On the Robustness of Text Vectorizers

Rémi Catellier [1 2] **Samuel Vaiter** [1 3] **Damien Garreau** [1 2]

Abstract

A fundamental issue in machine learning is the robustness of the model with respect to changes in the input. In natural language processing, models typically contain a first embedding layer, transforming a sequence of tokens into vector representations. While the robustness with respect to changes of continuous inputs is well-understood, the situation is less clear when considering discrete changes, for instance replacing a word by another in an input sentence. Our work formally proves that popular embedding schemes, such as concatenation, TF-IDF, and Paragraph Vector (*a.k.a.* `doc2vec`), exhibit robustness in the Hölder or Lipschitz sense with respect to the Hamming distance. We provide quantitative bounds for these schemes and demonstrate how the constants involved are affected by the length of the document. These findings are exemplified through a series of numerical examples.

1. Introduction

Recent advances in natural language processing (NLP) have exceeded all expectations. In particular, the advent of large language models such as BERT (Devlin et al., 2018) and GPT (Brown et al., 2020) are transforming radically the way we interact with computers. They typically rely on a deep neural network (DNN) architecture and are trained on a variety of tasks such as sentiment analysis, translation, and text summarization.

A known issue with DNNs is the existence of *adversarial examples*: examples modified in order to radically change the output of the model. Initially popularized in the context of image classification (Szegedy et al., 2014), such examples also exist in NLP and a flourishing literature exists on this topic (Zhang et al., 2020). This problem has sparked a tremendous interest into the *robustness* of models with respect to small changes in the input. In this paper, we focus on the robustness of the *vectorization* NLP pipelines: the transformation of the input document into a vector representation. We will consider documents as ordered sequences of tokens, not necessarily corresponding to words. For instance, GPT 2 uses Byte Pair encoding (Gage, 1994; Sennrich et al., 2016), which relies on tokens corresponding to sub-words.

As far as we reckon, there are essentially three main schools of thought when it comes to vectorization:
(i) **concatenation** of vectors corresponding to each token of the document. These vectors are often called *word vectors* when the tokens are individual words. They can either be one-hot representations of the tokens, or obtained by a mapping learned from data. A celebrated approach to produce word vectors is `word2vec` (Mikolov et al., 2013a;b), which transports semantic properties to the embedding space. Many other methods exist, such as GloVe (Pennington et al., 2014), EMF (Li et al., 2015), WordPiece (Wu et al., 2016), FastText (Bojanowski et al., 2017), and ELMo (Peters et al., 2018). Positional information is typically added to the token embeddings.
(ii) **TF-IDF (term frequency - inverse document frequency)**, taking words as tokens and simply considering the frequencies of each individual word in the document. These frequencies are reweighted by an overall importance term to take into account the lesser importance of frequently appearing words such as articles. This is the historical approach to text vectorization (Luhn, 1957; Jones, 1972).
(iii) ***ad hoc* approaches**. Notably, Paragraph Vector (also known as `doc2vec` (Le & Mikolov, 2014)) extends the ideas of `word2vec`. Although we will focus on `doc2vec` in this work, we emphasize that there exists other *ad hoc* approaches, such as skip-thought vectors (Kiros et al., 2015), quick-thought (Logeswaran & Lee, 2018), or universal sentence encoder (Cer et al., 2018).

A priori, vectorizers are not designed to be robust to small changes. Even when modifying a single word of the input document, the embedding could change drastically. Thus, we ask the following question:

> *Are text vectorizers* ***provably*** *robust with respect to modifying a small subset of the document?*

[1] Université Côte d'Azur, CNRS, LJAD, France [2] Inria, France [3] CNRS, France. Correspondence to: Damien Garreau <damien.garreau@unice.fr>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Typical notions of robustness in machine learning deals with *continuous* input data: changing slightly the observation means that for instance its ℓ^2-norm evolves infinitesimally. The challenge of our analysis is the fundamentally *discrete* nature of text data. Changing a word in a document is usually not innocuous – one can think of extreme cases where the meaning of this word is flipped – and vectorizers sensitive to the semantics of input documents should capture this phenomenon. Nevertheless, we show that the answer is positive for all vectorizers that we study. Another difficulty is that the mathematical formalization of some of these vectorizers was not the main concern of the community. A necessary first step is thus to give an unequivocal definition of our objects of interest.

Contributions. In this paper, we analyze the robustness of vectorizers as their local regularity (Lipschitz, Hölder) with respect to the **Hamming distance** (Section 2). We prove:
- the $1/2$-Hölder continuity of **concatenation of token and positional embeddings** (Proposition 3.1);
- the Lipschitz continuity of **TF-IDF** (Proposition 4.1), and the $1/2$-Hölder continuity of it normalized variant (Proposition 4.2);
- the Lipschitz continuity of **`doc2vec`** (Theorem 5.1). As a necessary step to derive the latter, we make two new mathematical contributions (see Appendix), we propose:
- a **local Lipschitz analysis of the softmax** (Theorem H.6);
- a **Grönwall–Bellman–Bahouri result** (Theorem G.1) needed when casting the `doc2vec` analysis as an ODE problem. The code for all experiments of the paper is available at `https://github.com/dgarreau/vectorizer-robustness`.

Related work. *(Adversarial examples).* A major motivation for studying robustness is its impact on the existence of adversarial examples. In the case of DNNs, robustness often means Lipschitz continuity with respect to the inputs. For instance, one can show that a network having a small Lipschitz constant prevents the existence of small adversarial changes. More precisely, Hein & Andriushchenko (2017) provide a lower bound on the norm of the input manipulation needed to change the classifier decision inversely proportional to the Lipschitz constant of the network. This was later extended by Weng et al. (2018b) to DNNs with ReLU activations. Quantitatively, Weng et al. (2018a) show that fully connected layers have a Lipschitz constant potentially as large as the operator norm of the weight matrix. From a practical point of view, it has also been noticed that enforcing the Lipschitz constants of the layers to remain low does improve the robustness (Cisse et al., 2017).
(Generalization & interpolation). It is known that robust algorithms generalize better. In particular, Xu & Mannor (2012) derive generalization bounds for generic algorithms depending in their robustness. The definition of robustness here includes Lipschitz continuous DNNs. More recently, Bubeck & Sellke (2021) extending (Bubeck et al., 2020) showed that in order to train Lipschitz continuous models, one has to take a large number of parameters.
(Theory of vectorizers). Surprisingly, the robustness of vectorizers received little attention until now on the theoretical side, and all previous works on robustness assume *continuous* input. Nevertheless, there exist some theoretical works on similar problems. Most notably, Arora et al. (2016) analyze a large class of word vectorizers and explain how the intriguing alignment properties observed experimentally appear.

