ is its Euclidean norm. Given a matrix A, $\|A\|_2 := \|\mathrm{flat}(A)\|_2 \equiv \|A\|_F$ is its Frobenius norm. We indicate with e_i the i-th element of the canonical basis, and denote $[n] := \{1, \ldots, n\}$. All the complexity notations $\Omega(\cdot), \omega(\cdot), \mathcal{O}(\cdot), o(\cdot)$ and $\Theta(\cdot)$ are understood for sufficiently large context length n, token embedding dimension d, number of neurons k, and number of input samples N. We indicate with $C, c > 0$ numerical constants, independent of n, d, k, N. Throughout the paper, we make the following assumption, which is easily achieved by pre-processing the raw data.

Assumption 3.1 (Normalization of token embedding). For every token x_i, we assume $\|x_i\|_2 = \sqrt{d}$.

Random Features (RF). A fully connected layer with random weights is commonly referred to as a *random features* map (Rahimi & Recht, 2007). The map acts from a vector of covariates to a feature space $\mathbb{R}^k$, where k denotes the number of neurons. Thus, we flatten the context $\mathrm{flat}(X) \in \mathbb{R}^D$ before feeding it in the layer, and the RF map $\varphi_{\mathrm{RF}} : \mathbb{R}^{n \times d} \to \mathbb{R}^k$ takes the form

$$\varphi_{\mathrm{RF}}(X) = \phi(V \mathrm{flat}(X)), \qquad (1)$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a non-linearity applied componentwise and $V \in \mathbb{R}^{k \times D}$ is the random features matrix, with $V_{i,j} \sim_{\mathrm{i.i.d.}} \mathcal{N}(0, 1/D)$. This scaling of the variance of V ensures that the entries of $V \mathrm{flat}(X)$ have unit variance, as $\|\mathrm{flat}\, X\|_2 = \sqrt{D}$ by Assumption 3.1. We later consider a similar model with several random layers, referred to as *deep random features* (DRF) model and recently considered by Bosch et al. (2023); Schröder et al. (2023).

Random Attention Features (RAF). We consider a single-head sequence-to-sequence self-attention layer without biases $\varphi_{\mathrm{QKV}}(X)$ (Vaswani et al., 2017), given by

$$\varphi_{\mathrm{QKV}}(X) = \mathrm{softmax}\left(\frac{XW_Q^\top W_K X^\top}{\sqrt{d'}}\right) XW_V^\top, \quad (2)$$

where the $\mathrm{softmax}$ is applied row-wise and defined as $\mathrm{softmax}(s)_i = e^{s_i}/\sum_j e^{s_j}$; $W_Q, W_K, W_V \in \mathbb{R}^{d' \times d}$ are respectively the queries, keys and values weight matrices. As in previous related work (Fu et al., 2023), we simplify the above expression with the re-parameterization $W := W_Q^\top W_K \in \mathbb{R}^{d \times d}$, and by removing the values weight matrix. This is for convenience of presentation, and our results can be generalized to the case where queries, keys and values are random independent features, see Remark 5.2 at the end of Section 5. Thus, we define the *random attention features* layer $\varphi_{\mathrm{RAF}} : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d}$ as

$$\varphi_{\mathrm{RAF}}(X) = \mathrm{softmax}\left(\frac{XWX^\top}{\sqrt{d}}\right) X, \qquad (3)$$

where $W_{i,j} \sim_{\mathrm{i.i.d.}} \mathcal{N}(0, 1/d)$. We refer to the argument of the $\mathrm{softmax}$ with the shorthand $S(X) := XWX^\top/\sqrt{d} \in \mathbb{R}^{n \times n}$. The scaling of the variance of W ensures that the entries of $S(X)$ have unit variance. Differently from (Fu et al., 2023), we do not consider the biased initialization discussed in (Trockman & Kolter, 2023), designed to make the diagonal elements of W positive in expectation. As remarked in (Trockman & Kolter, 2023), this initialization is aimed at replicating the final attention scores of vision transformers, and its utility on language models is less discussed. Furthermore, we remark that our simplified model does not include any of the architectural tweaks introduced to allow transformer models to process longer contexts (Bertsch et al., 2023; Beltagy et al., 2020), as it aims to capture the specific properties of a single attention layer.

Word sensitivity (WS). The word sensitivity (WS) measures how the embedding of a given mapping $\varphi(X)$ is impacted by a change in a single word. Given a sample $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times d}$, the *perturbed* sample $X^i(\Delta)$ is obtained from X by setting its i-th row to $x_i + \Delta$ and keeping the remaining rows the same. Here, $\Delta \in \mathbb{R}^d$ denotes the perturbation of the i-th token, and its magnitude does

not exceed the scaling of Assumption 3.1, i.e., $\|\Delta\|_2 \leq \sqrt{d}$. Formally, given a mapping φ, we are interested in

$$\mathcal{S}_\varphi(X) = \sup_{i \in [n], \|\Delta\|_2 \leq \sqrt{d}} \frac{\left\|\varphi(X^i(\Delta)) - \varphi(X)\right\|_2}{\left\|\varphi(X)\right\|_2}. \quad (4)$$

In words, $\mathcal{S}_\varphi(X)$ denotes the highest relative change of $\varphi(X)$, upon changing a single token in the input. Our goal is to study how $\mathcal{S}_\varphi(X)$ behaves for the RF and RAF models, with respect to the context length n. Informally, if $\mathcal{S}_\varphi(X) = o(1)$ for any X, the mapping φ has *low word sensitivity*. On the contrary, if $\mathcal{S}_\varphi(X) = \Omega(1)$, φ has *high word sensitivity*.

We remark that, in language models, tokens are elements of a discrete vocabulary. However, working directly in the embedding space (as in definition (4)) is common in the theoretical literature (Kim et al., 2021), and critical steps of practical algorithms for language models also take place in the embedding space (Ebrahimi et al., 2018; Shin et al., 2020; Zou et al., 2023).

The definition of WS recalls similar notions of sensitivity in the context of adversarial robustness (Dohmatob & Bietti, 2022; Bubeck & Sellke, 2021; Wu et al., 2021; Bombari et al., 2023), as well as the literature that designs adversarial prompting schemes for language models (Ebrahimi et al., 2018; Guo et al., 2021; Zou et al., 2023). However, in contrast with the definition in (4), these works consider a *trained* model and, therefore, the results implicitly depend on the training dataset. Our analysis of the WS dissects the impact of the feature map φ, and it does not involve any training process. The consequences on training (and generalization error) will be considered in Section 6.

4. Low WS of random features

We start by showing that the word sensitivity for the RF map in (1) is *low*, i.e., $\mathcal{S}_{\mathrm{RF}} = o(1)$.

Theorem 4.1. *Let φ_{RF} be the random features map defined in (1), where ϕ is Lipschitz and not identically 0. Let $X \in \mathbb{R}^{n \times d}$ be a generic input sample s.t. Assumption 3.1 holds, and assume $k = \Omega(D)$. Let $\mathcal{S}_{\mathrm{RF}}(X)$ denote the word sensitivity of φ_{RF} defined in (4). Then, we have*

$$\mathcal{S}_{\mathrm{RF}}(X) = \mathcal{O}\left(1/\sqrt{n}\right) = o(1), \quad (5)$$

with probability at least $1 - \exp(-cD)$ over V.

Theorem 4.1 shows that the RF model has a low word sensitivity, vanishing with the length of the context n. This means that, regardless of how any word is modified, the RF mapping is *not sensitive* to this modification, when length of the context is large. The proof follows from an upper bound on the numerator of (4), due to the Lipschitz continuity of ϕ, and a concentration result on the norm at the denominator. The details are deferred to Appendix B.
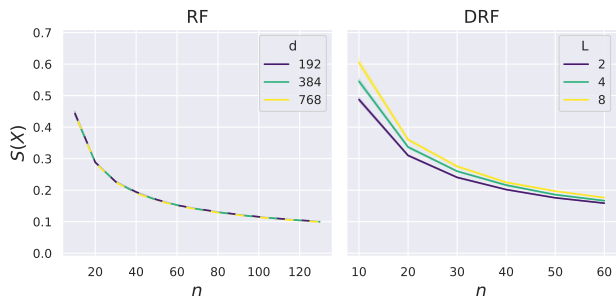

Figure 2. Numerical estimate of the WS for the RF (left) and DRF (right) map, with a ReLU activation function. We estimate $\mathcal{S}_\varphi$ looking for the perturbation $\Delta^* = \arg\sup_{\|\Delta\|_2 \leq \sqrt{d}} \left\|\varphi(X^1(\Delta)) - \varphi(X)\right\|_2$, where we fix the first token for symmetry. We find Δ^* by optimizing our objective with constrained gradient ascent. For RF, we consider $d \in \{192, 384, 768\}$, as n increases (taking the first d dimensions of the embeddings of the first n tokens). For DRF, we repeat the experiment for different depths $L \in \{2, 4, 8\}$ and fixed $d = 768$, as n increases. As textual data X, we use the BERT-Base token embeddings of samples from the imdb dataset, after a pre-processing to adapt the dimensions and fulfill Assumption 3.1. We plot the average over 10 independent trials and the confidence band at 1 standard deviation. In the figure on the left, we intentionally dash the plotted lines to ease the visualization, as they overlap.

Remark 4.2. Theorem 4.1 is readily extended to the case where the number of words m that can be modified in the context is $o(n)$. In fact, to achieve a sensitivity of constant order, a constant fraction of rows in X must be changed, i.e. $m = \Theta(n)$. This extension of Theorem 4.1 is also proved in Appendix B, and it further illustrates that a fully connected architecture is unable to capture the change of few (namely, $m = o(n)$) semantically relevant words in a long sentence.

We numerically validate this result in Figure 2 (left), where we estimate the value of $\mathcal{S}_{\mathrm{RF}}$ for different token embedding dimensions d, as the context length n increases. Clearly, $\mathcal{S}_{\mathrm{RF}}$ decreases as the context length increases, regardless of the embedding dimension d. In fact, the curves corresponding to different values of d (in different colors) basically coincide.

Deep Random Features (DRF). As an extension of Theorem 4.1, we show that the *word sensitivity is low* also for *deep* random features. We consider the DRF map

$$\varphi_{\mathrm{DRF}}(X) := \phi(V_L \phi(V_{L-1}(...V_2 \phi(V_1 \mathrm{flat}(X))...))), \quad (6)$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is the component-wise non-linearity, k is the number of neurons at each layer, and $V_l \in \mathbb{R}^{k \times k}$ are the random weights at layer l, with $[V_1]_{i,j} \sim_{\mathrm{i.i.d.}} \mathcal{N}(0, \beta/D)$ and $[V_l]_{i,j} \sim_{\mathrm{i.i.d.}} \mathcal{N}(0, \beta/k)$ for $l > 1$. We set β according to He's initialization (He et al., 2015), which ensures

$$\mathbb{E}_{\rho \sim \mathcal{N}(0,\beta)}\left[\phi^2(\rho)\right] = 1, \quad (7)$$

as done in (Hanin, 2018) to avoid the problem of vanishing/exploding gradients in the analysis of a deep network.

4

Theorem 4.3. *Let φ_{DRF} be the deep random feature map defined in* (6)*, where ϕ is Lipschitz and β is chosen s.t.* (7) *holds. Let $X \in \mathbb{R}^{n \times d}$ be a generic input sample s.t. Assumption 3.1 holds, and assume $k = \Theta(D)$, $L = o(\log k)$. Let $\mathcal{S}_{\mathrm{DRF}}(X)$ denote the word sensitivity of φ_{DRF} defined in* (4)*. Then, we have*

$$\mathcal{S}_{\mathrm{DRF}} = \mathcal{O}\left(\frac{e^{CL}}{\sqrt{n}}\right), \tag{8}$$

with probability at least $1 - \exp(-c \log^2 d)$ over $\{V_l\}_{l=1}^{L}$.

Theorem 4.3 shows that, as in the shallow case, the word sensitivity decreases with the length of the context n. The proof requires showing that $\|\varphi_{\mathrm{DRF}}(X)\|_2$ concentrates to $\sqrt{k}$, which is achieved via (7). The strategy to bound the term $\|\varphi_{\mathrm{DRF}}(X^i(\Delta)) - \varphi_{\mathrm{DRF}}(X)\|_2$ is similar to that of the shallow case. The details are deferred to Appendix C.

The exponential dependence on L comes from our worst-case analysis of the Lipschitz constant of the model, and it does not fully exploit the independence between the V_i's. Thus, we expect the actual dependence of the WS on L to be milder. This is confirmed by the numerical results of Figure 2 (right), where we numerically estimate $\mathcal{S}_{\mathrm{DRF}}$ for different depths L as the context length n increases. For all values of L, the word sensitivity quickly decreases with n.

5. High WS of random attention features

In contrast with random features, we show that the *word sensitivity is high* for the RAF map in (3), i.e., $\mathcal{S}_{\mathrm{RAF}} = \Omega(1)$.

Theorem 5.1. *Let φ_{RAF} be the random attention features map defined in* (3)*. Let $X \in \mathbb{R}^{n \times d}$ be a generic input sample s.t. Assumption 3.1 holds, and assume $d/\log^4 d = \Omega(n)$. Let $\mathcal{S}_{\mathrm{RAF}}(X)$ denote the word sensitivity of φ_{RAF} defined in* (4)*. Then, we have*

$$\mathcal{S}_{\mathrm{RAF}}(X) = \Omega(1), \tag{9}$$

with probability at least $1 - \exp(-c \log^2 d)$ over W.

Theorem 5.1 shows that the RAF model has a high word sensitivity, regardless of the length of the context n, as long as $d/\log^4 d = \Omega(n)$. This requires a number of tokens n that grows slower than the embedding dimension d. Models such as BERT-Base or BERT-Large have an embedding dimension of 768 and 1024 (Devlin et al., 2019), which allows our results to hold for fairly large context lengths. In fact, we prove a stronger statement: for *any* index $i \in [n]$, there exists a perturbation Δ^* (possibly dependent on i) s.t. the RAF map changes significantly when evaluated in $X^i(\Delta^*)$. The proof of Theorem 5.1 is deferred to Appendix D, and a sketch follows.

Proof sketch. The argument is not constructive, and the difficulty in finding a closed-form solution for the perturbation Δ^* is due to the lack of assumptions on the sample X, which is entirely generic. We follow the steps below.

Step 1: Find a direction δ^* aligned with many words x_j's. Using the probabilistic method, we prove the existence of a vector $\delta^* \in \mathbb{R}^d$, with $\|\delta^*\|_2 \leq \sqrt{d}$, s.t. its inner product with the tokens embeddings x_j's is large for a constant fraction of the words in the context. In particular, Lemma D.1 shows that $\left(x_j^\top \delta^*\right)^2 = \Omega(d^2/n)$ for at least $\Omega(n)$ indices $j \in [n]$.

Step 2: Exhibit two directions Δ_1^* and Δ_2^* both aligned with many words in the feature space $\{W^\top x_j\}_{j=1}^n$. Exploiting the properties of the Gaussian attention features W, we deduce the existence of two vectors Δ_1^* and Δ_2^* that *(i)* are far from each other, and *(ii)* ensure a constant fraction of the entries of $XW\Delta_k^*/\sqrt{d}$, $k \in \{1, 2\}$, to be large. In particular, by exploiting our assumption $d/\log^4 d = \Omega(n)$, Lemmas D.2 and D.3 show that $[XW\Delta_k^*/\sqrt{d}]_j = \Omega\left(\log^2 d\right)$ for at least $\Omega(n)$ indices $j \in [n]$, with $\|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d})$.

Step 3: Show that the attention concentrates towards the perturbed word. Recall that $s(X) := \mathrm{softmax}(XWX^\top/\sqrt{d})$ denotes the attention scores matrix. Then, Lemma D.4 proves that the attention scores $[s(X^i(\Delta_k^*))]_{j:}^\top$ are well approximated by the canonical basis vector e_i, for $k \in \{1, 2\}$ and an $\Omega(n)$ number of rows $j \in [n]$. This intuitively means that a constant fraction of tokens x_j *moves all their attention* towards the i-th modified token $x_i + \Delta_k^*$. This step critically exploits the $\mathrm{softmax}$ function in the RAF map: if one entry in its argument is $\Omega\left(\log^2 d\right)$, then the attention scores concentrate as described above.