Notations. For $u \in \mathbb{R}^p$, we denote by $\|u\|$ its Euclidean norm. Let $g : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ be a function. The derivative in the time variable (μ) is denoted by $\partial_\mu g$ whereas ∇g (resp. $\nabla^2 g$) denotes the Jacobian (resp. the Hessian) of g in the space variable. We let $\mathbb{1} = (1, \ldots, 1)^\top \in \mathbb{R}^d$. For a matrix R, $\sigma_{\min}(R)$ is its smallest singular value. For a given set $\mathcal{S}$, $|\mathcal{S}|$ is its cardinal.

2. Framework

Let us now present the mathematical framework in which we perform our analysis. We consider tokens from a finite dictionary $\mathcal{D}$, identified as $[D] := \{1, \ldots, D\}$. A *document* x built on $\mathcal{D}$ is a finite sequence of elements of $\mathcal{D}$, and we write $[D]^*$ for the set of all documents. Thus the central object of our work, a vectorizer, is simply a mapping $\varphi : [D]^* \to \mathbb{R}^d$, where d is the dimension of the embedding. The *length* of x will be denoted by $T(x)$, and therefore x can be written as $(x_1, \ldots, x_{T(x)})$. The set of all documents over $\mathcal{D}$ of length T will be denoted $[D]^T \subset [D]^*$. When there is no ambiguity, we remove the dependency in x from our notation, *e.g.*, $T(x)$ becomes T.

As discussed in the related work, robustness is often synonym with *Lipschitz continuity* of the model – distance between outputs lies within a constant factor of the distance between inputs. As distance between input documents x and $\tilde{x}$ of same length, we consider the *Hamming distance*, which is the number of indices such that x_t and $\tilde{x}_t$ differ:

$$\mathrm{d_H}(x, \tilde{x}) := |\{t \in [T] : x_t \neq \tilde{x}_t\}| \, .$$

The distance between outputs will simply be measured by the Euclidean norm in $\mathbb{R}^d$. In definitive, for a given document length T, what we call **Lipschitz continuity** of the vectorizer φ can be written as

$$\forall x, \tilde{x} \in [D]^T, \quad \|\varphi(x) - \varphi(\tilde{x})\| \leq C \, \mathrm{d_H}(x, \tilde{x}) \, , \qquad (1)$$

where C is called the Lipschitz constant. Another way to quantify robustness is to allow for an exponent in Eq. (1):

$$\forall x, \tilde{x} \in [D]^T, \quad \|\varphi(x) - \varphi(\tilde{x})\| \leq C \, \mathrm{d_H}(x, \tilde{x})^\beta \, , \qquad (2)$$

with $1 \geq \beta > 0$. This is known as **Hölder continuity**, and coincides with Lipschitz continuity whenever $\beta = 1$. While it is known that Lipschitz continuity implies Hölder continuity on the real line when $\beta \leq 1$, this is not the case here, since $\mathrm{d_H}$ takes values in $\mathbb{N}$. Thus in our setting, **Lipschitz continuity is a weaker notion of robustness than Hölder continuity.**

Often we obtain more precise results, depending explicitly on the set of indices such that the documents differ. To this extent, for a given subset $\mathcal{S}$ of $[T]$, we define the *set of $\mathcal{S}$-close documents* $B_{\mathcal{S}}(x)$ of $x \in [D]^T$ as

$$B_{\mathcal{S}}(x) = \{\tilde{x} \in [D]^T \,:\, x_i = \tilde{x}_i \text{ for } i \notin \mathcal{S}\}\,.$$

Said alternatively, $\tilde{x} \in B_{\mathcal{S}}(x)$ if it is obtained by replacing the tokens of x with indices belonging to $\mathcal{S}$ by arbitrary tokens in $\mathcal{D}$. We note that $B_{\mathcal{S}}(x)$ is a subset of the Hamming ball of radius $|S|$. Let us consider for instance the document $x =$ "the quick brown fox" and the set of perturbed indices $\mathcal{S} = \{2, 3\}$ Here, x has length $T = 4$, $|\mathcal{S}| = 2$, and an element of $B_{\mathcal{S}}(x)$ is the document $\tilde{x} =$ "the slow blue fox."

3. Warm-up: concatenation

Concatenation embeddings generally proceed by first mapping each token x_t of x to a vector $u(x_t, t) \in \mathbb{R}^d$. In a second step, these vector representations are concatenated together to form $\varphi(x)$. We assume that the representation $u(x_t, t)$ can be written as

$$u(x_t, t) = [u_e(x_t); u_p(t)] \in \mathbb{R}^d\,, \tag{3}$$

where $u_e \in \mathbb{R}^{d_e}$ denotes vector representations of individual tokens, while $u_p \in \mathbb{R}^{d_p}$ encodes positional information, and we define $d := d_e + d_p$.

Token embeddings. As noted in the introduction, there are essentially two widespread choices for u_e: either use *sparse* representations for individual tokens or use *dense* representations. The first approach is often synonymous with the use of *one-hot encodings*, hence considering the mapping $u_e : j \mapsto \mathbb{1}_j$ as a building brick, where, for any $j \in \mathcal{D}$, we define $\mathbb{1}_j$ the j-th vector of the canonical basis of $\mathbb{R}^D$. This has the advantage of simplicity. One caveat is that, although sparse, one-hot vectors have dimensionality $d_e = D$—the size of the dictionary. Regarding dense embeddings, as discussed in the introduction, the mapping $j \mapsto u_e(j)$ is learned from data and can encompass some semantic properties. In all these examples, $u_e(j)$ typically has dimensionality $d_e \ll D$ (for instance, `gensim` takes $d_e = 100$ in its `word2vec` implementation).

Positional embeddings. A common choice is to learn positional embeddings, jointly with token embeddings. It is also possible to use deterministic positional embeddings, such as one-hot vectors — $u_p(t) = \mathbb{1}_t \in \mathbb{R}^{T_{\max}}$, where $T_{\max}$ is a maximal document size, or more complicated functions of t. For instance, the original transformers architecture uses a sinusoidal transformation of t as positional embedding (Vaswani et al., 2017). Further, it is also possible to incorporate additional positional information in the embedding – for instance BERT incorporates segment position information corresponding to the index of the sentence the token belongs to (Devlin et al., 2018, Figure 2). Finally, one can simply ignore u_p altogether, relying simply on the order of the $u(x_t)$ to convey the positional information. Let us note that when $d_e = d_p$, one can add u_e and u_p in Eq. (3) instead of concatenating them, a possibility to which our analysis is robust.