Step 4: Conclude with at least one perturbation between Δ_1^* and Δ_2^*. Finally, exploiting the fact that $\|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d})$, Lemma D.5 proves that at least one of them gives a Δ^* s.t. $\|\varphi_{\mathrm{RAF}}(X) - \varphi_{\mathrm{RAF}}(X^i(\Delta^*))\|_F = \Omega(\sqrt{dn})$. This, together with the upper bound $\|\varphi_{\mathrm{RAF}}(X)\|_F = \mathcal{O}(\sqrt{dn})$, concludes the argument. $\square$

In Figure 3 (first plot), we estimate the value of $\mathcal{S}_{\mathrm{RAF}}$ for different token embedding dimensions d, as the context length n increases. In contrast with the random features map, even for large values of n, the WS remains larger than 1. In the same figure, in the second plot, we repeat the experiment for the ReLU-RAF map, which replaces the $\mathrm{softmax}$ with a ReLU activation. $\mathcal{S}_{\mathrm{ReLU\text{-}RAF}}(X)$ seems to decrease with the context length n, and has in general smaller values than $\mathcal{S}_{\mathrm{RAF}}(X)$. This highlights the importance of the $\mathrm{softmax}$ function, as discussed in *Step 3* of the proof sketch above. The condition $d/\log^4 d = \Omega(n)$ is required to allow *step 2* to go through. We believe this to be a reasonable assumption, as the maximum context length tends to be smaller than the embedding dimension. Popular examples include BERT-Base ($n = 512$, $d = 768$), BERT-Large ($n = 512$, $d = 1024$), and the Llama-2 family ($n = 4096$, $d = 5120$).
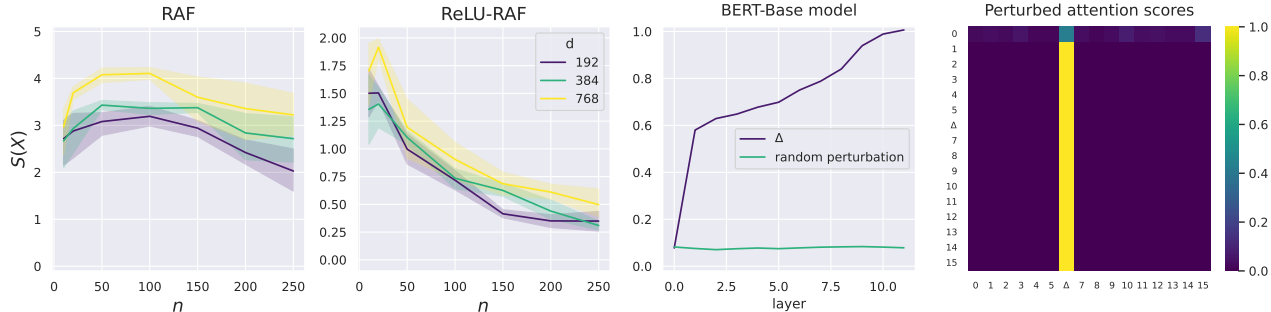
Figure 3. Numerical estimate of the WS for the RAF (first plot) and ReLU-RAF (second plot) map. The ReLU-RAF map is defined as the RAF one, but the $\mathrm{softmax}$ is replaced with a ReLU activation over the entries of $S(X)$, followed by a re-normalization to ensure that the attention scores sum up (on every row) to 1, as in the $\mathrm{softmax}$ case. We consider $d = 192, 384, 768$, as n increases. The rest of the setup is equivalent to the one described in Figure 2. In the third plot, we present the relative change of the pre-trained BERT-Base model layer embeddings, evaluated on the abstract of this paper (~ 200 tokens), when the 42nd token is modified in embedding space with a vector Δ. The 0-th layer represents the input itself and, as a comparison, we report the results when the perturbation is chosen to be Gaussian noise with the same norm. In the fourth plot, we present the attention scores in the first head of the first layer of BERT-Base model, evaluated on the title of this paper, when the 6th token is modified in embedding space with a vector Δ. In the third and fourth plots, Δ is chosen to be a perturbation that *attracts* all the attention on the perturbed key token, which follows the proof idea of Theorem 5.1.

Remark 5.2. The re-parameterization of the attention layer through the features W, which removes the dependence on queries, keys and values, does not substantially change the problem. In fact, *Step 2* of the argument uses that W acts as an approximate isometry on δ^*. This would also hold for the product of two independent Gaussian matrices $W_Q^\top W_K$. Similarly, introducing the independent Gaussian matrix W_V would not interfere with our conclusion in *Step 4*. Additionally, the results obtained on the RAF model seem to extend to realistic, pre-trained, transformer architectures. In fact, in the third plot of Figure 3, we provide a lower bound on the sensitivity of the BERT-Base architecture, evaluated on the abstract of this paper, when the 42nd token is modified by a perturbation Δ. The blue line shows how after the first layer the embeddings are heavily modified, and how this change increases deeper in the architecture. In this case, the perturbation Δ follows from the proof idea of Theorem 5.1, as it is chosen to *move all the attention* towards x_i (see the fourth plot in Figure 3). To do so, we set Δ to be aligned with the right singular vector of $W_Q^\top W_K$ associated with the largest singular value. We do this on all the heads in the first layer separately, and then average and re-normalize.

6. Generalization on context modification

The study of the word sensitivity is motivated by understanding the capabilities of a model to learn to distinguish two contexts X and $X^i(\Delta)$ that only differ by a word. In fact, in practice, modifying a single row/word of X can lead to a significant change in the meaning, see Figure 1. We formalize the problem in a supervised learning setting, and character-

ize whether a *generalized linear model* (GLM) induced by a feature map φ generalizes over a sample $(X^i(\Delta), y_\Delta)$, after being trained on (X, y). Crucially, the index $i \in [n]$ and the perturbation Δ are s.t. the label y_Δ is different from y (e.g., $y \in \{-1, +1\}$ and $y_\Delta = -y$), namely, perturbing the i-th word changes the meaning of the context.

Supervised learning with generalized linear models (GLMs). Let $(\mathcal{X}, \mathcal{Y})$ be a labelled training dataset, where $\mathcal{X} = (X_1, \ldots, X_N)$ contains the training data $X_i \in \mathbb{R}^{n \times d}$ and $\mathcal{Y} = [y_1, \ldots, y_N]^\top \in \{-1, 1\}^N$ the corresponding binary labels. The sample (X, y) does not belong to $(\mathcal{X}, \mathcal{Y})$, and it is introduced later in the training set. Let $\varphi : \mathbb{R}^{n \times d} \to \mathbb{R}^p$ be a feature map, and consider the GLM

$$f_\varphi(\cdot, \theta) = \varphi(\cdot)^\top \theta, \tag{10}$$

where $\theta \in \mathbb{R}^p$ are trainable parameters of the model. We define the feature matrix as $\Phi_\varphi := [\varphi(X_1), \ldots, \varphi(X_N)]^\top \in \mathbb{R}^{N \times p}$, and focus on the quadratic loss

$$\mathcal{L}(\theta) := \frac{1}{N} \sum_{j=1}^N \left(\varphi(X_j)^\top \theta - y_j \right)^2. \tag{11}$$

Minimizing (11) with gradient descent gives (see equation (33) in (Bartlett et al., 2021))

$$\theta^* = \theta_0 + \Phi_\varphi^+ (\mathcal{Y} - \Phi_\varphi \theta_0), \tag{12}$$

where θ^* is the gradient descent solution, θ_0 is the initialization, and Φ_φ^+ is the Moore-Penrose inverse of Φ_φ.

Our goal is to establish whether the additional training on the sample (X, y) allows the model f_φ to generalize on

$(X^i(\Delta), y_\Delta)$, with $y_\Delta = -y$. Thus, we do *not* focus on the case where $f_\varphi(X, \theta^*) = y$ and $f_\varphi(X^i(\Delta), \theta^*) = y_\Delta$, i.e., $f_\varphi(\cdot, \theta^*)$ already generalizes well on the two new samples. Instead, we look at how $f_\varphi(\cdot, \theta^*)$ *extrapolates* the information contained in the pair (X, y) to the perturbed sample $X^i(\Delta)$. This motivates the following assumption.

Assumption 6.1. There exists a parameter $\gamma \in [0, 2)$ s.t.

$$\left| f_\varphi(X^i(\Delta), \theta^*) - f_\varphi(X, \theta^*) \right| \leq \gamma. \quad (13)$$

The parameter γ captures the degree over which the trained model $f_\varphi(\cdot, \theta^*)$ can distinguish between X and $X^i(\Delta)$. If $\gamma = 0$, $f_\varphi(\cdot, \theta^*)$ does not recognize any difference between X and $X^i(\Delta)$; instead, if $f_\varphi(\cdot, \theta^*)$ correctly classifies the two samples X and $X^i(\Delta)$ (i.e., $f_\varphi(X, \theta^*) = y$ and $f_\varphi(X^i(\Delta), \theta^*) = y_\Delta = -y$), then $\gamma = 2$, which is beyond the scope of our analysis.

We consider training on (X, y) in the following two ways.

(a) Fine-tuning. First, we look at the model obtained after *fine-tuning* the solution θ^* defined in (12) over the new sample (X, y). This means that θ^* is the initialization of a gradient descent algorithm trained on this single sample. As $\left(\varphi(X)^\top\right)^+ = \varphi(X)/\|\varphi(X)\|_2^2$, the fine-tuned solution is given by

$$\theta_f^* = \theta^* + \frac{\varphi(X)}{\|\varphi(X)\|_2^2}\left(y - \varphi(X)^\top\theta^*\right). \quad (14)$$

(b) Re-training. Second, we re-train the model from scratch, after adding the pair (X, y) to the training set. The new training set is denoted by $(\mathcal{X}_r, \mathcal{Y}_r)$, where $\mathcal{X}_r = (X_1, \ldots, X_N, X)$ contains the training data and $\mathcal{Y}_r = [y_1, \ldots, y_N, y]$ the binary labels. Thus, denoting by $\Phi_{\varphi,r} := [\varphi(X_1), \ldots, \varphi(X_N), \varphi(X)]^\top \in \mathbb{R}^{(N+1)\times p}$ the new feature matrix, the re-trained solution θ_r^* takes the form

$$\theta_r^* = \theta_0 + \Phi_{\varphi,r}^+(\mathcal{Y}_r - \Phi_{\varphi,r}\theta_0). \quad (15)$$

The quantity of interest is the *test error* on $(X^i(\Delta), y_\Delta)$:

$$\mathrm{Err}_\varphi(X^i(\Delta), \theta) := \left(f_\varphi(X^i(\Delta), \theta) - y_\Delta\right)^2, \quad (16)$$

where $\theta \in \{\theta_f^*, \theta_r^*\}$ is the vector of parameters obtained either after fine-tuning or re-training.

6.1. Random features do not generalize

By exploiting the low word sensitivity of random features, we show that both the fine-tuned and retrained solutions generalize poorly.

Theorem 6.2. *Let φ_{RF} be the random features map defined in* (1)*, with ϕ Lipschitz and not identically 0, and let $f_{\mathrm{RF}}(\cdot, \theta_f^*) = \varphi_{\mathrm{RF}}(\cdot)^\top\theta_f^*$ be the corresponding model*

fine-tuned on the sample (X, y), where $X \in \mathbb{R}^{n\times d}$ satisfies Assumption 3.1 and θ_f^* is given by (14). Assume $k = \Omega(D)$, $|f_{\mathrm{RF}}(X, \theta^*)| = \mathcal{O}(1)$, and that Assumption 6.1 holds with $\gamma \in [0, 2)$. Let $\mathrm{Err}_{\mathrm{RF}}(X^i(\Delta), \theta_f^*)$ be the test error of $f_{\mathrm{RF}}(\cdot, \theta_f^*)$ on $(X^i(\Delta), y_\Delta)$ as defined in (16). Then, for any Δ s.t. $\|\Delta\|_2 \leq \sqrt{d}$ and any $i \in [n]$, we have

$$\mathrm{Err}_{\mathrm{RF}}(X^i(\Delta), \theta_f^*) > (2 - \gamma)^2 - \mathcal{O}\left(1/\sqrt{n}\right), \quad (17)$$

with probability at least $1 - \exp(-cD)$ over V.

Proof sketch. The idea is to use (14) to obtain that

$$\begin{aligned} & f_{\mathrm{RF}}(X^i(\Delta), \theta_f^*) - f_{\mathrm{RF}}(X^i(\Delta), \theta^*) \\ & = \frac{\varphi_{\mathrm{RF}}(X^i(\Delta))^\top\varphi_{\mathrm{RF}}(X)}{\|\varphi_{\mathrm{RF}}(X)\|_2^2}\left(y - f_{\mathrm{RF}}(X, \theta^*)\right). \end{aligned} \quad (18)$$

Next, we note that

$$\left| \frac{\varphi_{\mathrm{RF}}(X^i(\Delta))^\top\varphi_{\mathrm{RF}}(X)}{\|\varphi_{\mathrm{RF}}(X)\|_2^2} - 1 \right| \leq \mathcal{S}_{\mathrm{RF}}(X). \quad (19)$$

Theorem 4.1 gives that the RHS of (19) is small, which combined with (18) implies that $f_{\mathrm{RF}}(X^i(\Delta), \theta_f^*)$ is close to

$$y + f_{\mathrm{RF}}(X^i(\Delta), \theta^*) - f_{\mathrm{RF}}(X, \theta^*). \quad (20)$$

By upper bounding $f_{\mathrm{RF}}(X^i(\Delta), \theta^*) - f_{\mathrm{RF}}(X, \theta^*)$ via Assumption 6.1, we obtain that (20) cannot be far from y. This implies that $f_{\mathrm{RF}}(X^i(\Delta), \theta_f^*)$ cannot be close to the correct label $y_\Delta = -y$. The complete proof is in Appendix B. $\square$

Theorem 6.3. *Let φ_{RF} be the random features map defined in* (1)*, with ϕ Lipschitz and non-linear, and let $f_{\mathrm{RF}}(\cdot, \theta_r^*) = \varphi_{\mathrm{RF}}(\cdot)^\top\theta_r^*$ be the corresponding model re-trained on the dataset $(\mathcal{X}_r, \mathcal{Y}_r)$ that contains the pair (X, y), thus with θ_r^* defined in* (15)*. Assume the training data to be sampled i.i.d. from a distribution $\mathcal{P}_X$ s.t. $\mathbb{E}_{X\sim\mathcal{P}_X}[X] = 0$, Assumption 3.1 holds, and the Lipschitz concentration property is satisfied. Let $N\log^3 N = o(k)$, $N\log^4 N = o(D^2)$ and $k = \Omega(D)$. Assume that $|f_{\mathrm{RF}}(X, \theta^*)| = \mathcal{O}(1)$ and that Assumption 6.1 holds with $\gamma \in [0, 2)$. Let $\mathrm{Err}_{\mathrm{RF}}(X^i(\Delta), \theta_r^*)$ be the test error of $f_{\mathrm{RF}}(\cdot, \theta_r^*)$ on $(X^i(\Delta), y_\Delta)$ defined in* (16)*. Then, for any Δ s.t. $\|\Delta\|_2 \leq \sqrt{d}$ and any $i \in [n]$, we have*

$$\mathrm{Err}_{\mathrm{RF}}(X^i(\Delta), \theta_r^*) > (2 - \gamma)^2 - \mathcal{O}\left(1/\sqrt{n}\right), \quad (21)$$

with probability at least $1 - \exp\left(-c\log^2 N\right)$ over $V, \mathcal{X}_r$.

Proof sketch. The idea is to leverage the stability analysis in Bombari & Mondelli (2024), which gives

$$\begin{aligned} & f_{\mathrm{RF}}(X^i(\Delta), \theta_r^*) - f_{\mathrm{RF}}(X^i(\Delta), \theta^*) \\ & = \mathcal{F}_{\mathrm{RF}}(X, X^i(\Delta))\left(f_{\mathrm{RF}}(X, \theta_r^*) - f_{\mathrm{RF}}(X, \theta^*)\right), \end{aligned} \quad (22)$$

where

$$\mathcal{F}_{\mathrm{RF}}(X, X^i(\Delta)) := \frac{\varphi_{\mathrm{RF}}(X^i(\Delta))^\top P_\Phi^\perp \varphi_{\mathrm{RF}}(X)}{\left\| P_\Phi^\perp \varphi_{\mathrm{RF}}(X) \right\|_2^2} \quad (23)$$

is the *feature alignment* between X and $X^i(\Delta)$ induced by φ_{RF} and P_Φ the projector over $\mathrm{Span}\{\varphi_{\mathrm{RF}}(X_1), \ldots, \varphi_{\mathrm{RF}}(X_N)\}$. After some manipulations, we have

$$\left| \mathcal{F}_{\mathrm{RF}}(X, X^i(\Delta)) - 1 \right| \leq \mathcal{S}_{\mathrm{RF}}(X) \frac{\|\varphi_{\mathrm{RF}}(X)\|_2}{\sqrt{\lambda_{\min}(K_{\mathrm{RF},r})}}, \quad (24)$$

where $K_{\mathrm{RF},r} := \Phi_{\mathrm{RF},r} \Phi_{\mathrm{RF},r}^\top$ is the kernel of the model. A lower bound on its smallest eigenvalue $\lambda_{\min}(K_{\mathrm{RF},r})$ follows from the fact that the kernel is well-conditioned (see Lemma B.2), which crucially relies on the assumptions on the data (i.i.d. and Lipschitz concentrated) and the scalings $N \log^3 N = o(k)$, $N \log^4 N = o(D^2)$. As the word sensitivity $\mathcal{S}_{\mathrm{RF}}(X)$ is upper bounded by Theorem 4.1, from (24) we conclude that $\mathcal{F}_{\mathrm{RF}}(X, X^i(\Delta))$ is close to 1.

Since $K_{\mathrm{RF},r}$ is invertible, the re-trained model $f_{\mathrm{RF}}(\cdot, \theta_r^*)$ interpolates the dataset $(\mathcal{X}_r, \mathcal{Y}_r)$, giving that $f(X, \theta_r^*) = y$. Thus, as $\mathcal{F}_{\mathrm{RF}}(X, X^i(\Delta)) \approx 1$, $f_{\mathrm{RF}}(X^i(\Delta), \theta_r^*)$ is close to (20), and we conclude from the same argument used for Theorem 6.2. The complete proof is in Appendix B. $\square$

In a nutshell, by exploiting the low word sensitivity of random features, Theorems 6.2-6.3 show that, after either fine-tuning or re-training, the model does not learn to "separate" the predictions on the samples X and $X^i(\Delta)$. As a consequence, the test error is lower bounded by $(2-\gamma)^2$. In fact, γ is the distance between the predictions on X and $X^i(\Delta)$ before fine-tuning/re-training (see (13)), and the ground-truth labels have distance 2 ($y_\Delta, y \in \{-1, 1\}$ and $y_\Delta = -y$).

While Theorem 6.2 does not require distributional assumptions on the data, Theorem 6.3 considers i.i.d. training data, satisfying Lipschitz concentration. This property corresponds to having well-behaved tails, and it is common in the related theoretical literature (Bubeck & Sellke, 2021; Nguyen et al., 2021; Bombari et al., 2022), see Appendix A for the formal definition and a discussion.

We remark that Assumption 6.1 requires the model $f_{\mathrm{RF}}(\cdot, \theta^*)$ to give a similar output when evaluated on the two new samples X and $X^i(\Delta)$. Thus, we are asking if the model generalizes on $X^i(\Delta)$ *only* from the additional training on X. Now, one could design an adversarial Δ s.t. $f(X^i(\Delta), \theta_r^*)$ and $f(X, \theta_r^*)$ are different from each other (so that $f_{\mathrm{RF}}(X^i(\Delta), \theta_r^*) = y_\Delta$ while $f_{\mathrm{RF}}(X, \theta_r^*) = y$), by exploiting the adversarial vulnerability of random features (Dohmatob & Bietti, 2022; Dohmatob, 2022; Bombari et al., 2023). However, if we restrict the possible Δ's to those that satisfy Assumption 6.1, Theorems 6.2 and 6.3 prove that such adversarial patch cannot be found. We finally note that,

when the context length n is comparable or larger than the number of training samples N, the model becomes adversarially robust to any token modification and Assumption 6.1 automatically holds, see Appendix E for details.

6.2. Random attention features can generalize

Next, the behavior of random features is contrasted with that of random attention features. Let us consider the RAF model $f_{\mathrm{RAF}}(\cdot, \theta) := \mathrm{flat}(\varphi_{\mathrm{RAF}}(\cdot))^\top \theta$, where $\varphi_{\mathrm{RAF}}(\cdot)$ is defined in (3). Theorem 5.1 proves that the word sensitivity of $\varphi_{\mathrm{RAF}}(\cdot)$ is large. This suggests that the RAF model is capable of extrapolating the information contained in (X, y) to correctly classify the perturbed sample $X^i(\Delta)$. While proving a rigorous statement on a fine-tuned/re-trained RAF model remains challenging, we provide experimental evidence of this generalization capability.

Figure 4 (first row) shows that, after fine-tuning on (X, y), $f_{\mathrm{RAF}}(X^i(\Delta), \theta_f^*)$ can be close to the perturbed label $y_\Delta = -y$, even if the model before fine-tuning was unable to distinguish between X and $X^i(\Delta)$. Specifically, the two central sub-plots consider the RAF model for two values of the context length $n \in \{40, 120\}$: here, the loss on the perturbed sample can be close to 0, even when the parameter γ in (13) is close to 0, i.e., X and $X^i(\Delta)$ were indistinguishable before fine-tuning; in general, the test error $\mathrm{Err}_{\mathrm{RAF}}(X^i(\Delta), \theta_f^*)$ is often smaller than the lower bound of $(2 - \gamma)^2$ (dashed black line), which holds for random features. The left sub-plot considers the RF model for $n = 40$: here, the loss on the perturbed sample is close to 0 only if the model before fine-tuning was already able to perfectly distinguish between X and $X^i(\Delta)$, i.e., γ is not far from 2; in general, the test error $\mathrm{Err}_{\mathrm{RF}}(X^i(\Delta), \theta_f^*)$ always respects the lower bound of $(2 - \gamma)^2$ proved in Theorem 6.2. Finally, the right sub-plot considers the ReLU-RAF model (which replaces the softmax with a ReLU activation, as described at the end of Section 5) for $n = 120$: here, even if the lower bound of Theorem 6.2 is often violated, the model still cannot reach small error unless γ is large, i.e., X and $X^i(\Delta)$ could be distinguished already before fine-tuning. This confirms the impact of the softmax on the capability of attention layers to understand the context. Analogous results hold when models are re-trained (instead of being fine-tuned), as reported in the second row of the same figure.

In a nutshell, our results show that, when a RAF model is not able to distinguish two points with opposite labels (that only differ in one word), fine-tuning or retraining on one of these points allows the loss on also the other point to decrease. In contrast, this is not the case for the RF model, where the loss on the second point can be lower bounded according to Theorems 6.2 and 6.3. This is shown in Figure 4, where most of the points for the RAF model are below the dashed line representing the lower bound for the RF model.
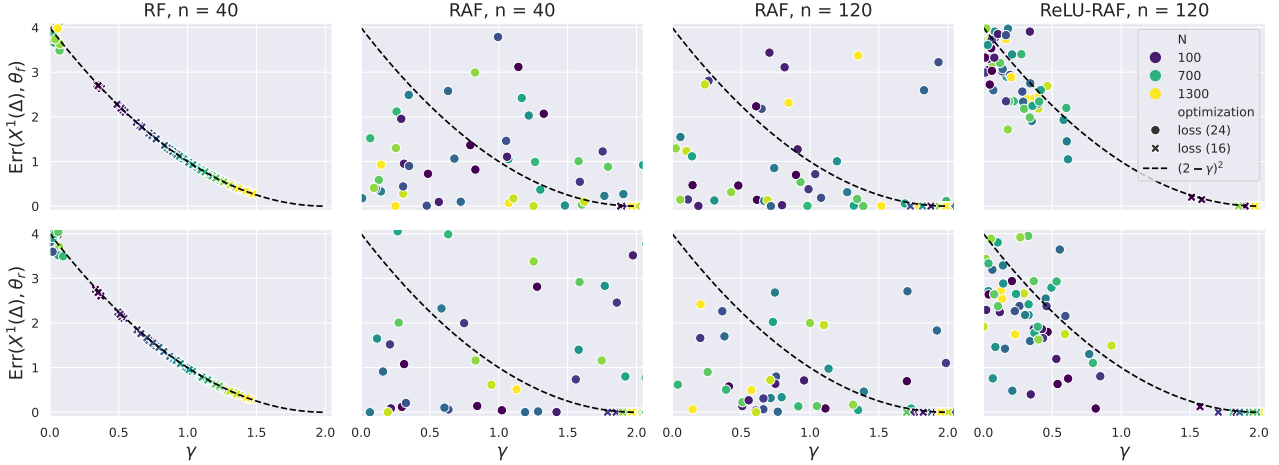
8

Figure 4. Test error (as defined in (16) taking $i = 1$) for the RF (left subplot), RAF (two central sub-plots) and ReLU-RAF (right subplot) maps, as a function of the smallest γ s.t. Assumption 6.1 holds. The first (resp. second) row considers the fine-tuned solution θ_f^* (resp. re-trained solution θ_r^*). Every sub-plot has a fixed embedding dimension $d = 768$, and context length $n \in \{40, 120\}$, taking the first n token embeddings for each sample. Different colors correspond to a different number of training samples $N \in \{100, 700, 1300\}$. Every point in the scatter-plots is an independent simulation where (X, y) and $(\mathcal{X}, \mathcal{Y})$ are the BERT-Base embeddings of a random subset of the imdb dataset (after pre-processing to fulfill Assumption 3.1). Circular markers correspond to obtaining Δ via gradient descent optimization of the losses in (25); cross markers correspond to minimizing directly the test error in (16).

6.3. Experimental details

The experiments of Figure 4 are performed on multiple independent trials, for different choices of the training data. We report in cross markers the results obtained by choosing Δ after optimizing the test error in (16) via gradient descent. While this approach directly minimizes the metric of interest, it results in low test error only when γ is rather large, regardless of the model taken into account (RF, RAF, or ReLU-RAF). In contrast, optimizing a different loss controls the value of γ, while still achieving small error for RAF (and, to a smaller extent, ReLU-RAF). We report in circular markers the results obtained by minimizing the following two losses (respectively, for fine-tuning and re-training):

$$\ell_{\theta_f^*}(\Delta) := \left(\frac{\varphi_{\text{RAF}}(X^i(\Delta))^\top \varphi_{\text{RAF}}(X)}{\|\varphi_{\text{RAF}}(X)\|_2^2} + 1 \right)^2, \quad (25)$$

$$\ell_{\theta_r^*}(\Delta) := \left(\mathcal{F}_{\text{RAF}}(X, X^i(\Delta)) + 1 \right)^2.$$

This choice is suggested by (18) and (22) which, after assuming for simplicity that $f_{\text{RAF}}(X^i(\Delta), \theta^*) = f_{\text{RAF}}(X, \theta^*) = 0$, can be re-written as

$$f_{\text{RAF}}(X^i(\Delta), \theta_f^*) = \frac{\varphi_{\text{RAF}}(X^i(\Delta))^\top \varphi_{\text{RAF}}(X)}{\|\varphi_{\text{RAF}}(X)\|_2^2} \, y,$$

$$f_{\text{RAF}}(X^i(\Delta), \theta_r^*) = \mathcal{F}_{\text{RAF}}(X, X^i(\Delta)) \, y. \quad (26)$$

Achieving small error means that the LHS of (26) is close to $-y$, which corresponds to making the losses in (25) small. Further numerical results and comparisons between different optimization algorithms for finding Δ are in Appendix F.

7. Conclusions

This work provides a formal characterization of the fundamental difference between fully connected and attention layers. To do so, we consider the prototypical setting of random features and study the *word sensitivity*, which captures how the output of a map changes after perturbing a single row/word of the input. On the one hand, the sensitivity of standard random features decreases with the context length and, in order to obtain to a significant change in the output of the map, a constant fraction of the words needs to be perturbed. On the other hand, the sensitivity of random attention features is large, regardless of the context length, thus indicating the suitability of attention layers for NLP tasks. These bounds on the word sensitivity translate into formal negative generalization results for random features, which are contrasted by positive empirical evidence of generalization for the attention layer.

Our analysis allows the perturbations to be any (bounded) vector in the embedding space. Taking the tokenization process (and, hence, the discrete nature of the textual samples) into account offers an exciting avenue for future work.

Acknowledgements

Impact statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., and Yang, G. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., and Wu, D. Learning in the presence of low-dimensional structure: A spiked random matrix perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Bartlett, P. L., Montanari, A., and Rakhlin, A. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021.

Belkin, M. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.

Ben-Shaul, I., Galanti, T., and Dekel, S. Exploring the approximation capabilities of multiplicative neural networks for smooth functions. *Transactions on Machine Learning Research*, 2023.

Bertsch, A., Alon, U., Neubig, G., and Gormley, M. R. Unlimiformer: Long-range transformers with unlimited length input. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Bietti, A., Cabannes, V., Bouchacourt, D., Jegou, H., and Bottou, L. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Bombari, S. and Mondelli, M. How spurious features are memorized: Precise analysis for random and ntk features. In *International Conference on Machine Learning (ICML)*, 2024.

Bombari, S., Amani, M. H., and Mondelli, M. Memorization and optimization in deep neural networks with minimum over-parameterization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Bombari, S., Kiyani, S., and Mondelli, M. Beyond the universal law of robustness: Sharper laws for random features and neural tangent kernels. In *International Conference on Machine Learning (ICML)*, 2023.

Bosch, D., Panahi, A., and Hassibi, B. Precise asymptotic analysis of deep random feature models. In *Annual Conference on Learning Theory (COLT)*, 2023.

Brown, T. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Bubeck, S. and Sellke, M. A universal law of robustness via isoperimetry. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Cabannes, V., Dohmatob, E., and Bietti, A. Scaling laws for associative memories. In *International Conference on Learning Representations (ICLR)*, 2024.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Damian, A., Lee, J., and Soltanolkotabi, M. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory (COLT)*, 2022.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

Dohmatob, E. Fundamental tradeoffs between memorization and robustness in random features and neural tangent regimes. *arXiv preprint arXiv:2106.02630*, 2022.

Dohmatob, E. and Bietti, A. On the (non-) robustness of two-layer neural networks in different learning regimes. *arXiv preprint arXiv:2203.11864*, 2022.

Ebrahimi, J., Rao, A., Lowd, D., and Dou, D. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Edelman, B. L., Goel, S., Kakade, S., and Zhang, C. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning (ICML)*, 2022.

Fu, H., Guo, T., Bai, Y., and Mei, S. What can a single attention layer learn? a study through the random features lens. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. When do neural networks outperform kernel methods? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

Guo, C., Sablayrolles, A., Jégou, H., and Kiela, D. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Hanin, B. Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Hanin, B. Universal function approximation by deep neural nets with bounded width and relu activations. *Mathematics*, 7(10), 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Javanmard, A., Mondelli, M., and Montanari, A. Analysis of a two-layer neural network via displacement convexity. *The Annals of Statistics*, 48(6):3619–3642, 2020.

Jelassi, S., Sander, M. E., and Li, Y. Vision transformers provably learn spatial structure. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kim, H., Papamakarios, G., and Mnih, A. The lipschitz constant of self-attention. In *International Conference on Machine Learning (ICML)*, 2021.

Kim, Y., Denton, C., Hoang, L., and Rush, A. M. Structured attention networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Li, H., Wang, M., Liu, S., and Chen, P.-Y. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. In *International Conference on Learning Representations (ICLR)*, 2023.