Concatenation. For a given u, the embedding $\varphi(x)$ of a document x is formed by *concatenating* the $u(x_t, t)$s for $t \in [T]$. Formally, if $T \geq T_{\max}$, then the concatenation $\varphi(x)$ of $(x_1, \ldots, x_T)$ is defined as

$$\varphi(x) := [u(x_1, 1); \ldots ; u(x_{T_{\max}}, T_{\max})] \in \mathbb{R}^{dT_{\max}}\,,$$

and if $T < T_{\max}$, as (*zero-padding*),

$$\varphi(x) := [u(x_1, 1); \ldots ; u(x_T, T); 0; \ldots ; 0] \in \mathbb{R}^{dT_{\max}}\,.$$

Since the embedding is explicit in this case, it is straightforward to show the following:

Proposition 3.1 (Robustness of concatenation). *Let $x \in [D]^T$, $\mathcal{S} \subseteq [T]$, and $\tilde{x} \in B_{\mathcal{S}}(x)$. Then*

$$\|\varphi(x) - \varphi(\tilde{x})\| \leq \max_{j \neq k} \|u_e(j) - u_e(k)\| \cdot \sqrt{|\mathcal{S}| \wedge T_{\max}}\,.$$

In particular, for small perturbation of the input document, **concatenation is $1/2$-Hölder with respect to the Hamming distance**. Closer inspection of the proof reveals that the constant depends only on the perturbed tokens: if the changes made are close from the point of view of u_e, then $\varphi(x)$ and $\varphi(\tilde{x})$ remain close.

4. TF-IDF transform

Let x be a document of length T built on $\mathcal{D}$. In this section, we will assume that tokens correspond to individual words. Forgetting the sequential nature of natural language, one can simply look at the words appearing in x with repetitions – this is informally called a *bag-of-words* representation. Any given word $j \in \mathcal{D}$ appears in this representation with *multiplicity* $m_j(x)$. The TF-IDF transform of x is a vector $\varphi(x) \in \mathbb{R}^D$, with each coordinate of $\varphi(x)$ corresponding to a word of the dictionary. Component-wise, $\varphi(x)$ is a product of two terms: the *term frequency* f_j and the *inverse document frequency* v_j:

$$\forall j \in \mathcal{D}, \quad \begin{cases} f_j & := \frac{m_j}{T}\,, \\ v_j & := \log \frac{|\mathcal{C}|}{|\{z \in \mathcal{C} \text{ s.t. } j \in z\}|}\,, \end{cases} \tag{4}$$

where $\mathcal{C}$ is a set of documents. We will assume that $v_j > 0$. The exact expressions appearing in Eq. (4) can vary depending on implementation, we use here the most common definitions (in particular, they are the default choices used by `scikit-learn` (Pedregosa et al., 2011)). The (non-normalized) TF-IDF of x can be written $\varphi(x)_j = f_j v_j$ for all $j \in \mathcal{D}$. Intuitively, one wants to quantify the importance of each word in the document, while ignoring common words appearing in many documents such as articles. Finally, it is common to normalize $\varphi(x)$, generally using the Euclidean norm. We denote by $\phi(x) := \varphi(x)/\left\|\varphi(x)\right\|$ the normalized TF-IDF of x.

4.1. Robustness results

As we saw in the previous section, the TF-IDF transform of a given document can be given *in closed-form* as a function of the word multiplicities and the given coefficients. This allows a simple analysis, at least in the non-normalized case.

Proposition 4.1 (Robustness of non-normalized TF-IDF). *Let $x \in [D]^T$, $\mathcal{S} \subseteq [T]$, and $\tilde{x} \in B_{\mathcal{S}}(x)$. Let $m_{\max}$ be the maximal word multiplicity in x and $v_{\max}$ be the maximal inverse document frequency over $\mathcal{D}$. Then*

$$\left\|\varphi(x) - \varphi(\tilde{x})\right\| \leq 4 m_{\max} v_{\max} \frac{|\mathcal{S}|}{T} \, .$$

In other words, non-normalized **TF-IDF is Lipschitz continuous for the Hamming distance**, with Lipschitz constant inversely proportional to the common length of the documents. In reality, the dependency in T is slightly more complicated since nothing prevents $m_{\max}$ from being as large as T in pathological cases (when all the words of the document are identical). In any case, we uncover a satisfying fact about TF-IDF: **small changes in long documents do not matter much.** Taking into account the normalization, we have a similar result:

Proposition 4.2 (Robustness of normalized TF-IDF). *Let $x \in [D]^T$. Let $v_{\min}$ be the minimal inverse document frequency associated to the words of x. Let $\mathcal{S} \subseteq [T]$ such that $|\mathcal{S}| \leq \left\|\varphi(x)\right\| / (4 m_{\max} v_{\max})$ and $\tilde{x} \in B_{\mathcal{S}}(x)$. Then*

$$\left\|\phi(x) - \phi(\tilde{x})\right\| \leq \frac{4 m_{\max}^{1/2} v_{\max}^{1/2} D^{1/4}}{v_{\min}^{1/2}} \sqrt{\frac{|\mathcal{S}|}{T}} \, .$$

In plain words, **normalized TF-IDF is $1/2$-Hölder with respect to the Hamming distance**. Again, the constant appearing decreases with the length of the base document. A close inspection of the proof also reveals that the D is actually equal to $D(x)$, the size of the local dictionary.

4.2. Experimental validation

In order to check the accuracy of Proposition 4.2, we ran some numerical experiments. We considered movie reviews from the IMDB dataset as documents and the TF-IDF implementation from `scikit-learn` with L^2 normalization.

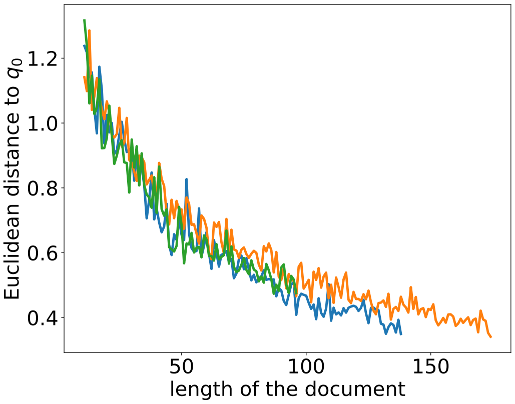

Figure 1. Normalized TF-IDF, influence of T. Documents of increasing length t, 5 random replacements. Proposition 4.2 gives a bound in $\mathcal{O}(1/\sqrt{T})$.