Louart, C., Liao, Z., and Couillet, R. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2011.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Moniri, B., Lee, D., Hassani, H., and Dobriban, E. A theory of non-linear feature learning with one gradient step in two-layer neural networks. In *International Conference on Machine Learning (ICML)*, 2024.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations (ICLR)*, 2020.

Nguyen, Q., Mondelli, M., and Montufar, G. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep ReLU networks. In *International Conference on Machine Learning (ICML)*, 2021.

Oymak, S., Rawat, A. S., Soltanolkotabi, M., and Thrampoulidis, C. On the role of attention in prompt-tuning. In *International Conference on Machine Learning (ICML)*, 2023.

Pennington, J. and Worah, P. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Prach, B. and Lampert, C. H. Almost-orthogonal layers for efficient general-purpose lipschitz networks. In *European Conference on Computer Vision (ECCV)*, 2022.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

Rigollet, P. and Hütter, J.-C. High-dimensional statistics. *arXiv preprint arXiv:2310.19244*, 2023.

Rotskoff, G. M. and Vanden-Eijnden, E. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Schröder, D., Cui, H., Dmitriev, D., and Loureiro, B. Deterministic equivalent and error universality of deep random features learning. In *International Conference on Machine Learning (ICML)*, 2023.

Seddik, M. E. A., Louart, C., Tamaazousti, M., and Couillet, R. Random matrix theory proves that deep learning representations of GAN-data behave as gaussian mixtures. In *International Conference on Machine Learning (ICML)*, 2020.

Shevchenko, A. E., Kungurtsev, V., and Mondelli, M. Mean-field analysis of piecewise linear solutions for wide relu networks. *Journal of Machine Learning Research*, 23: 130:1–130:55, 2021.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

Tian, Y., Wang, Y., Chen, B., and Du, S. S. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Touvron, H. et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Trauger, J. and Tewari, A. Sequence length independent norm-based generalization bounds for transformers. *arXiv preprint arXiv:2310.13088*, 2023.

Trockman, A. and Kolter, J. Z. Mimetic initialization of self-attention layers. In *International Conference on Machine Learning (ICML)*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018.

Weng, T.-W., Zhang, H., Chen, P.-Y., Yi, J., Su, D., Gao, Y., Hsieh, C.-J., and Daniel, L. Evaluating the robustness of neural networks: An extreme value theory approach. In *International Conference on Learning Representations (ICLR)*, 2018.

Wu, B., Chen, J., Cai, D., He, X., and Gu, Q. Do wider neural networks really help adversarial robustness? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Wu, Y., Liu, F., Chrysos, G., and Cevher, V. On the convergence of encoder-only shallow transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S., and Kumar, S. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations (ICLR)*, 2020.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A. Additional notation

Given a sub-Gaussian random variable, let $\|X\|_{\psi_2} = \inf\{t > 0 \ : \ \mathbb{E}[\exp(X^2/t^2)] \leq 2\}$, see Section 2.5 of Vershynin (2018). Given a sub-exponential random variable X, let $\|X\|_{\psi_1} = \inf\{t > 0 \ : \ \mathbb{E}[\exp(|X|/t)] \leq 2\}$, see Section 2.7 of Vershynin (2018)). We recall the property that, if X and Y are scalar random variables, then $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}$, see Lemma 2.7.7 of Vershynin (2018).

We use the term *standard Gaussian vector* in $\mathbb{R}^d$ to indicate a vector ρ such that $\rho_i \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$. We recall that the maximum of n Gaussian (not necessarily independent) random variables is smaller than $\log n$ with probability at least $1 - \exp(c \log^2 n)$, see, e.g., Section 1.4 of (Rigollet & Hütter, 2023).

Given a vector $u \in \mathbb{R}^d$, we denote by u_i its i-th component. Given a matrix $A \in \mathbb{R}^{n \times d}$, we denote by $[A]_{i:}$ its i-th row, by $[A]_{:j}$ its j-th column, and by $[A]_{ij}$ its entry at position (i,j).

We say that a random variable or vector respects the Lipschitz concentration property if there exists an absolute constant $c > 0$ such that, for every Lipschitz continuous function $\varphi : \mathbb{R}^d \to \mathbb{R}$, we have $\mathbb{E}|\varphi(X)| < +\infty$ and for all $t > 0$,

$$\mathbb{P}\left(|\varphi(x) - \mathbb{E}_X[\varphi(x)]| > t\right) \leq 2e^{-ct^2/\|\varphi\|_{\text{Lip}}^2}. \tag{27}$$

The family of Lipschitz concentrated distributions covers a number of important cases, e.g., standard Gaussian (Vershynin, 2018), uniform on the sphere and on the unit (binary or continuous) hypercube (Vershynin, 2018), or data obtained via a Generative Adversarial Network (GAN) (Seddik et al., 2020).

B. Proofs for random features

In this section, we provide the proofs for our results on the random features model. Thus, we will drop the sub-script "RF" in all the quantities of this section, for the sake of a cleaner notation. We consider a single textual data-point $X = [x_1, x_2, ..., x_n]^\top \in \mathbb{R}^{n \times d}$ that satisfies Assumption 3.1. We consider the random features model defined in (1), i.e.,

$$\varphi(X) = \phi(V \operatorname{flat}(X)), \tag{28}$$

where $V_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/D)$, $D = nd$ and ϕ is the activation function, applied component-wise to the pre-activations $V \operatorname{flat}(X)$.

Further in the section, we will investigate the generalization capabilities of the RF model $f(\cdot, \theta) := \varphi(\cdot)^\top \theta$ on token modification. We use the notation $(\mathcal{X}, \mathcal{Y})$ to indicate the original labelled training dataset, with $\mathcal{X} = (X_1, \ldots, X_N)$, $X_i \in \mathbb{R}^{n \times d}$ and $\mathcal{Y} = [y_1, \ldots, y_N]^\top \in \{-1, +1\}^N$. We will use the short-hand $\Phi := [\varphi(X_1), ...\varphi(X_N)]^\top \in \mathbb{R}^{N \times p}$ for the feature matrix and $K := \Phi\Phi^\top$ for the kernel. According to (12), minimizing the quadratic loss over this dataset returns the parameters

$$\theta^* = \theta_0 + \Phi^+(\mathcal{Y} - \Phi\theta_0). \tag{29}$$

We consider the test error on the modified sample $X^i(\Delta)$ given by (see also (16))

$$\operatorname{Err}(X^i(\Delta), \theta) := \left(f(X^i(\Delta), \theta) - y_\Delta\right)^2. \tag{30}$$

We will investigate this quantity for both the fine-tuned and the re-trained model. In particular, the new solution obtained after fine-tuning over the sample (X, y) gives

$$\theta_f^* = \theta^* + \frac{\varphi(X)}{\|\varphi(X)\|_2^2}\left(y - \varphi(X)^\top \theta^*\right), \tag{31}$$

while retraining the model with initialization θ_0 on the new dataset $(\mathcal{X}_r, \mathcal{Y}_r)$ returns

$$\theta_r^* = \theta_0 + \Phi_r^+(\mathcal{Y}_r - \Phi_r\theta_0), \tag{32}$$

where we denote by $\Phi_r := [\varphi(X_1), ...\varphi(X_N), \varphi(X)]^\top \in \mathbb{R}^{(N+1) \times p}$ the new feature matrix. The corresponding kernel is denoted by $K_r = \Phi_r \Phi_r^\top$.

The outline of this section is the following:

1. We use Lemma B.1 to upper bound the numerator of the word sensitivity $\mathcal{S}(X)$ defined in (4), which readily allows to prove the desired result Theorem 4.1.

2. We prove Theorem 6.2, where we lower bound $\mathrm{Err}(X^i(\Delta), \theta_f^*)$ for the fine-tuned solution θ_f^* as a function of γ.

3. We prove Theorem 6.3, where we lower bound $\mathrm{Err}(X^i(\Delta), \theta_r^*)$ for the retrained solution θ_r^* as a function of γ. This result requires additional assumptions and analysis:

   - We report in our notation Lemma 4.1 from (Bombari & Mondelli, 2024), which defines the *feature alignment* $\mathcal{F}(X, X^i(\Delta))$ between the two samples X and $X^i(\Delta)$.

   - In Lemma B.2, we exploit the additional assumptions to prove a lower bound on the smallest eigenvalue of the kernel $\lambda_{\min}(K_r) = \Omega(k)$.

   - In Lemma B.3, we show that $\left|\mathcal{F}(X, X^i(\Delta)) - 1\right| = \mathcal{O}(1/\sqrt{n})$ with high probability. This step is useful as here this term assumes the role previously taken by $\mathcal{S}(X)$.

Lemma B.1. *Let $\varphi(X)$ be the random feature map defined in* (1)*, with ϕ Lipschitz and $k = \Omega(D)$. Let $\Delta \in \mathbb{R}^d$ be such that $\|\Delta\|_2 \le \sqrt{d}$. Then, for every $i \in [n]$, we have*

$$\left\|\varphi(X^i(\Delta)) - \varphi(X)\right\|_2 = \mathcal{O}\left(\sqrt{\frac{k}{n}}\right), \tag{33}$$

with probability at least $1 - \exp(-cD)$ over V.

Proof. Let's condition on the event

$$\|V\|_{\mathrm{op}} = \mathcal{O}\left(\sqrt{\frac{D+k}{D}}\right), \tag{34}$$

which happens with probability at least $1 - \exp(-c_1 D)$ over V, by Theorem 4.4.5 of (Vershynin, 2018). Thus, for every i, we have

$$\begin{aligned}
\left\|\varphi(X^i(\Delta)) - \varphi(X)\right\|_2 &= \left\|\phi(V \operatorname{flat}(X^i(\Delta))) - \phi(V \operatorname{flat}(X))\right\|_2 \\
&\le M \left\|V(\operatorname{flat}(X^i(\Delta)) - \operatorname{flat}(X))\right\|_2 \\
&\le M \|V\|_{\mathrm{op}} \|\Delta\|_2 \\
&\le C_1 \sqrt{\frac{D+k}{D}} \sqrt{d} \\
&= C_1 \sqrt{\frac{D+k}{n}} \\
&\le C_2 \sqrt{\frac{k}{n}},
\end{aligned} \tag{35}$$

where the first inequality comes from the Lipschitz continuity of ϕ, and the last step is a consequence of $k = \Omega(D)$. $\qquad\square$

Theorem 4.1 Let $\varphi(X)$ be the random feature map defined in (1), where ϕ is Lipschitz and not identically 0. Let $X \in \mathbb{R}^{n \times d}$ be a generic input sample s.t. Assumption 3.1 holds, and assume $k = \Omega(D)$. Let $\mathcal{S}(X)$ denote the the word sensitivity defined in (4). Then, we have

$$\mathcal{S}(X) = \mathcal{O}\left(1/\sqrt{n}\right) = o(1), \tag{36}$$

with probability at least $1 - \exp(-cD)$ over V.

Proof. As ϕ is Lipschitz and non-0, we can apply the result in Lemma C.3 of (Bombari & Mondelli, 2024), getting

$$\|\varphi(X)\|_2 = \Theta(\sqrt{k}), \tag{37}$$

with probability at least $1 - \exp(-c_1 D)$ over V. Thus, the thesis readily follows from Lemma B.1. $\qquad\square$

Proof of Remark 4.2. The only difference with respect to the argument for Theorem 4.1 is in (35). Now, Δ is replaced by a new set of m perturbations $\Delta_1, \ldots, \Delta_m$. Thus, the modified context takes the form

$$X(\Delta_1, \ldots, \Delta_m) = X + \sum_{j=1}^{m} e_{i_j} \Delta_j^\top, \tag{38}$$

where $\{e_{i_j}\}_{j \in [m]}$ represent different elements of the canonical basis. Thus, we have

$$\|\text{flat}(X(\Delta_1, \ldots, \Delta_m)) - \text{flat}(X)\|_2 = \left\| \text{flat}\left(\sum_{j=1}^{m} e_{i_j} \Delta_j^\top \right) \right\|_2 = \left\| \sum_{j=1}^{m} e_{i_j} \Delta_j^\top \right\|_F. \tag{39}$$

Since the e_{i_j}'s are all distinct (as we are modifying m different words), we obtain

$$\left\| \sum_{j=1}^{m} e_{i_j} \Delta_j^\top \right\|_F^2 = \sum_{j=1}^{m} \|\Delta_j\|_2^2 \leq md, \tag{40}$$

where in the last step we use that $\|\Delta_j\|_2 \leq \sqrt{d}$ for all $j \in [m]$.

This allows to replace the $\sqrt{d}$ in the fourth line of (35) with $\sqrt{md}$, increasing the final bound in the statement of Lemma B.1 by a factor $\sqrt{m}$. Thus, the upper bound on the sensitivity is given by $\sqrt{m/n} = o(1)$, which concludes the proof. $\qquad \square$

Theorem 6.2 Let $f(\cdot, \theta_f^*)$ be the RF model fine-tuned on the sample (X, y), where ϕ in (1) is Lipschitz and not identically 0, $X \in \mathbb{R}^{n \times d}$ is a generic sample s.t. Assumption 3.1 holds and θ_f^* is defined in (14). Assume $k = \Omega(D)$, $|f(X, \theta^*)| = \mathcal{O}(1)$, and that Assumption 6.1 holds with $\gamma \in [0, 2)$. Let $\text{Err}(X^i(\Delta), \theta_f^*)$ be the test error of the model on the sample $X^i(\Delta)$ defined in (16). Then, for any Δ such that $\|\Delta\|_2 \leq \sqrt{d}$ and any $i \in [n]$, we have

$$\text{Err}(X^i(\Delta), \theta_f^*) > (2 - \gamma)^2 - \mathcal{O}\left(1/\sqrt{n}\right), \tag{41}$$

with probability at least $1 - \exp(-cD)$ over V.

Proof. We have

$$f(X^i(\Delta), \theta_f^*) = \varphi(X^i(\Delta))^\top \theta_f^* = \varphi(X^i(\Delta))^\top \theta^* + \frac{\varphi(X^i(\Delta))^\top \varphi(X)}{\|\varphi(X)\|_2^2} \left(y - \varphi(X)^\top \theta^*\right), \tag{42}$$

where the second step is justified by (14). As $\left|\varphi(X)^\top \theta^* - \varphi(X^i(\Delta))^\top \theta^*\right| = \left|f(X, \theta^*) - f(X^i(\Delta), \theta^*)\right| \leq \gamma$ by Assumption 6.1, we can write

$$\left|f(X^i(\Delta), \theta_f^*) - y\right| = \left| f(X^i(\Delta), \theta^*) - f(X, \theta^*) + \left(\frac{\varphi(X^i(\Delta))^\top \varphi(X)}{\|\varphi(X)\|_2^2} - 1 \right)(y - f(X, \theta^*)) \right|$$
$$\leq \gamma + \left| \frac{\varphi(X^i(\Delta))^\top \varphi(X)}{\|\varphi(X)\|_2^2} - 1 \right| |y - f(X, \theta^*)|. \tag{43}$$

By Cauchy-Schwartz inequality, we have

$$\left| \frac{\varphi(X^i(\Delta))^\top \varphi(X)}{\|\varphi(X)\|_2^2} - 1 \right| = \left| \frac{(\varphi(X^i(\Delta)) - \varphi(X))^\top \varphi(X)}{\|\varphi(X)\|_2^2} \right| \leq \frac{\|\varphi(X^i(\Delta)) - \varphi(X)\|_2 \|\varphi(X)\|_2}{\|\varphi(X)\|_2^2} \leq \mathcal{S}(X). \tag{44}$$

By Theorem 4.1, we have that $\mathcal{S}(X) = \mathcal{O}(1/\sqrt{n})$ with probability at least $1 - \exp(-cD)$ over V. Conditioning on such high probability event, we can write

$$\begin{aligned} \left|f(X^i(\Delta), \theta_f^*) - y_\Delta\right| &\geq |y_\Delta - y| - \left|f(X^i(\Delta), \theta_f^*) - y\right| \\ &\geq |y_\Delta - y| - \gamma - \mathcal{S}(X) |y - f(X, \theta^*)| \\ &= 2 - \gamma - S(X) |y - f(X, \theta^*)| \\ &= 2 - \gamma - \mathcal{O}\left(1/\sqrt{n}\right), \end{aligned} \tag{45}$$

15

where the third step is a consequence of $|y| = |y_\Delta| = 1$, with $y = -y_\Delta$, and the fourth step comes from $|f(X, \theta^*)| = \mathcal{O}(1)$. Thus, we can conclude

$$\mathrm{Err}(X^i(\Delta), \theta_f^*) = \left(f(X^i(\Delta), \theta_f^*) - y_\Delta\right)^2 \geq \left(2 - \gamma - \mathcal{O}\left(1/\sqrt{n}\right)\right)^2 > (2 - \gamma)^2 - \mathcal{O}\left(1/\sqrt{n}\right). \tag{46}$$

$\square$

Lemma 4.1 (Bombari & Mondelli, 2024) Let the kernel $K_r \in \mathbb{R}^{(N+1)\times(N+1)}$ be invertible, and let $P_\Phi \in \mathbb{R}^{k \times k}$ be the projector over $\mathrm{Span}\{\mathrm{rows}(\Phi)\}$. Let us denote by

$$\mathcal{F}(X, X^i(\Delta)) := \frac{\varphi(X^i(\Delta))^\top P_\Phi^\perp \varphi(X)}{\left\|P_\Phi^\perp \varphi(X)\right\|_2^2} \tag{47}$$

the *feature alignment* between X and $X^i(\Delta)$. Then, we have

$$f(X^i(\Delta), \theta_r^*) - f(X^i(\Delta), \theta^*) = \mathcal{F}(X, X^i(\Delta)) \left(f(X, \theta_r^*) - f(X, \theta^*)\right). \tag{48}$$

Notice that

$$\left\|P_\Phi^\perp \varphi(X)\right\|_2^2 \geq \lambda_{\min}(K_r) > 0 \tag{49}$$

is directly implied by the invertibility of K_r, as shown in Lemma B.1 from (Bombari & Mondelli, 2024).

Lemma B.2. *Let ϕ be a non-linear, Lipschitz function. Let all the training data in $\mathcal{X}_r$ be sampled i.i.d. according to a distribution $\mathcal{P}_X$ s.t. $\mathbb{E}_{X \sim \mathcal{P}_X}[X] = 0$, Assumption 3.1 holds, and the Lipschitz concentration property is satisfied. Let $N \log^3 N = o(k)$ and $N \log^4 N = o(D^2)$. Then, we have*

$$\lambda_{\min}(K_r) = \Omega(k), \tag{50}$$

with probability at least $1 - \exp\left(-c \log^2 N\right)$ over V and $\mathcal{X}_r$.

Proof. The desired result follows from Lemma D.2 in (Bombari & Mondelli, 2024). Notice that for their argument to go through, they need their Lemma D.1 to hold, which requires our assumptions on the data distribution $\mathcal{P}_X$, and Assumption 3.1. They further require the scalings $N = o(D^2/\log^4 D)$ and $N \log^4 N = o(D^2)$, which are both given by our assumption $N \log^4 N = o(D^2)$. Finally, in their Lemma D.2 they require the activation function ϕ to be Lipschitz and non-linear, and the over-parameterized setting $N \log^3 N = o(k)$. $\square$

Lemma B.3. *Let ϕ be a non-linear, Lipschitz function. Let all the training data in $\mathcal{X}_r$ be sampled i.i.d. according to a distribution s.t. $\mathbb{E}_{X \sim \mathcal{P}_X}[X] = 0$, Assumption 3.1 holds, and the Lipschitz concentration property is satisfied. Let $N \log^3 N = o(k)$, $N \log^4 N = o(D^2)$ and $k = \Omega(D)$. Let $\mathcal{F}(X, X^i(\Delta))$ be defined as in (47). Then, we have*

$$\left|\mathcal{F}(X, X^i(\Delta)) - 1\right| = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right), \tag{51}$$

with probability at least $1 - \exp\left(-c \log^2 N\right)$ over $\mathcal{X}_r$ and V.

Proof. By Cauchy-Schwartz inequality, we have

$$\begin{aligned}
\left|\mathcal{F}(X, X^i(\Delta)) - 1\right| &= \left|\frac{\left(\varphi(X^i(\Delta)) - \varphi(X)\right)^\top P_\Phi^\perp \varphi(X)}{\left\|P_\Phi^\perp \varphi(X)\right\|_2^2}\right| \\
&\leq \frac{\left\|\varphi(X^i(\Delta)) - \varphi(X)\right\|_2}{\left\|\varphi(X)\right\|_2} \frac{\left\|\varphi(X)\right\|_2}{\left\|P_\Phi^\perp \varphi(X)\right\|_2} \\
&\leq \mathcal{S}(X) \frac{\left\|\varphi(X)\right\|_2}{\sqrt{\lambda_{\min}(K_r)}},
\end{aligned} \tag{52}$$

where the last step is a consequence of (49). As ϕ is Lipschitz and non-0, we can apply the result in Lemma C.3 of (Bombari & Mondelli, 2024), getting

$$\|\varphi(X)\|_2 = \Theta(\sqrt{k}), \tag{53}$$

with probability at least $1 - \exp(-c_1 D)$ over V. Due to Lemma B.2, we can also write

$$\lambda_{\min}(K_r) = \Omega(k), \tag{54}$$

with probability at least $1 - \exp\left(-c_2 \log^2 N\right)$ over V and $\mathcal{X}_r$. Thus, (52) promptly gives

$$\left|\mathcal{F}(X, X^i(\Delta)) - 1\right| \leq \mathcal{S}(X)\frac{\|\varphi(X)\|_2}{\sqrt{\lambda_{\min}(K_r)}} = \mathcal{O}\left(\mathcal{S}(X)\right) = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right), \tag{55}$$

where the last step comes from Theorem 4.1, and holds with probability at least $1 - \exp(-c_3 D)$. This gives the desired result. $\qquad\square$

Theorem 6.3 Let $f(\cdot, \theta_r^*)$ be the RF model re-trained on the dataset $(\mathcal{X}_r, \mathcal{Y}_r)$ that contains the pair (X, y), thus with θ_r^* defined in (15). Let ϕ in (1) be a non-linear, Lipschitz function. Assume the training data to be sampled i.i.d. from a distribution $\mathcal{P}_X$ s.t. $\mathbb{E}_{X \sim \mathcal{P}_X}[X] = 0$, Assumption 3.1 holds, and the Lipschitz concentration property is satisfied. Let $N \log^3 N = o(k)$, $N \log^4 N = o(D^2)$ and $k = \Omega(D)$. Assume that $|f(X, \theta^*)| = \mathcal{O}(1)$ and that Assumption 6.1 holds with $\gamma \in [0, 2)$. Let $\text{Err}(X^i(\Delta^*), \theta_r^*)$ be the test error of the model on the sample $X^i(\Delta)$ defined in (16). Then, for any Δ s.t. $\|\Delta\|_2 \leq \sqrt{d}$ and any $i \in [n]$, we have

$$\text{Err}(X^i(\Delta^*), \theta_r^*) > (2 - \gamma)^2 - \mathcal{O}\left(1/\sqrt{n}\right), \tag{56}$$

with probability at least $1 - \exp\left(-c \log^2 N\right)$ over V and $\mathcal{X}_r$.

Proof. Let's condition on K_r being invertible, which by Lemma B.2 happens with probability at least $1 - \exp\left(-c_1 \log^2 N\right)$ over V and $\mathcal{X}_r$. Then, by (48), we have

$$f(X^i(\Delta), \theta_r^*) - f(X^i(\Delta), \theta^*) = \mathcal{F}(X, X^i(\Delta))\left(y - f(X, \theta^*)\right), \tag{57}$$

since $f(\cdot, \theta_r^*)$ fully interpolates the training data, thus giving $f(X, \theta_r^*) = y$. As $\left|\varphi(X)^\top \theta^* - \varphi(X^i(\Delta))^\top \theta^*\right| = \left|f(X, \theta^*) - f(X^i(\Delta), \theta^*)\right| \leq \gamma$ by Assumption 6.1, we can write

$$\begin{aligned}
\left|f(X^i(\Delta), \theta_r^*) - y\right| &= \left|f(X^i(\Delta), \theta^*) - f(X, \theta^*) + \left(\mathcal{F}(X, X^i(\Delta)) - 1\right)\left(y - f(X, \theta^*)\right)\right| \\
&\leq \gamma + \left|\mathcal{F}(X, X^i(\Delta)) - 1\right| \left|y - f(X, \theta^*)\right|.
\end{aligned} \tag{58}$$

By Lemma B.3 we have that $\left|\mathcal{F}(X, X^i(\Delta)) - 1\right| = \mathcal{O}\left(1/\sqrt{n}\right)$ with probability at least $1 - \exp\left(-c \log^2 N\right)$ over $\mathcal{X}_r$ and V. Conditioning on such high probability event, we can write

$$\begin{aligned}
\left|f(X^i(\Delta), \theta_r^*) - y_\Delta\right| &\geq |y_\Delta - y| - \left|f(X^i(\Delta), \theta_r^*) - y\right| \\
&\geq |y_\Delta - y| - \gamma - \left|\mathcal{F}(X, X^i(\Delta)) - 1\right| \left|y - f(X, \theta^*)\right| \\
&= 2 - \gamma - \left|\mathcal{F}(X, X^i(\Delta)) - 1\right| \left|y - f(X, \theta^*)\right| \\
&= 2 - \gamma - \mathcal{O}\left(1/\sqrt{n}\right),
\end{aligned} \tag{59}$$

where the third step is a consequence of $|y| = |y_\Delta| = 1$, with $y = -y_\Delta$, and the fourth step comes from $|f(X, \theta^*)| = \mathcal{O}(1)$. Thus, we can conclude

$$\text{Err}(X^i(\Delta), \theta_r^*) = \left(f(X^i(\Delta), \theta_r^*) - y_\Delta\right)^2 \geq \left(2 - \gamma - \mathcal{O}\left(1/\sqrt{n}\right)\right)^2 > (2 - \gamma)^2 - \mathcal{O}\left(1/\sqrt{n}\right). \tag{60}$$

$\qquad\square$

C. Proofs for deep random features

In this section, we provide the proofs for our results on the deep random features model. We consider a single textual data-point $X = [x_1, x_2, ..., x_n]^\top \in \mathbb{R}^{n \times d}$ that satisfies Assumption 3.1. We consider the deep random features model defined in (6), i.e.,

$$\varphi_{\text{DRF}}(X) := \phi(V_L \phi(V_{L-1}(...V_2 \phi(V_1 \text{ flat } X)...))), \tag{61}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is the non-linearity applied component-wise at each layer, and $V_l \in \mathbb{R}^{D \times D}$ are the random weights at layer l, sampled independently and such that $[V_1]_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta/D)$ and $[V_l]_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta/k)$ for $l > 1$. We set β according to He's (or Kaiming) initialization (He et al., 2015), i.e.,

$$\mathbb{E}_{\rho \sim \mathcal{N}(0, \beta)}\left[\phi^2(\rho)\right] = 1. \tag{62}$$

Thus, we require the activation function ϕ to guarantee at least one value β for which the previous equation is respected. We will consider β to be a positive constant dependent only on the activation ϕ.

Let's introduce the shorthands

$$\begin{aligned}
\varphi_0(X) &= \text{flat}(X), \\
\varphi_l(X) &= \phi\left(V_l \, \varphi_{l-1}(X)\right), \qquad \text{for } l \in [L].
\end{aligned} \tag{63}$$

The outline of this section is the following:

1. In Lemma C.1, we prove that at every layer, the norm of the features $\|\varphi_l(X)\|_2$, with $l > 1$, concentrates to $\sqrt{k}$. This is the step where He's initialization is necessary.

2. In Lemma C.2, we show that $\left\|\varphi_l(X^i(\Delta)) - \varphi_l(X)\right\|_2$ can be upper bounded by a term that grows exponentially with the depth of the layer l.

3. In Theorem 5.1, we upper bound $\mathcal{S}_{\text{DRF}}$, concluding the argument.

Lemma C.1. *Let $\varphi_l(X)$ be defined in (63), and let ϕ be a Lipschitz function such that (7) admits at least one solution β. Let $X \in \mathbb{R}^{n \times d}$ be a generic input sample such that Assumption 3.1 holds, and let $k = \Theta(D)$. Then, for every $l > 0$, we have*

$$\left|\|\varphi_l(X)\|_2 - \sqrt{k}\right| \leq e^{Cl} \log D, \tag{64}$$

with probability at least $1 - 2L \exp(-c \log^2 D)$ over $\{V_l\}_{l=1}^L$.

Proof. By Assumption 3.1 the statement trivially holds for $l = 0$, if we replace $\sqrt{k}$ with $\sqrt{d}$. Let's consider this the base case of an induction argument to prove the statement for $l \in [L]$.

Thus, using the notation $k_0 = D$ and $k_l = k$ for $l \in [L]$, the inductive hypothesis becomes

$$\left|\|\varphi_{l-1}(X)\|_2 - \sqrt{k_{l-1}}\right| \leq e^{C(l-1)} \log D, \tag{65}$$

with probability at least $1 - 2(l-1) \exp(-c \log^2 D)$ on $\{V_m\}_{m=1}^{l-1}$. Let's condition on this high probability event and on $\|V_l\|_{\text{op}} \leq C_1$ until the end of the proof. By Theorem 4.4.5 of (Vershynin, 2018) and since $k = \Theta(D)$, there exists C_1 large enough and independent from l, such that this holds with probability at least $1 - 2 \exp(-c_1 k)$ over V_l. We therefore aim to prove the thesis for

$$e^C := \max\left(1, MC_1 + 1\right), \tag{66}$$

where we denote by M the Lipschitz constant of ϕ. (66) explicitly shows that C is a natural constant dependent only on the activation function ϕ.

We have

$$\varphi_l(X) = \phi(V_l \, \varphi_{l-1}(X)) = \phi\left(V_l \frac{\sqrt{k_{l-1}} \, \varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2} + V_l\left(\varphi_{l-1}(X) - \frac{\sqrt{k_{l-1}} \, \varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2}\right)\right), \tag{67}$$

which gives

$$\left\|\varphi_l(X) - \phi\left(V_l \frac{\sqrt{k_{l-1}} \, \varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2}\right)\right\|_2 \leq MC_1 \left|\|\varphi_{l-1}(X)\|_2 - \sqrt{k_{l-1}}\right| \leq MC_1 e^{C(l-1)} \log D, \tag{68}$$

where the first step is true since ϕ is M-Lipschitz, and the last step is a direct consequence of the inductive hypothesis (65).

Let's now consider the second term in the left hand side of (68). Let's define the shorthand $\rho = V_l \frac{\sqrt{k_{l-1}}\, \varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2} \in \mathbb{R}^D$. In the probability space of V_l, ρ is distributed as a Gaussian random vector, such that all its entries ρ_i are i.i.d. Gaussian with variance β. Thus, we have

$$\mathbb{E}_{V_l}\left[\left\|\phi\left(V_l \frac{\sqrt{k_{l-1}}\, \varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2}\right)\right\|_2^2\right] = \mathbb{E}_\rho\left[\|\phi(\rho)\|_2^2\right] = k_l \mathbb{E}_{\rho_1}\left[\phi^2(\rho_1)\right] = k, \tag{69}$$

where the last step is a consequence of (62), and of $k_l = k$ for $l \in [L]$.

As the ρ_i's are independent and ϕ is Lipschitz, the random variables $\left(\phi^2(\rho_i) - 1\right)$ are independent, mean-0, and sub-exponential, such that $\left\|\phi^2(\rho_i) - 1\right\|_{\psi_1} \leq C_2$. Thus, by Bernstein inequality (cf. Theorem 2.8.1. in (Vershynin, 2018)), we have

$$\mathbb{P}_\rho\left(\left|\sum_{i=1}^k \left(\phi^2(\rho_i) - 1\right)\right| \geq \sqrt{k}\log k\right) \leq 2\exp(-c_2 \log^2 k), \tag{70}$$

which gives

$$\left|\left\|\phi\left(V_l \frac{\sqrt{k}\, \varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2}\right)\right\|_2^2 - k\right| \leq \sqrt{k}\log k, \tag{71}$$

with probability at least $1 - 2\exp(-c_2 \log^2 k)$ over V_l. We will condition on such high probability event until the end of the proof.