Influence of the document length. In a first set of experiments, we investigated the behavior of $\left\|\phi(x) - \phi(\tilde{x})\right\|$ with respect to the length T of x. To this extent, for several documents, we created a sequence of growing documents by considering the first t words of the documents, with t ranging from 5 to T. For each value of t, we replaced 5 words in the intermediary document and repeated this experiment several time. The words to replace were chosen uniformly at random in the document, and the replacements uniformly at random in $\mathcal{D}$, and we estimated the supremum of $\left\|\varphi(x) - \varphi(\tilde{x})\right\|$ by taking the maximum over these repetitions. Proposition 4.2 predicts that, since $|\mathcal{S}|$ is kept constant here, the supremum of $\left\|\varphi(x) - \varphi(\tilde{x})\right\|$ over all possible replacements should be upper bounded by $1/\sqrt{T}$ (up to numerical constants). This appears to be empirically true (see Figure 1).

Influence of the number of removals. In a second set of experiments, we looked at the dependency of $\left\|\phi(x) - \phi(\tilde{x})\right\|$ with respect to $|\mathcal{S}|$. This time keeping x fixed, we gradually increased the number of replaced words from 1 to T. Since T is fixed, Proposition 4.2 predicts that the supremum of $\left\|\phi(x) - \phi(\tilde{x})\right\|$ over all possible replacements should behave at most as $\sqrt{|\mathcal{S}|}$. This also appears to be empirically true, see Figure 2.

5. Paragraph Vector (`doc2vec`)

We now turn to the most challenging part of our analysis, `doc2vec`. On a high level, a token embedding matrix is learned jointly with a document embedding matrix on a corpus, aiming to predict correctly a missing token in a

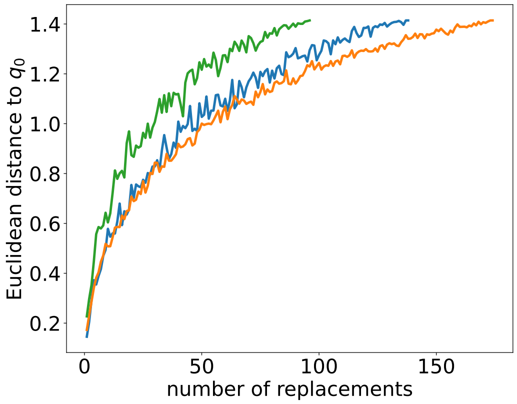

Figure 2. Normalized TF-IDF, influence of $|S|$. For a given document, s words are replaced at random with s ranging from 1 to T. Proposition 4.2 gives a bound in $\mathcal{O}\left(\sqrt{|S|}\right)$.

given context. The key difference with other vectorizers is that, at inference time, **another minimization problem is solved** by the model. Different documents yield different optimization problems, and therefore it is quite challenging to see where the resulting minimizer is located with respect to the original embedding.

5.1. A primer on doc2vec

The key idea underlying paragraph vector is neural probabilistic language modeling (Bengio et al., 2000): **predict words of a document** knowing (i) the **context** of the missing word in the document, and (ii) some **global information** about the document, **encoded as a vector** $q \in \mathbb{R}^d$. Thus the key concept is the probability of observing word j at position t given some context $c(t)$ and vector q. This is written informally as $\mathbb{P}\left(j|c(t), q\right)$, and we describe its exact formulation in the next paragraphs. Two models are proposed in Le & Mikolov (2014): *distributed memory* (PVDM) model, similar to the *continuous bag of words* model of Mikolov et al. (2013a), and *distributed bag of words* (PVDBOW) model, similar to the *skip gram* model. We first focus on the PVDM model, PVDBOW being a simplified version thereof, referring to Figure 3 for a visual help.

Local information. For a document x with length T, for any $\nu < t < T - \nu$, we define the *neighborhood* of t as

$$\gamma(t) := (t-\nu, \ldots, t-1, t+1, \ldots, t+\nu)\,. \tag{5}$$

Here, ν is an hyperparameter often called *context size* (or window size), quantifying the breath of the context considered by the model. To this neighborhood corresponds the *context*

$$c(t) := (x_{t-\nu}, \ldots, x_{t-1}, x_{t+1}, \ldots, x_{t+\nu})\,. \tag{6}$$

Intuitively, $c(t)$ corresponds to the tokens surrounding x_t in the document x. The tokens contained in $c(t)$ are then mapped to their one-hot representations, which are aggregated together. There are two natural ways to do this, either computing the *mean* (PVDMmean) or the *concatenation* of these vectors (PVDMconcat). Thus, at this stage, the local information at index t is summarized as a vector h_t, with

$$h_t := \frac{1}{2\nu}\sum_{s\in\gamma(t)} \mathbb{1}_{x_s} \in \mathbb{R}^D$$

if average is used, and

$$h_t := [\mathbb{1}_{x_{t-\nu}}; \ldots; \mathbb{1}_{x_{t-1}}; \mathbb{1}_{x_{t+1}}; \ldots; \mathbb{1}_{x_{t+\nu}}] \in \mathbb{R}^{2\nu D}$$

if concatenation is used (see bottom layer of Figure 3).

Projecting and lifting. This local information is then projected into $\mathbb{R}^d$, with $d \ll D$, the embedding space. At this stage, the document vector $q \in \mathbb{R}^d$ is added to the local representation. This intermediary representation is lifted back to $\mathbb{R}^D$. PVDM relies on two matrices P and R such that each context is mapped to

$$y_t := R(Ph_t + q) = \pi_t + Rq \in \mathbb{R}^D\,,$$

where $\pi_t := RPh_t \in \mathbb{R}^D$. Here, P has size $d \times D$ for PVDMmean, and $d \times 2\nu D$ for PVDMconcat, while R has size $D \times d$. When tokens are words, the columns of P are called *word vectors*, since they correspond to d dimensional embeddings for individual words. We refer to the intermediate layers of Figure 3 for a visual help.