Thus, we have

$$\left\|\phi\left(V_l \frac{\sqrt{k}\, \varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2}\right)\right\|_2 \leq \sqrt{k + \sqrt{k}\log k} = \sqrt{k}\sqrt{1 + \frac{\log k}{\sqrt{k}}} \leq \sqrt{k}\left(1 + \frac{\log k}{\sqrt{k}}\right) = \sqrt{k} + \log k, \tag{72}$$

and

$$\left\|\phi\left(V_l \frac{\sqrt{k}\, \varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2}\right)\right\|_2 \geq \sqrt{k - \sqrt{k}\log k} = \sqrt{k}\sqrt{1 - \frac{\log k}{\sqrt{k}}} \geq \sqrt{k}\left(1 - \frac{\log k}{\sqrt{k}}\right) = \sqrt{k} - \log k. \tag{73}$$

Putting together the last two equations gives

$$\left|\left\|\phi\left(V_l \frac{\sqrt{k}\, \varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2}\right)\right\|_2 - \sqrt{k}\right| \leq \log k. \tag{74}$$

Applying (68), (74) and the triangle inequality gives

$$\left|\|\varphi_l(X)\|_2 - \sqrt{k}\right| \leq MC_1 e^{C(l-1)}\log k + \log k \leq \left(MC_1 e^{C(l-1)} + e^{C(l-1)}\right)\log k \leq e^{Cl}\log k, \tag{75}$$

where the second and the third step are both consequences of (66). This inequality, performing a union bound on the high probability events we considered so far, holds with probability at least $1 - 2(l-1)\exp(-c\log^2 k) - 2\exp(-c_1 k) - 2\exp(-c_2 \log^2 k) \geq 1 - 2l\exp(-c\log^2 k)$ over $\{V_m\}_{m=1}^l$, as soon as we consider $c = \min(c_1, c_2)$. $\qquad\square$

Lemma C.2. *Let $\varphi_l(X)$ be defined in (63), let ϕ be a Lipschitz function and $X \in \mathbb{R}^{n \times d}$ a generic input sample such that Assumption 3.1 holds. Assume $k = \Theta(D)$. Then, for every $l \geq 0$, for every $i \in [n]$ and for any $\Delta \in \mathbb{R}^d$ such that $\|\Delta\|_2 \leq \sqrt{d}$, we have*

$$\left\|\varphi_l(X^i(\Delta)) - \varphi_l(X)\right\|_2 \leq \sqrt{d}e^{Cl}, \tag{76}$$

with probability at least $1 - 2L\exp(-ck)$ over $\{V_l\}_{l=1}^L$.

Proof. Let's prove the statement by induction over l. The base case $l = 0$ is a direct consequence of $\left\|\text{flat}(X^i(\Delta)) - \text{flat}(X)\right\|_2 = \|\Delta\|_2 \leq \sqrt{d}$, which makes the thesis true for any $C \geq 0$.

In the inductive step, the inductive hypothesis becomes

$$\left\|\varphi_{l-1}(X^i(\Delta)) - \varphi_{l-1}(X)\right\|_2 \leq \sqrt{d}e^{C(l-1)}, \tag{77}$$

with probability at least $1 - 2(l-1)\exp(-ck)$ on $\{V_m\}_{m=1}^{l-1}$. Let's condition on this high probability event and on $\|V_l\|_{\text{op}} \leq C_1$ until the end of the proof. By Theorem 4.4.5 of (Vershynin, 2018) and since $k = \Theta(D)$, there exists C_1 large enough and independent from l, such that this holds with probability at least $1 - 2\exp(-c_1 k)$ over V_l.

We aim to prove the thesis for $e^C = MC_1$, and $c = c_1$, where M is the Lipschitz constant of ϕ. We have

$$\begin{aligned}
\left\|\varphi_l(X^i(\Delta)) - \varphi_l(X)\right\|_2 &= \left\|\phi(V_l\varphi_{l-1}(X^i(\Delta))) - \phi(V_l\varphi_{l-1}(X))\right\|_2 \\
&\leq MC_1 \left\|\varphi_{l-1}(X^i(\Delta)) - \varphi_{l-1}(X)\right\|_2 \\
&\leq MC_1\sqrt{d}e^{C(l-1)} \\
&\leq \sqrt{d}e^{Cl},
\end{aligned} \tag{78}$$

with probability at least $1 - 2(l-1)\exp(-ck) - 2\exp(-c_1 k) = 1 - 2l\exp(-ck)$ over $\{V_m\}_{m=1}^{l}$, which gives the thesis. $\square$

Theorem 4.3. Let $\varphi_{\text{DRF}}(X)$ be the deep random feature map defined in (6), and let ϕ be a Lipschitz function, such that (7) admits at least one solution β. Let $X \in \mathbb{R}^{n \times d}$ be a generic input sample s.t. Assumption 3.1 holds, and assume $k = \Theta(D)$, $L = o(\log k)$. Let $\mathcal{S}_{\text{DRF}}(X)$ be the the word sensitivity defined in (4). Then, we have

$$\mathcal{S}_{\text{DRF}} = \mathcal{O}\left(\frac{e^{CL}}{\sqrt{n}}\right), \tag{79}$$

with probability at least $1 - \exp(-c\log^2 k)$ over $\{V_m\}_{m=1}^{l}$.

Proof. By Lemma C.1 (for the case $l = L$) and since $L = o(\log k)$, we have

$$\|\varphi_{\text{DRF}}(X)\|_2 = \Theta(\sqrt{k}), \tag{80}$$

with probability at least $1 - 2L\exp(-c_1\log^2 k) \geq 1 - \exp(-c_2\log^2 k)$ over $\{V_l\}_{l=1}^{L}$.

By Lemma C.2 (for the case $l = L$), we have that, for every $i \in [n]$, we have

$$\sup_{\|\Delta\|_2 \leq \sqrt{d}} \left\|\varphi_{\text{DRF}}(X^i(\Delta)) - \varphi_{\text{DRF}}(X)\right\|_2 \leq \sqrt{d}e^{CL}, \tag{81}$$

with probability at least $1 - 2L\exp(-c_3 k) \geq 1 - \exp(-c_4\log^2 k)$ over $\{V_l\}_{l=1}^{L}$.

Then, putting together (80), (81), and (4), and recalling that $k = \Theta(D)$, we get the thesis. $\square$

D. Proofs for random attention features

In this section, we provide the proofs for our results on the random attention features model.

We consider a single textual data-point $X = [x_1, x_2, ..., x_n]^\top \in \mathbb{R}^{n \times d}$ that satisfies Assumptions 3.1. We assume $d/\log^4 d = \Omega(n)$. We consider the random attention features model defined in (3), i.e.,

$$\varphi_{\text{RAF}}(X) = \text{softmax}\left(\frac{XWX^\top}{\sqrt{d}}\right)X, \tag{82}$$

where $W_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$ sampled independently from X. We define the argument of the softmax as $S(X) \in \mathbb{R}^{n \times n}$ given by

$$S(X) = \frac{XWX^\top}{\sqrt{d}}, \tag{83}$$

and its evaluation on the perturbed sample as

$$S(X^i(\Delta)) = \frac{X^i(\Delta)W X^i(\Delta)^\top}{\sqrt{d}}. \tag{84}$$

The attention scores are therefore defined as the row-wise $\mathrm{softmax}$ of the previous term, i.e.,

$$[s(X)]_{j:} = \mathrm{softmax}\left([S(X)]_{j:}\right), \tag{85}$$

which allows us to rewrite (3) as

$$\varphi_{\mathrm{RAF}}(X) = s(X)X. \tag{86}$$

Through this section, we will always consider this notation and assumptions, which will therefore not be repeated in the statement of the lemmas before our final result Theorem 5.1.

The outline of this section is the following:

1. In Lemma D.1, we prove the existence of a vector $\delta^* \in \mathbb{R}^d$, such that its inner product with the token embeddings x_i's is large for a constant fraction of the words in the context.

2. In Lemma D.2, we show that the result in the previous Lemma can be extended to two different vectors δ_1^* and δ_2^*, that are very different from each other, i.e., $\|\delta_1^* - \delta_2^*\|_2 = \sqrt{d}$. The reason why we would like this statement to hold on two vectors will be clear in Lemma D.5.

3. In Lemma D.3, exploiting the properties of the Gaussian attention features W, we show that there exist two different vectors Δ_1^* and Δ_2^*, that are very different from each other, i.e., $\|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d})$, such that a constant fraction of the entries of the vector $XW\Delta_k^*/\sqrt{d}$, $k \in \{1, 2\}$, is large.

4. In Lemma D.4, we prove that, for the two vectors Δ_1^* and Δ_2^* and an $\Omega(n)$ number of rows $j \in [n]$, the attention scores $[s(X^i(\Delta_k^*))]_{j:}^\top$ are well approximated by the canonical basis vector e_i for $k \in \{1, 2\}$. This intuitively means that a constant fraction of tokens x_j puts all their attention towards the i-th modified token $x_i + \Delta_k^*$.

5. In Lemma D.5, using an argument by contradiction, we prove that at least one between Δ_1^* and Δ_2^* is such that $\left\|\varphi_{\mathrm{RAF}}(X) - \varphi_{\mathrm{RAF}}(X^i(\Delta^*))\right\|_F = \Omega(\sqrt{dn})$.

6. In Theorem 5.1, we upper bound $\|\varphi_{\mathrm{RAF}}(X)\|_F$, concluding the argument.

Lemma D.1. *There exists a vector $\delta^* \in \mathbb{R}^d$, such that*

$$\|\delta^*\|_2 \le \sqrt{d}, \tag{87}$$

and

$$\left(x_i^\top \delta^*\right)^2 = \Omega\left(\frac{d^2}{n}\right), \tag{88}$$

for at least $\Omega(n)$ indices $i \in [n]$.

Proof. Let $X^\top = UDV^\top$ the singular value decomposition of $X^\top$, where $U \in \mathbb{R}^{d \times d}$, $D \in \mathbb{R}^{d \times n}$, and $V \in \mathbb{R}^{n \times n}$, and let $U_r \in \mathbb{R}^{d \times r}$ be the matrix obtained keeping only the first $r = \mathrm{rank}(X)$ columns of U. Let δ_ε be defined as follows

$$\delta_\varepsilon := \sqrt{\frac{d}{n}} U_r \varepsilon, \tag{89}$$

where we consider $\varepsilon \in \mathbb{R}^r$ uniformly distributed on the sphere $\sqrt{r}\, \mathbb{S}^{r-1}$ of radius $\sqrt{r}$. By definition we have $U_r^\top U_r = I$, which implies

$$\|U_r \varepsilon\|_2 = \|\varepsilon\|_2 = \sqrt{r}. \tag{90}$$

As $r \le n$, we have that

$$\|\delta_\varepsilon\|_2 \le \sqrt{d}, \tag{91}$$

for every ε. Let's call Z_i^ε the random variable $x_i^\top \delta_\varepsilon$.

By contradiction, let's assume the thesis to be false. Then, in particular, by (91), there is no ε such that δ_ε respects the second part of the thesis. This implies that, for every ε, there are at least $\lceil n/2 \rceil$ values of i such that $(Z_i^\varepsilon)^2 < 0.01\, d^2/n$. Let's define the indicators

$$\chi_i = \begin{cases} 1, & \text{if } (Z_i^\varepsilon)^2 < 0.01\, d^2/n, \\ 0, & \text{if } (Z_i^\varepsilon)^2 \geq 0.01\, d^2/n. \end{cases} \tag{92}$$

Thus, by contradiction, we have

$$\sum \chi_i \geq \left\lceil \frac{n}{2} \right\rceil \geq \frac{n}{2}, \tag{93}$$

for every ε. Thus, by the probabilistic method,

$$\frac{n}{2} \leq \mathbb{E}_\varepsilon \left[\sum \chi_i \right] = \sum \mathbb{E}_\varepsilon \left[\chi_i \right] = n\, \mathbb{E}_\varepsilon \left[\chi_1 \right] = n\, \mathbb{P}_\varepsilon \left((Z_1^\varepsilon)^2 \leq 0.01\, d^2/n \right), \tag{94}$$

where the second equality is true as the Z_i^ε are identically distributed, and the last step comes from the definition of the indicators in (92). This implies, for every i,

$$\mathbb{P}_\varepsilon \left((Z_i^\varepsilon)^2 \leq 0.01\, d^2/n \right) \geq 0.5. \tag{95}$$

Let $\rho \in \mathbb{R}^r$ be a standard Gaussian vector and define $\rho_r = \|\rho\|_2 / \sqrt{r}$. We have that $\mathbb{E}[\rho_r^2] = 1$ and $\text{Var}(\rho_r^2) = 1/r$. Thus, for every r, by Chebyshev's inequality, we can write

$$\mathbb{P}_{\rho_r} \left(\rho_r^2 \leq 3 \right) \geq 0.75 \geq 0.5. \tag{96}$$

By rotational invariance of the Gaussian measure, we have that $\rho_r \varepsilon$ is distributed as a standard Gaussian vector. In particular, we have

$$\rho_r Z_i^\varepsilon = \rho_r x_i^\top \delta_\varepsilon = \left(\sqrt{\frac{d}{n}} U_r^\top x_i \right)^\top (\rho_r \varepsilon). \tag{97}$$

As $U_r^\top x_i$ is a fixed vector independent from ε and ρ_r, we have that $\rho_r Z_i^\varepsilon$ is a Gaussian random variable (in the probability space of ε and ρ_r), with variance $d \left\| U_r^\top x_i \right\|_2^2 /n = d^2/n$, as $x_i \in \text{Span}\{\text{rows}(U_r^\top)\}$ by construction. Thus, $\sqrt{n} \rho_r Z_i^\varepsilon / d$ is distributed as a standard Gaussian random variable g_i.

Therefore, putting together (95) and (96) gives

$$\mathbb{P}_{g_i} \left(g_i^2 \leq 0.03 \right) \geq 0.25. \tag{98}$$

However, we can upper-bound the left hand side of the previous equation exploiting the closed form of the pdf of a standard Gaussian random variable. This gives

$$\mathbb{P}_{g_i} \left(|g_i| \leq \sqrt{0.03} \right) < \frac{1}{\sqrt{2\pi}} \cdot 2 \cdot \sqrt{0.03} < 0.14 < 0.25, \tag{99}$$

which leads to the desired contradiction. $\qquad\square$

Lemma D.2. *There exist two vectors $\delta_1^* \in \mathbb{R}^d$ and $\delta_2^* \in \mathbb{R}^d$ such that*

$$\|\delta_1^*\|_2 \leq \sqrt{d}, \qquad \|\delta_2^*\|_2 \leq \sqrt{d}, \qquad \|\delta_1^* - \delta_2^*\|_2 = \sqrt{d}, \tag{100}$$

and $x_i^\top \delta_1^* > 0$, $x_i^\top \delta_2^* > 0$, and

$$x_i^\top \delta_1^* = \Omega \left(\sqrt{d} \log^2 d \right), \tag{101}$$

$$x_i^\top \delta_2^* = \Omega \left(\sqrt{d} \log^2 d \right), \tag{102}$$

for at least $\Omega(n)$ indices $i \in [n]$.

Proof. By Lemma D.1, we have that there exists a vector δ^*, with $\|\delta^*\|_2 \leq \sqrt{d}$ such that

$$\left(x_i^\top \delta^*\right)^2 = \Omega\left(\frac{d^2}{n}\right),\tag{103}$$

for at least $\Omega(n)$ indices $i \in [n]$. This also implies (up to considering $-\delta^*$ instead of δ^*) that

$$x_i^\top \delta^* = \Omega\left(\frac{d}{\sqrt{n}}\right) = \Omega\left(\sqrt{d}\log^2 d\right),\tag{104}$$

where the last step is justified by $d/\log^4 d = \Omega(n)$.