Prediction. Finally, the prediction for x_t is encoded as the *softmax* of y_t, where the softmax $\sigma : \mathbb{R}^D \to \mathbb{R}^D$ is defined for $u \in \mathbb{R}^D$ as

$$\sigma(u) = \left(\frac{\mathrm{e}^{u_j}}{\sum_{k=1}^D \mathrm{e}^{u_k}}\right)_{1\leq j\leq D}\,. \tag{7}$$

In particular, all components of $\sigma(y_t)$ lie between 0 and 1 and sum to one, and reading coordinate j of $\sigma(y_t)$ can be interpreted as reading the predicted probability of token j. To summarize, $\sigma(y_t)$ encodes a discrete distribution over $\mathcal{D}$ that depends on the context of x_t and the document vector q (topmost layer of Figure 3).

Training. Let us call $x^{(1)}, \ldots, x^{(N)}$ the documents in our training set, with lengths $T_1, \ldots, T_N$. To each of these documents correspond an embedding $q^{(i)} \in \mathbb{R}^d$, which can be seen as the columns of a matrix $Q \in \mathbb{R}^{d\times N}$, each giving rise to $y_t^{(i)}$. The columns of Q are often referred to as *document vectors*. The key idea here is to learn P, Q, and R so that the predicted tokens at position t are accurate for all documents. Seeing $\sigma(y_t^{(i)})$ as a discrete probability distribution on $\mathcal{D}$,

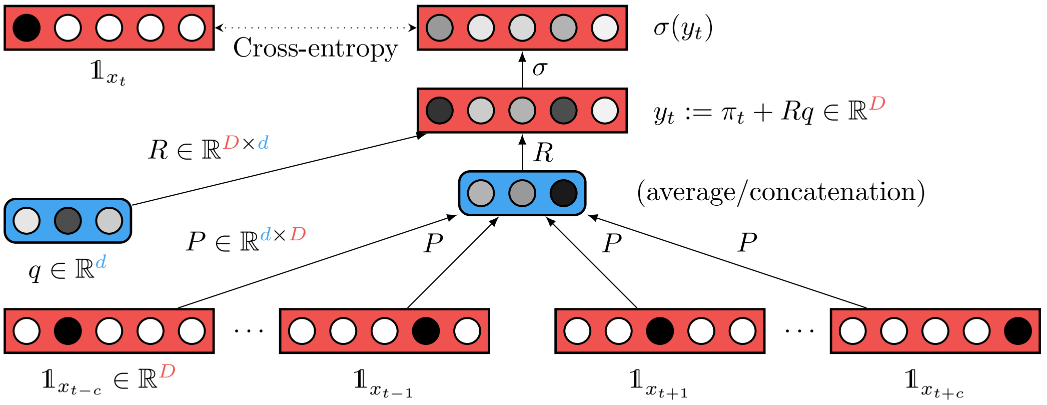

Figure 3. Overview of the doc2vec vectorizer, PVDM model. For a given document, for each token position t, the model considers the context $c(t)$ of x_t. The one-hot representation of the tokens in $c(t)$ (which are of size D) are either average or concatenated, then projected to the embedding layer ($\mathbb{R}^d$, in blue). At this stage, the document embedding $q \in \mathbb{R}^d$ is added to this local representation, which is then lifted back to $y_t \in \mathbb{R}^D$. Taking a softmax transform of y_t yields a discrete distribution on $\mathcal{D}$, which is compared to the truth (x_t) using cross-entropy (top part, dotted line). During training, PV minimizes objective (8) to find satisfying token embeddings P, document embedding q, and lifting R. At inference time, P and R are frozen and only q is allowed to vary.

a natural way to compare it to the groundtruth ($x_t^{(i)}$) is to compute the *cross-entropy* between the distribution putting mass one at $x_t^{(i)}$ and $\sigma(y_t^{(i)})$, that is,

$$\ell_t^{(i)} := -\log \sigma(y_t^{(i)})_{x_t^{(i)}} := \psi_{x_t^{(i)}}(y_t^{(i)}) \,,$$

where we defined $\psi := -\log \sigma$ coordinate-wise. The optimization problem solved by PV is written

$$\operatorname*{Minimize}_{P,Q,R} \sum_{i=1}^{N} \frac{1}{T_i} \sum_{t\in x^{(i)}} \psi_{x_t^{(i)}}(y_t^{(i)}) \,, \tag{8}$$

where $t \in x^{(i)}$ means t ranging from $\nu + 1$ to $T_i - \nu - 1$. Problem (8) is solved by stochastic gradient descent, or ADAM (Kingma & Ba, 2015).

Inference. Let us describe the embedding of a new document x, assuming that the model was trained on a corpus. The way inference works for the PV model is **to keep** P **and** R **fixed**, and to optimize solely in $q \in \mathbb{R}^d$

$$\operatorname*{Minimize}_{q\in\mathbb{R}^d} \frac{1}{T} \sum_{t\in x} \psi_{x_t}(y_t) \,. \tag{9}$$

An important observation is that $q \mapsto \psi_{x_t}(\pi_t + Rq)$ is a convex function, although not strictly (see Appendix). Therefore, a regularization term is often added to Eq. (9), a point which we will clarify in the next section. Also noting that q has only d parameters, solving PV inference (9) efficiently is not too challenging.

The case of PVDBOW. PVDBOW is another model falling under the PV umbrella. In a nutshell, following the idea of the distributed bag of word model, PVDBOW works the other way around and uses only the representation of the document to predict tokens. At position t, no local information is taken into account and we put $\pi_t = 0$ in that case. The predicted token distribution for the document is encoded as before (as $\sigma(y_t) = \sigma(Rq)$), and its quality also measured as $\psi_{x_t}(y_t)$ for all tokens in the document, leading to the same optimization problems. To summarize, PVDBOW is a simplified, lightweight version of PVDM, simply obtained by taking $\pi_t = 0$ in our framework. In particular, there is no matrix P, which leads to fewer parameters, and thus easier training and inference, a fact which was pointed out by Le & Mikolov (2014). Nevertheless, they recognize that PVDBOW still performs well as an embedding, and recommend considering as an embedding the concatenation of PVDM and PVDBOW.

Hierarchical softmax and negative sampling. In practice, as advocated by Le & Mikolov (2014), two additional expedients are used. First, the softmax is replaced by *hierarchical softmax* (Morin & Bengio, 2005). In a nutshell, each call of σ has a computational cost linear in D, which can be as large as 10^5 in practice. A solution is to replace the softmax by a tree-based approximation thereof, which computation is much faster. Second, following Mikolov et al. (2013a), it is common to incorporate tokens with a negative association to the token to predict when computing ℓ_t, leading to faster training. These two possibilities are non-trivial modifications to the PV model and we do not consider them in our analysis.