Let's now define

$$\delta_1^* = \frac{\delta^*}{2} + \frac{v}{2},\tag{105}$$

and

$$\delta_2^* = \frac{\delta^*}{2} - \frac{v}{2},\tag{106}$$

where $v \in \mathbb{R}^d$ is a generic fixed vector such that $\|v\|_2 = \sqrt{d}$, and $\|Xv\|_2 = 0$, i.e. $v \in \mathrm{Span}\{\mathrm{rows}(X)\}^\perp$. This is again possible because of $d/\log^4 d = \Omega(n)$. As $\|\delta^*\|_2 \leq \sqrt{d}$ by Lemma D.1, the first part of the thesis follows from a straightforward application of the triangle inequality.

Since $v^\top x_i = 0$ for every $1 \leq i \leq n$, we also have

$$x_i^\top \delta_1^* = x_i^\top \delta_2^* = x_i^\top \delta^*/2,\tag{107}$$

which readily gives the desired result. $\square$

Lemma D.3. *Let's define $\sigma(\Delta) \in \mathbb{R}^n$ as*

$$\sigma(\Delta) := \frac{XW\Delta}{\sqrt{d}},\tag{108}$$

where $\Delta \in \mathbb{R}^d$. Then, with probability at least $1 - \exp(-c\log^2 d)$ over W, there exist Δ_1^* and Δ_2^* such that,

$$\|\Delta_1^*\|_2 \leq \sqrt{d}, \qquad \|\Delta_2^*\|_2 \leq \sqrt{d}, \qquad \|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d}),\tag{109}$$

and

$$\sigma(\Delta_1^*)_i = \Omega\left(\log^2 d\right).\tag{110}$$

$$\sigma(\Delta_2^*)_i = \Omega\left(\log^2 d\right).\tag{111}$$

for at least $\Omega(n)$ indices $i \in [n]$.

Proof. Let's set

$$\Delta = CW^\top \delta,\tag{112}$$

where $\delta \in \mathbb{R}^d$ is independent from W, and C is an absolute constant that will be fixed later in the proof. For every i, we can write

$$\delta = \frac{\delta^\top x_i}{\|x_i\|_2^2} x_i + \delta_i^\perp,\tag{113}$$

with $\delta_i^\perp$ being orthogonal with x_i by construction. Let's for now suppose $\left\|\delta_i^\perp\right\|_2 \neq 0$. Let $a(\delta)_i := \delta^\top x_i$. We have

$$\begin{aligned}
\sigma(\Delta)_i &\geq C\frac{a(\delta)_i}{\sqrt{d}\,\|x_i\|_2^2}\left\|W^\top x_i\right\|_2^2 - C\left|\frac{x_i^\top WW^\top \delta_i^\perp}{\sqrt{d}}\right| \\
&= C\frac{a(\delta)_i}{\sqrt{d}\,\|x_i\|_2^2}\left\|W^\top x_i\right\|_2^2 - C\frac{\left\|\delta_i^\perp\right\|_2}{d}\left|x_i^\top WW^\top \frac{\sqrt{d}\,\delta_i^\perp}{\left\|\delta_i^\perp\right\|_2}\right|.
\end{aligned}\tag{114}$$

In the probability space of W, as the entries of W are i.i.d. and Gaussian $\mathcal{N}(0, 1/d)$, we have that $W^\top x_i$ and $W^\top \frac{\sqrt{d}\,\delta_i^\perp}{\|\delta_i^\perp\|_2}$ are two independent standard Gaussian vectors, namely, ρ_1 and ρ_2. This implies

$$\left\|W^\top x_i\right\|_2^2 = \Omega(d), \qquad \left|x_i^\top W W^\top \frac{\sqrt{d}\,\delta_i^\perp}{\|\delta_i^\perp\|_2}\right| =: |\rho_1^\top \rho_2| = \mathcal{O}\left(\sqrt{d}\log d\right), \tag{115}$$

with probability at least $1 - \exp(-c_1 \log^2 d)$, over the probability space of W. The first statement holds because of Theorem 3.1.1. of (Vershynin, 2018), and the second one because, in the probability space of ρ_1, $\rho_1^\top \rho_2$ is a Gaussian random variable with variance $\|\rho_2\|_2^2$, which in turn is $\mathcal{O}(d)$ with probability at least $1 - \exp(-c_2 d)$ over ρ_2. Using $\|x_i\|_2^2 = d$, we get, for two positive absolute constants C_1 and C_2,

$$\sigma(\Delta)_i \geq C_1 \frac{a(\delta)_i}{\sqrt{d}} - C_2 \frac{\|\delta_i^\perp\|_2 \log d}{\sqrt{d}} \geq C_1 \frac{a(\delta)_i}{\sqrt{d}} - C_2 \frac{\|\delta\|_2 \log d}{\sqrt{d}}. \tag{116}$$

Notice that the previous equation would still hold even in the case $\|\delta_i^\perp\|_2 = 0$, which is therefore a case now included in our derivation.

By Lemma D.2, we can choose two vectors δ_1^* and δ_2^*, independently from W such that, for $k \in \{1, 2\}$, $\|\delta_k^*\|_2 \leq \sqrt{d}$ and $a(\delta_k^*)_i = \Omega(\sqrt{d}\log^2 d)$ for $\Omega(n)$ indices $i \in [n]$. This gives, for another absolute positive constant C_3,

$$\sigma(\Delta_1^*)_i \geq C_3 \log^2 d - C_2 \log d = \Omega(\log^2 d), \tag{117}$$

$$\sigma(\Delta_2^*)_i \geq C_3 \log^2 d - C_2 \log d = \Omega(\log^2 d), \tag{118}$$

for $\Omega(n)$ indices $i \in [n]$.

Let's now verify that $\|\Delta_1^* - \Delta_2^*\|_2 = C \|W^\top(\delta_1^* - \delta_2^*)\|_2 = \Omega(\sqrt{d})$. As $\delta_1^* - \delta_2^*$ is independent on W, and $\|\delta_1^* - \delta_2^*\|_2 = \sqrt{d}$, we have that $W^\top(\delta_1^* - \delta_2^*)$ is a standard Gaussian random vector, in the probability space of W. Thus, by Theorem 3.3.1 of (Vershynin, 2018), we have

$$\|\Delta_1^* - \Delta_2^*\|_2 = C \|W^\top(\delta_1^* - \delta_2^*)\|_2 = \Omega(\sqrt{d}), \tag{119}$$

with probability at least $1 - \exp(-c_3 d)$, on the probability space of W.

We are left to verify that, for $k \in \{1, 2\}$, $\|\Delta_k^*\|_2 = C \|W^\top \delta_k^*\|_2 \leq \sqrt{d}$. This is readily implied by Theorem 4.4.5 of (Vershynin, 2018), which gives $\|W\|_{\mathrm{op}} = \mathcal{O}(1)$ with probability at least $1 - \exp(-c_4 d)$. Taking the intersection between this high probability event and the ones in (115) and (119), and setting C small enough to have $C \|W\|_{\mathrm{op}} \leq 1$, we get the thesis. $\qquad\square$

Lemma D.4. *For every $i \in [n]$, with probability at least $1 - \exp(-c \log^2 d)$ over W, there exist Δ_1^* and Δ_2^* such that,*

$$\|\Delta_1^*\|_2 \leq \sqrt{d}, \qquad \|\Delta_2^*\|_2 \leq \sqrt{d}, \qquad \|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d}), \tag{120}$$

and

$$\left\|[s(X^i(\Delta_1^*))]_{j:}^\top - e_i\right\|_2 = o\left(\frac{1}{\sqrt{d}}\right), \tag{121}$$

$$\left\|[s(X^i(\Delta_2^*))]_{j:}^\top - e_i\right\|_2 = o\left(\frac{1}{\sqrt{d}}\right), \tag{122}$$

for at least $\Omega(n)$ indices $j \in [n]$, where e_i represents the i-th element of the canonical basis in $\mathbb{R}^n$.

Proof. As we can write $X^i(\Delta) = X + e_i \Delta^\top$, we can rewrite (84) as

$$\sqrt{d}\, S(X^i(\Delta)) = X^i(\Delta) W X^i(\Delta)^\top = XWX^\top + XW\Delta e_i^\top + e_i \Delta^\top W X + e_i \Delta^\top W \Delta e_i^\top. \tag{123}$$

Let's look at the j-th row of $S(X^i(\Delta))$, with $j \neq i$. We have

$$[S(X^i(\Delta))]_{j:} = \frac{x_j^\top W X^\top}{\sqrt{d}} + \frac{x_j^\top W \Delta e_i^\top}{\sqrt{d}}. \tag{124}$$

With probability at least $1 - \exp(-c_1 \log^2 d)$ over W, we have that all the n entries of the vector $x_j^\top W X^\top / \sqrt{d}$ are smaller than $\log d$, as they are all standard Gaussian random variables. By Lemma D.3, we have that there exist Δ_1^* and Δ_2^* such that $\|\Delta_1^*\|_2 \le \sqrt{d}$, $\|\Delta_2^*\|_2 \le \sqrt{d}$, $\|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d})$ and

$$\frac{x_j^\top W \Delta_k^*}{\sqrt{d}} = \Omega(\log^2 d), \qquad k \in \{1, 2\}, \tag{125}$$

for $\Omega(n)$ indices $j \in [n]$, with probability at least $1 - \exp(-c_2 \log^2 d)$ over W. Let's consider this set of indices until the end of the proof, where we have that, for $k \in \{1, 2\}$, $[S(X^i(\Delta_k^*))]_{ji} = \Omega(\log^2 d)$.

After applying the $\mathrm{softmax}$ function to the row in (124) as indicated in (85), we get, for every index $k \in [n]$ and $k \ne i$

$$
\begin{aligned}
[s(X^i(\Delta_1^*))]_{jk} &= \mathrm{softmax}\left([S(X^i(\Delta_1^*))]_{j:}\right)_k \\
&= \frac{\exp\left([S(X^i(\Delta_1^*))]_{jk}\right)}{\sum_l \exp\left([S(X^i(\Delta_1^*))]_{jl}\right)} \\
&\le \frac{d}{\exp\left([S(X^i(\Delta_1^*))]_{ji}\right)} \\
&\le \frac{d}{C d^{\log d}} = o\left(\frac{1}{d}\right),
\end{aligned}
\tag{126}
$$

and

$$
\begin{aligned}
[s(X^i(\Delta_1^*))]_{ji} &= \mathrm{softmax}\left([S(X^i(\Delta_1^*))]_{j:}\right)_i \\
&= 1 - \frac{\sum_{l \ne i} \exp\left([S(X^i(\Delta_1^*))]_{jl}\right)}{\sum_l \exp\left([S(X^i(\Delta_1^*))]_{jl}\right)} \\
&\ge 1 - \frac{nd}{\exp\left([S(X^i(\Delta_1^*))]_{ji}\right)} \\
&\ge 1 - \frac{nd}{C d^{\log d}} = 1 - o\left(\frac{1}{d}\right),
\end{aligned}
\tag{127}
$$

where the last step is justified by the fact that $d / \log^4 d = \Omega(n)$.

Thus, for the same reason, we have

$$\left\|[s(X^i(\Delta_1^*))]_{j:}^\top - e_i\right\|_2^2 = \sum_{k \ne i}[s(X^i(\Delta_1^*))]_{jk}^2 + \left(1 - [s(X^i(\Delta_1^*))]_{ji}\right)^2 = no\left(\frac{1}{d^2}\right) = o\left(\frac{1}{d}\right). \tag{128}$$

As (128) can be written also with respect to Δ_2^*, the thesis readily follows. $\qquad\square$

Lemma D.5. *For every $i \in [n]$, with probability at least $1 - \exp(-c \log^2 d)$ over W, there exist Δ^* such that $\|\Delta^*\|_2 \le \sqrt{d}$, and*

$$\left\|\varphi_{\mathrm{RAF}}(X) - \varphi_{\mathrm{RAF}}(X^i(\Delta^*))\right\|_F = \Omega(\sqrt{dn}). \tag{129}$$

Proof. By (86), we have that,
$$\varphi_{\mathrm{RAF}}(X^i(\Delta)) = s(X^i(\Delta)) X^i(\Delta), \tag{130}$$

By Lemma D.4, we have that there exists Δ_1^*, with $\|\Delta_1^*\|_2 \le \sqrt{d}$, such that there are $\Omega(n)$ indices j such that

$$\left\|[s(X^i(\Delta_1^*))]_{j:}^\top - e_i\right\|_2 = o\left(\frac{1}{\sqrt{d}}\right), \tag{131}$$

with probability at least $1 - \exp(-c_1 \log^2 d)$ over W. Until the end of the proof, we will refer to j as a generic index for which the previous equation holds, and we will condition on the high probability event that the equations in the statement of Lemma D.4 hold.

As we can write

$$\left[\varphi_{\mathrm{RAF}}(X^i(\Delta_1^*))\right]_{j:} = [s(X^i(\Delta_1^*))]_{j:}X^i(\Delta_1^*) = e_i^\top X^i(\Delta_1^*) + \left([s(X^i(\Delta_1^*))]_{j:} - e_i^\top\right)X^i(\Delta_1^*), \tag{132}$$

and $e_i^\top X^i(\Delta_1^*) = x_i^\top + \Delta_1^{*\top}$, we have

$$
\begin{aligned}
\left\|\left[\varphi_{\mathrm{RAF}}(X^i(\Delta_1^*))\right]_{j:} - (x_i + \Delta_1^*)^\top\right\|_2 &\le \left\|[s(X^i(\Delta_1^*))]_{j:} - e_i^\top\right\|_2\left\|X^i(\Delta_1^*)\right\|_{\mathrm{op}} \\
&\le o\left(\frac{1}{\sqrt{d}}\right)(\|X\|_F + \|\Delta_1^*\|_2) \\
&= o\left(\frac{1}{\sqrt{d}}\right)\mathcal{O}\left(\sqrt{dn}\right) = o\left(\sqrt{d}\right),
\end{aligned}
\tag{133}
$$

where the last step follows from our assumption $d/\log^4 d = \Omega(n)$. By Lemma D.4, there also exists Δ_2^* such that $\|\Delta_2^*\|_2 \le \sqrt{d}$, $\|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d})$ and (133) holds.

Let's now suppose by contradiction that, for some j,

$$\left\|[\varphi_{\mathrm{RAF}}(X)]_{j:} - [\varphi_{\mathrm{RAF}}(X^i(\Delta_1^*))]_{j:}\right\|_2 = o\left(\sqrt{d}\right), \qquad \left\|[\varphi_{\mathrm{RAF}}(X)]_{j:} - [\varphi_{\mathrm{RAF}}(X^i(\Delta_2^*))]_{j:}\right\|_2 = o\left(\sqrt{d}\right). \tag{134}$$

Then, we have

$$
\begin{aligned}
\|\Delta_1^* - \Delta_2^*\|_2 &= \|(x_i + \Delta_1^*) - (x_i + \Delta_2^*)\|_2 \\
&= \left\| \left((x_i + \Delta_1^*) - [\varphi_{\mathrm{RAF}}(X^i(\Delta_1^*))]_{j:}\right) + \left([\varphi_{\mathrm{RAF}}(X^i(\Delta_1^*))]_{j:} - [\varphi_{\mathrm{RAF}}(X)]_{j:}\right) \right. \\
&\quad\left. - \left((x_i + \Delta_2^*) - [\varphi_{\mathrm{RAF}}(X^i(\Delta_2^*))]_{j:}\right) - \left([\varphi_{\mathrm{RAF}}(X^i(\Delta_2^*))]_{j:} - [\varphi_{\mathrm{RAF}}(X)]_{j:}\right) \right\| \\
&\le \left\|(x_i + \Delta_1^*) - [\varphi_{\mathrm{RAF}}(X^i(\Delta_1^*))]_{j:}\right\|_2 + \left\|[\varphi_{\mathrm{RAF}}(X^i(\Delta_1^*))]_{j:} - [\varphi_{\mathrm{RAF}}(X)]_{j:}\right\|_2 \\
&\quad + \left\|(x_i + \Delta_2^*) - [\varphi_{\mathrm{RAF}}(X^i(\Delta_2^*))]_{j:}\right\|_2 + \left\|[\varphi_{\mathrm{RAF}}(X^i(\Delta_2^*))]_{j:} - [\varphi_{\mathrm{RAF}}(X)]_{j:}\right\|_2 \\
&= o(\sqrt{d}),
\end{aligned}
\tag{135}
$$

where the third step holds by triangle inequality, and the last comes from (133) and (134). As $\|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d})$ by Lemma D.4, we get the desired contradiction, and we have that at least one equation in (134) doesn't hold.