5.2. Robustness result

Before stating our robustness result, let us explain why it is challenging and outline the proof technique. As detailed in the previous section, the embedding of a document x of length T is found by solving

$$q_0 = \operatorname*{arg\,min}_{q \in \mathbb{R}^d} \left\{ F(q) + \frac{\alpha}{2} \left\| q \right\|^2 \right\} , \tag{10}$$

where $F(q) := \frac{1}{T} \sum_{t \in x} \psi_{x_t}(\pi_t + Rq)$. The regularization term $\alpha \left\| q \right\|^2 / 2$ with $\alpha > 0$ ensures uniqueness of the solution. Indeed, the softmax is invariant by translation by a vector proportional to $\mathbb{1}$, and solutions to (9) are not unique. As before, consider $\tilde{x}$, a modified version of x where tokens with indices in $\mathcal{S}$ have been replaced by others. The embedding q_1 of $\tilde{x}$ is found by solving

$$q_1 = \operatorname*{arg\,min}_{t \in x} \left\{ G(q) + \frac{\alpha}{2} \left\| q \right\|^2 \right\} , \tag{11}$$

where $G(q) := \frac{1}{T} \sum_{t \in x} \psi_{\tilde{x}_t}(\tilde{\pi}_t + Rq)$, and $\tilde{\pi}_t$ is defined analogously to π_t. The main challenge here is that q_0 **and** q_1 **are solutions of distinct optimization problems, which can be quite different if** $|\mathcal{S}|$ **is large**.

From discrete to continuous. The solution we propose to connect between these two problems is to interpolate smoothly between them. There are many ways to do this, and we settle for the simplest: linear interpolation. More precisely, we define for all $\mu \in [0, 1]$ and $q \in \mathbb{R}^d$ by

$$\Psi^{\mathrm{lin}}(\mu, q) := (1 - \mu) F(q) + \mu G(q) . \tag{12}$$

Subsequently, for all $\mu \in [0, 1]$, we can solve the following regularized optimization problem:

$$q(\mu) := \operatorname*{arg\,min}_{q \in \mathbb{R}^d} \left\{ \Psi^{\mathrm{lin}}(\mu, q) + \frac{\alpha}{2} \left\| q \right\|^2 \right\} , \tag{13}$$

giving rise to a continuous trajectory in the embedding space (see Figure 4 for an illustration). One can think of $q(\mu)$ as the embedding of a fictitious document traveling halfway between x and $\tilde{x}$ as μ ranges from 0 to 1.

Dynamics of interpolation. This approach is powerful, since it allows us to transform a problem which is discrete in nature (elements of a sum are modified) to a continuous one (time parameter varies). In particular, the dynamics of $\mu \mapsto q(\mu)$ **are described by an ordinary differential equation (ODE)**. Indeed, for all $\mu \in [0, 1]$, since $q \mapsto \Psi^{\mathrm{lin}}(\mu, q) + \frac{\alpha}{2} \left\| q \right\|^2$ is a strongly convex function, $q(\mu)$ is the (unique) critical point of $q \to \nabla \Psi^{\mathrm{lin}}(\mu, q) + \alpha q$, where ∇ denotes derivative with respect to the *space* coordinate (q). That is, for all $\mu \in [0, 1]$,

$$\nabla \Psi^{\mathrm{lin}}(\mu, q(\mu)) + \alpha q(\mu) = 0 .$$

Differentiating, we get that for all $\mu \in [0, 1]$,

$$\left(\nabla^2 \Psi^{\mathrm{lin}}(\mu, q(\mu)) + \alpha \, \mathrm{I} \right) q'(\mu) + \partial_\mu \nabla \Psi^{\mathrm{lin}}(\mu, q(\mu)) = 0 , \tag{14}$$

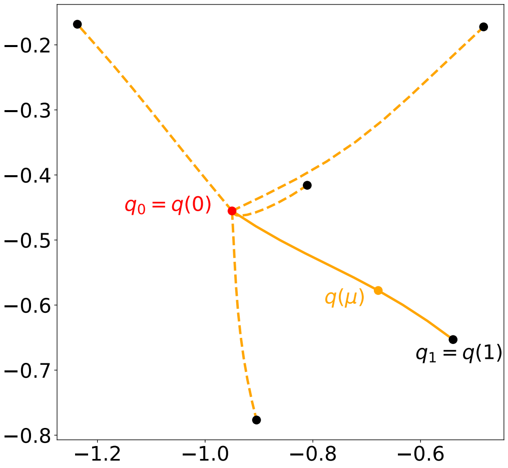

Figure 4. Continuously interpolating between q_0, the embedding of x (in red), and q_1, the embedding of $\tilde{x}$ (in black). Visualization in a 2-dimensional slice of $\mathbb{R}^d$. To each $\mu \in [0, 1]$ corresponds a solution to (13), appearing here as a point of the trajectory between q_0 and q_1 (solid orange line). Dynamics of this trajectory are described by Eq. (14). Different document perturbations lead to different embeddings and associated trajectories (dotted lines).

where g' denotes derivative with respect to the *time* coordinate (μ) and I the identity matrix. Let us set

$$\Phi^{\mathrm{lin}}(\mu, q) := - \left(\nabla^2 \Psi^{\mathrm{lin}}(\mu, q) + \alpha \, \mathrm{I} \right)^{-1} \partial_\mu \nabla \Psi^{\mathrm{lin}}(\mu, q) . \tag{15}$$

Then, Eq. (14) can be rewritten as $q'(\mu) = \Phi^{\mathrm{lin}}(q(\mu), \mu)$.

Spectrum of the Hessian of the log-softmax. Looking back at the ODE problem, it appears that one needs to understand precisely the behavior of Φ^{lin}. Intuitively, an ill-behaved function could lead to the explosion of the solution of the ODE, preventing the existence of reasonable bounds on $\left\| q(\mu) - q(0) \right\|$ for large μ. This understanding relies on the control of the smallest positive eigenvalue of $\nabla^2 \Psi^{\mathrm{lin}}$, $\lambda_1(\mu, q)$. Coming back to the definition of Ψ^{lin} (Eq. (12)), F, and G, we see that λ_1 closely related to $\lambda_{\min}$, the smallest positive eigenvalue of the Hessian of the log-softmax, for which we have precise results (Lemma H.5 and Theorem H.9).

Grönwall-type result. Once that a precise control is achieved on Φ^{lin}, one may have hoped to use standard Grönwall type inequalities such as Pachpatte (2004) to obtain quantitative bounds on $\left\| q(1) - q(0) \right\|$. However, in our setting, the growth of Φ^{lin} prevents us from getting explicit bounds and we had to prove a new result (Theorem G.1) which is actually true in a more general setting than that of `doc2vec`. Specifying this result, we get:

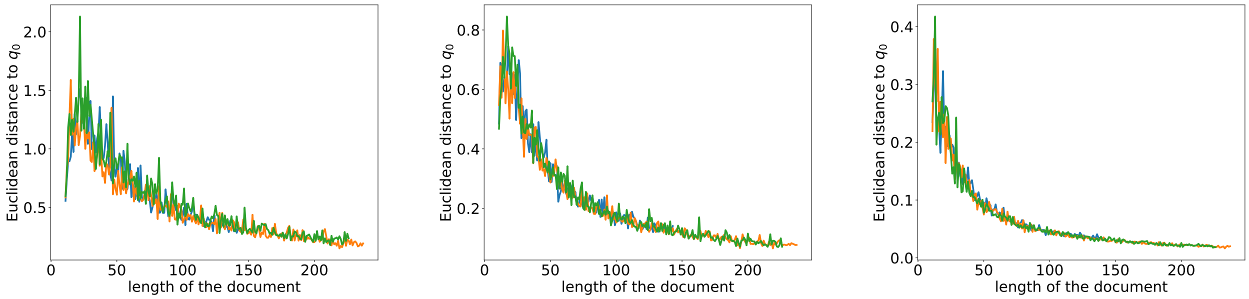

Figure 5. Influence of the length of the document on the robustness of `doc2vec`. Five random replacements, from left to right: PVDMmean, PVDMconcat, and PVDBOW.

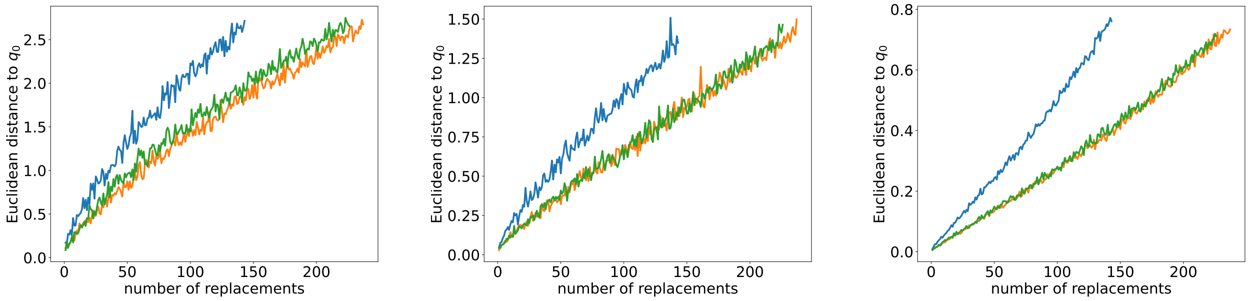

Figure 6. Influence of the number of words replaced on the robustness of `doc2vec`. From left to right: PVDMmean, PVDMconcat, and PVDBOW.

Theorem 5.1 (Bounded trajectories). *Let $x \in [D]^T$, $\mathcal{S} \subseteq [T]$, and $\tilde{x} \in B_{\mathcal{S}}(x)$. Suppose that $R \in \mathbb{R}^{D \times d}$ is such that $\sigma_{\min}(R) > 0$ and $\operatorname{Im}(R) \subset \mathbb{1}^{\perp}$. Let $\mu \mapsto q(\mu)$ be the solution of ODE* (14). *Then, there exist two constants $c = c(\alpha) > 0$ and $L = L(\|q(0)\|) > 0$ depending explicitly on P, R, ν, and D such that, whenever $|\mathcal{S}|/T \leq c$,*

$$\sup_{\mu \in [0,1]} \|q(\mu) - q(0)\| \leq L \frac{|\mathcal{S}|}{T} \,.$$

Since $\varphi(x) = q(0)$ and $\varphi(\tilde{x}) = q(1)$, a corollary of Theorem 5.1 is that the **`doc2vec` embedding is Lipschitz continuous with respect to the Hamming distance**, with Lipschitz constant at most inversely proportional to the document lengTheorem Coming back to our initial question, Theorem 5.1 guarantees that, for documents of reasonable length and small perturbations, `doc2vec` embeddings can not vary too greatly. We emphasize that Theorem 5.1 is true for all three `doc2vec` models.

The key assumption here is that $|\mathcal{S}|$ is small enough. We argue that it is only natural to ask so: indeed, if one is allowed to modify every single token of x, this yield a completely different document (although having the same length), which could *a priori* be embedded anywhere. The other main assumptions concern the matrix R. Experimentally, we observe that $\sigma_{\min}(R) > 0$ holds (see Section I.3). Requiring that $\operatorname{Im}(R) \subset \mathbb{1}^{\perp}$ is not too restricting: because of the translation invariance by $\mathbb{1}$ of the softmax, one can always normalize R by removing the average line from each line. The main limitation of Theorem 5.1 is the dependency of c and L in the problems parameters. Exact expression can be found in Appendix (Theorem F.7).

5.3. Experimental validation

In order to verify the validity of Theorem 5.1, we ran similar experiments to those presented in Section 4. We considered again movie reviews from the IMDB dataset. As vectorizer, we trained `doc2vec` models from scratch on a subset of the IMDB dataset (10^3 reviews). The associated dictionary has size $D = 18,416$: we took tokens as words of the English dictionary. Note that one can also consider sub-word tokens, but in that case replacing a word in the document usually implies replacing several tokens. We chose $d = 50$ as dimension of the embedding. We took $\nu = 5$ as context size parameter.

We present results of experiments regarding the influence of the document length in Figure 5. Theorem 5.1 predicts that, since $|\mathcal{S}|$ is kept constant here, the supremum of $\|\varphi(x) - \varphi(\tilde{x})\|$ over all replacements should be upper bounded by $1/T$ (up to numerical constants). This appears to be empirically true.

We present results of experiments regarding the influence of the number of replaced words in Figure 6. Here we took the number of replaced words from $\nu + 1$ to $T - \nu - 1$ to avoid for border effects. Since T is fixed, Theorem 5.1 predicts that the supremum of $\|\varphi(x) - \varphi(\tilde{x})\|$ over all possible replacements should behave at most linearly in $|\mathcal{S}|$. This appears to be empirically true.