Let's therefore denote by Δ^* the vector with $\|\Delta^*\|_2 \le \sqrt{d}$ such that

$$\left\|[\varphi_{\mathrm{RAF}}(X)]_{j:} - [\varphi_{\mathrm{RAF}}(X^i(\Delta^*))]_{j:}\right\|_2 = \Omega\left(\sqrt{d}\right) \tag{136}$$

holds for $\Omega(n)$ indices j. Due to the previous conditioning, such a vector exists with probability at least $1 - \exp(-c_1\log^2 d)$ over W. Let's denote the set of these indices by J, with $|J| = \Omega(n)$ (where we denote set cardinality by $|\cdot|$).

Thus, we have

$$\left\|\varphi_{\mathrm{RAF}}(X) - \varphi_{\mathrm{RAF}}(X^i(\Delta^*))\right\|_F^2 \ge \sum_{j \in J}\left\|[\varphi_{\mathrm{RAF}}(X)]_{j:} - [\varphi_{\mathrm{RAF}}(X^i(\Delta^*))]_{j:}\right\|_2^2 = \Omega(nd), \tag{137}$$

which concludes the proof.

$\square$

Theorem 5.1 Let $\varphi_{\mathrm{RAF}}(X)$ be the random attention features map defined in (3). Let $X \in \mathbb{R}^{n \times d}$ be a generic input sample s.t. Assumption 3.1 holds, and assume $d/\log^4 d = \Omega(n)$. Let $\mathcal{S}_{\mathrm{RAF}}(X)$ be the the word sensitivity defined in (4). Then, we have

$$\mathcal{S}_{\mathrm{RAF}}(X) = \Omega(1), \tag{138}$$

with probability at least $1 - \exp(-c\log^2 d)$ over W.

Proof. We have $\varphi_{\text{RAF}}(X) = s(X)X$, which implies

$$[\varphi_{\text{RAF}}(X)]_{j:} = [s(X)]_{j:} X, \tag{139}$$

and

$$\left\| [\varphi_{\text{RAF}}(X)]_{j:} \right\|_2 = \left\| \sum_{k=1}^n [s(X)]_{jk} x_k \right\|_2 \le \sum_{k=1}^n [s(X)]_{jk} \|x_k\|_2 = \sqrt{d}, \tag{140}$$

where the second step follows from triangle inequality, and the last step holds as $\sum_{k=1}^n [s(X)]_{jk} = 1$, and $\|x_k\|_2 = \sqrt{d}$ for every k. This readily implies

$$\|\varphi_{\text{RAF}}(X)\|_F \le \sqrt{nd}. \tag{141}$$

By Lemma D.5, we have that, with probability at least $1 - \exp(-c \log^2 d)$ over W, there exist Δ^* such that $\|\Delta^*\|_2 \le \sqrt{d}$, and

$$\left\| \varphi_{\text{RAF}}(X) - \varphi_{\text{RAF}}(X^i(\Delta^*)) \right\|_F = \Omega(\sqrt{dn}). \tag{142}$$

This, together with (141), concludes the proof. □

E. Assumption 6.1 and adversarial robustness

Assumption 6.1 requires the perturbation Δ to be such that the model $f_{\text{RF}}(\cdot, \theta^*)$ gives a similar output when evaluated on the two new samples X and $X^i(\Delta)$. This assumption is necessary to understand the different behaviour of the RF and the RAF model in our setting, as there in fact exists an adversarial patch Δ such that, for example, $f(X^i(\Delta), \theta_r^*)$ and $f(X, \theta_r^*)$ are very different from each other (e.g., $f_{\text{RF}}(X^i(\Delta), \theta_r^*) = y_\Delta$ while $f_{\text{RF}}(X, \theta_r^*) = y$).

This conclusion derives from the adversarial vulnerability of the RF model, extensively studied in previous work (Dohmatob & Bietti, 2022; Dohmatob, 2022; Bombari et al., 2023). We remark that this vulnerability depends on the scalings of the problem, i.e., n, d and N. In fact, (using the re-trained solution as example) we can write (assuming $\theta_0 = 0$ for simplicity)

$$\begin{aligned}
\left| f_{\text{RF}}(X^i(\Delta), \theta_r^*) - f_{\text{RF}}(X, \theta_r^*) \right| &= \left| \left(\varphi_{\text{RF}}(X^i(\Delta)) - \varphi_{\text{RF}}(X) \right)^\top \Phi_{\text{RF},r}^+ \mathcal{Y}_r \right| \\
&\le \left\| \varphi_{\text{RF}}(X^i(\Delta)) - \varphi_{\text{RF}}(X) \right\|_2 \left\| \Phi_{\text{RF},r}^+ \right\|_{\text{op}} \|Y_r\|_2.
\end{aligned} \tag{143}$$

If we bound the three terms on the RHS separately, we get:

- $\left\| \varphi_{\text{RF}}(X^i(\Delta)) - \varphi_{\text{RF}}(X) \right\|_2 = \mathcal{O}\left(\sqrt{k/n} \right)$ with probability at least $1 - \exp(-cD)$ over V, by Lemma B.1;

- $\left\| \Phi_{\text{RF},r}^+ \right\|_{\text{op}} = \lambda_{\min}^{-1/2}(K_{\text{RF},r}) = \mathcal{O}\left(\sqrt{1/k} \right)$, with probability at least $1 - \exp\left(-c \log^2 N \right)$ over V and $\mathcal{X}_r$, by Lemma B.2;

- $\|Y_r\|_2 = \sqrt{N+1}$, as we are considering labels in $\{-1, 1\}$.

Thus, we conclude that

$$\left| f_{\text{RF}}(X^i(\Delta), \theta_r^*) - f_{\text{RF}}(X, \theta_r^*) \right| = \mathcal{O}\left(\sqrt{\frac{N}{n}} \right), \tag{144}$$

with high probability. As a consequence, in the regime where $N = o(n)$, $f_{\text{RF}}(\cdot, \theta_r^*)$ cannot distinguish between the samples X and $X^i(\Delta)$, without the additional need for Assumption 6.1. This result is also shown in our experiments, in the left subplots of Figure 4, where the points approach smaller values of γ and consequently higher values of the error as N decreases, becoming comparable with n.

F. Further experiments

In Figure 4, we compute Δ^* by optimizing with respect to it the following two losses (for fine-tuning and re-training, respectively):

$$\ell_{\theta_f^*}(\Delta) := \left(\frac{\varphi_{\text{RAF}}(X^1(\Delta))^\top \varphi_{\text{RAF}}(X)}{\|\varphi_{\text{RAF}}(X)\|_2^2} + 1 \right)^2, \tag{145}$$
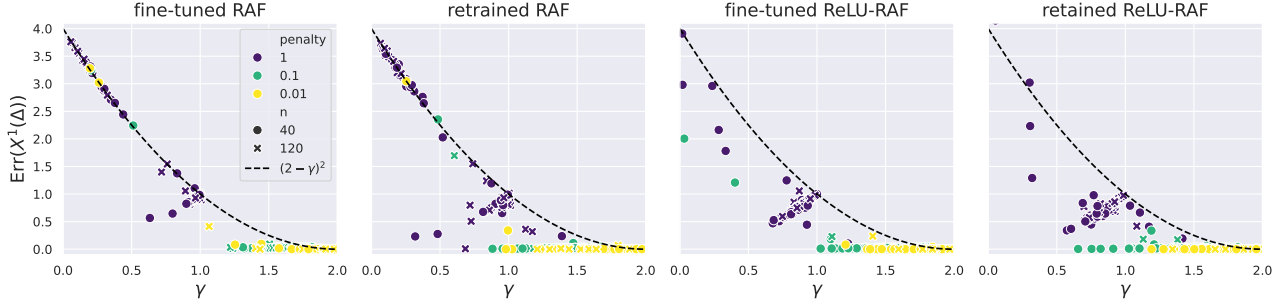
Figure 5. $\mathrm{Err}_\varphi(X^i(\Delta), \theta^*_{f/r})$ for the RAF (two left sub-plots) and ReLU-RAF (two right subplots) maps, as a function of the smallest γ for which Assumption 6.1 is satisfied. Every sub-plot has a fixed context length $n = \{40, 120\}$, embedding dimension $d = 768$ and number of training samples $N = 400$. Every point in the scatter-plots represents an independent simulation where (X, y) and $(\mathcal{X}, \mathcal{Y})$ are the BERT-Base embeddings of a random subset of the imdb dataset (after pre-processing to fulfill Assumption 3.1). For every point, Δ is obtained through constrained gradient descent optimization of $\ell_{\mathrm{Err},p}(\Delta)$, defined in (147), for different values of the penalty term $p = \{1, 0.1, 0.01\}$.

$$\ell_{\theta^*_r}(\Delta) := \left(\mathcal{F}_{\mathrm{RAF}}(X, X^i(\Delta)) + 1\right)^2, \tag{146}$$

subject to the constraint $\|\Delta^*\| \leq \sqrt{d}$. We also report with cross markers the points obtained optimizing the loss $\ell_{\mathrm{Err}}(\Delta) := \mathrm{Err}_{\mathrm{RAF}}(X^i(\Delta), \theta^*_{f/r})$, showing that this method provides less interesting results, as the error tends to be minimized at the expenses of a large value of γ.

An alternative approach is to still optimize with respect to the errors directly, but after introducing a penalty term p on the value of γ. In particular, we consider the following penalized loss

$$\ell_{\mathrm{Err},p}(\Delta) := \mathrm{Err}_{\mathrm{RAF}}(X^i(\Delta), \theta^*_{f/r}) + p\left(f_{\mathrm{RAF}}(X^i(\Delta), \theta^*) - f_{\mathrm{RAF}}(X, \theta^*)\right)^2. \tag{147}$$

We perform new experiments with this optimization algorithm, for different values of p, and we report the results in Figure 5. In this case, compared to the loss $\ell_{\mathrm{Err}}(\Delta)$, we can more easily obtain points that lie below the lower bound. However, it remains difficult to find points where both the error and γ are small.

Another optimization option is to introduce a penalty term to the losses in (145) and (146):

$$\ell_{\theta^*_f,p}(\Delta) := \left(\frac{\varphi_{\mathrm{RAF}}(X^1(\Delta))^\top \varphi_{\mathrm{RAF}}(X)}{\|\varphi_{\mathrm{RAF}}(X)\|_2^2} + 1\right)^2 + p\left(f_{\mathrm{RAF}}(X^i(\Delta), \theta^*) - f_{\mathrm{RAF}}(X, \theta^*)\right)^2 \tag{148}$$

and

$$\ell_{\theta^*_r,p}(\Delta) := \left(\mathcal{F}_{\mathrm{RAF}}(X, X^i(\Delta)) + 1\right)^2 + p\left(f_{\mathrm{RAF}}(X^i(\Delta), \theta^*) - f_{\mathrm{RAF}}(X, \theta^*)\right)^2, \tag{149}$$

for the fine-tuning and re-training case respectively. We perform new experiments with this optimization algorithm, for different values of p, and we report the results in Figure 6. In this case, it is easier to obtain final points that respect Assumption 6.1 with a lower value of γ.

Finally, we consider swapping the losses $\ell_{\theta^*_f}$ and $\ell_{\theta^*_r}$ defined above, i.e., employ the former for re-training and the latter for fine-tuning. Given the heuristic nature of these losses, it is a priori not obvious that they perform their best on the respective setting, as they could be interchangeable. In Figure 7, we report the results of this investigation. We report in deep-blue the points resulting from the optimization of $\ell_{\theta^*_f}$, and in yellow the points resulting from the optimization of $\ell_{\theta^*_r}$, for both the fine-tuned and re-trained setting. We note that $\ell_{\theta^*_f}$ performs better on the fine-tuned solution, and $\ell_{\theta^*_r}$ better on the re-trained one.
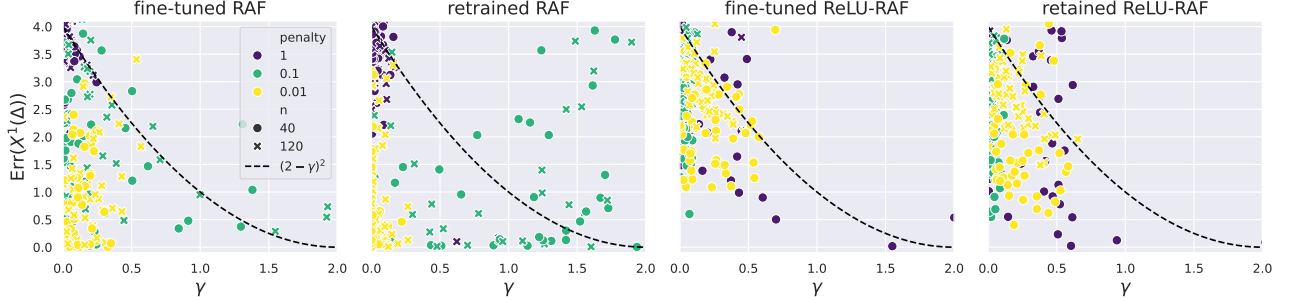
Figure 6. $\mathrm{Err}_\varphi(X^i(\Delta), \theta^*_{f/r})$ for the RAF (two left sub-plots) and ReLU-RAF (two right subplots) maps, as a function of the smallest γ for which Assumption 6.1 is satisfied. For every point, Δ is obtained through constrained gradient descent optimization of $\ell_{\theta^*_r, f}(\Delta)$ in the fine-tuned case, and $\ell_{\theta^*_r, p}(\Delta)$ in the re-trained case, defined in (148) and (149) respectively, for different values of the penalty term $p = \{1, 0.1, 0.01\}$. The rest of the setup is equivalent to the one described in Figure 5.
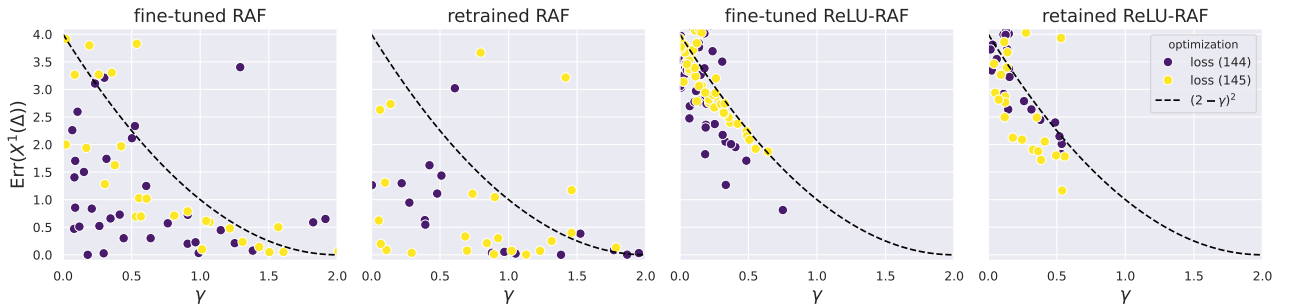

Figure 7. $\mathrm{Err}_\varphi(X^i(\Delta), \theta^*_{f/r})$ for the RAF (two left sub-plots) and ReLU-RAF (two right subplots) maps, as a function of the smallest γ for which Assumption 6.1 is satisfied. We consider fixed embedding dimension $d = 768$, context length $n = 120$, and number of training samples $N = 400$. Every point in the scatter-plots represents an independent simulation where (X, y) and $(\mathcal{X}, \mathcal{Y})$ are the BERT-Base embeddings of a random subset of the imdb dataset (after pre-processing to fulfill Assumption 3.1). For every point, Δ is obtained through constrained gradient descent optimization of either $\ell_{\theta^*_f}$, defined in (145), or $\ell_{\theta^*_r}$, defined in (146).