We present in Appendix (Section I) additional results with another implementation, `gensim` (Řehůřek & Sojka, 2010). In particular, this implementation uses hierarchical softmax. The results are consistent with the behavior presented here.

6. Conclusion

In this paper, we proved that several popular text vectorizers are robust, in the sense that they are either Lipschitz or Hölder continuous with respect to the Hamming distance. Proving this robustness was possible for concatenation and TF-IDF thanks to elementary computations, but required a much more challenging mathematical analysis for `doc2vec` requiring two new results (local Lipschitz continuity of the softmax and a new Grönwall–Bellman–Bahouri non-explosion lemma).

Let us outline future research directions. First, we studied the robustness of the *true* solution of (8) and (9). In practice, this problem is solved thanks to gradient descent, and it would be interesting to measure the impact of this approximation. A second line of work would consist in obtaining refined results when we put a random model on the distribution of the words of the document, similarly to what is done in (Arora et al., 2016).

Acknowledgements

This work was supported by the French government under the management of the Agence Nationale de la Recherche, grant agreements GraVa ANR-18-CE40-0005, NEMATIC ANR-21-CE45-0010, and NIM-ML ANR-21-CE23-0005-01. D.G. acknowledges the support of EU Horizon 2020 project AI4Media (contract no. 951911) and would like to thank Charbel Yachouchi for his preliminary work. S.V. would like to thank Nicolas Patry for fruitful discussion about NLP embeddings.

References

Agarwal, R. P., Deng, S., and Zhang, W. Generalization of a retarded Gronwall-like inequality and its applications. *Appl. Math. Comput.*, 165(3):599–612, 2005.

Alghamdi, W., Hsu, H., Jeong, H., Wang, H., Michalak, P. W., Asoodeh, S., and Calmon, F. P. Beyond adult and compas: Fairness in multi-class prediction. In *NeurIPS*, 2022.

Arnold, V. I. *Ordinary Differential Equations*. MIT Press, 1978.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. A latent variable model approach to PMI-based word embeddings. *TACL*, 4:385–399, 2016.

Bailleul, I. and Catellier, R. Non-explosion criteria for rough differential equations driven by unbounded vector fields. *Ann. Fac. Sci. Toulouse Math.*, 29(3):721–759, 2020.

Bengio, Y., Ducharme, R., and Vincent, P. A neural probabilistic language model. In *NeurIPS*, 2000.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *TACL*, 5: 135–146, 2017.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *NeurIPS*, 2020.

Bubeck, S. and Sellke, M. A universal law of robustness via isoperimetry. In *NeurIPS*, 2021.

Bubeck, S., Eldan, R., Lee, Y. T., and Mikulincer, D. Network size and size of the weights in memorization with two-layers neural networks. In *NeurIPS*, 2020.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In *ICML*, 2017.

Dannan, F. M. Integral inequalities of Gronwall-Bellman-Bihari type and asymptotic behavior of certain second order nonlinear differential equations. *J. Math. Anal. Appl.*, 108(1):151–164, 1985.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2018.

Gage, P. A new algorithm for data compression. *C Users Journal*, 12(2):23–38, 1994.

Gao, B. and Pavel, L. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Garreau, D. and Mardaoui, D. What does LIME really see in images? In *ICML*, 2021.

Hein, M. and Andriushchenko, M. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *NeurIPS*, 2017.

Jones, K. S. A statistical interpretation of term specificity and its application in retrieval. *J. Doc.*, 1972.

Kim, Y.-H. Gronwall, Bellman and Pachpatte type integral inequalities with applications. *Nonlinear Anal. Theory Methods Appl.*, 71(12):2641–2656, 2009.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. Skip-thought vectors. In *NeurIPS*, 2015.

Le, Q. and Mikolov, T. Distributed representations of sentences and documents. In *ICML*, 2014.

Li, Y., Xu, L., Tian, F., Jiang, L., Zhong, X., and Chen, E. Word embedding revisited: A new representation learning and explicit matrix factorization perspective. In *AISTAT*, 2015.

Logeswaran, L. and Lee, H. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, 2018.

Luhn, H. P. A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.*, 1 (4):309–317, 1957.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. *NeurIPS*, 2013b.

Morin, F. and Bengio, Y. Hierarchical probabilistic neural network language model. In *AISTAT*, 2005.

Pachpatte, B. G. On some new nonlinear retarded integral inequalities. *J. Inequal. Pure Appl. Math*, 5(3):80, 2004.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *JMLR*, 12: 2825–2830, 2011.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *EMNLP*, 2014.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In *NAACL*, 2018.

Řehůřek, R. and Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725. Association for Computational Linguistics, 2016.

Steele, J. M. *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *ICLR*, 2014.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Wei, D., Wu, H., Wu, M., Chen, P.-Y., Barrett, C., and Farchi, E. Convex bounds on the softmax function with applications to robustness verification. In *AISTATS*, 2023.

Weng, L., Zhang, H., Chen, H., Song, Z., Hsieh, C.-J., Daniel, L., Boning, D., and Dhillon, I. Towards fast computation of certified robustness for ReLU networks. In *ICML*, 2018a.

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 *ICLR*, 2018b.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Xu, H. and Mannor, S. Robustness and generalization. *Mach. Learn.*, 86(3):391–423, 2012.

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., and Li, C. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(3):1–41, 2020.

A. General organization

This Appendix is organized as follows: in Section B (resp. C) we collect the missing proofs for Section 3 (resp. 4) of the main paper.

The next five sections are dedicated to the proof of Theorem 5.1: First, in Section D, we formally prove that the dynamics of the interpolation scheme between two minimizers follow an ordinary differential equation (ODE). We actually show a more general result and provide technical conditions on the interpolation Ψ under which we are able to formulate the interpolation between minimization problems as an ODE. Next, in Section E, we derive quantitative bounds for the solution of this ODE. We show how to specialize this result in the `doc2vec` setting in Section F, proving Theorem 5.1 in the process. The main tool used to obtain these bounds is a general Grönwall-Bellman-Bahouri type result for ODE with exponentially-growing coefficients. This result (Theorem G.1), as well as all other technical results concerning ODEs, is stated and proved in Section G . In order to specialize our result to the `doc2vec` setting, we needed a fine-grained study of the (log-)softmax function. In particular, we derive a new bound on the softmax function (Theorem H.9), which is proved in Section H.

We conclude this Appendix with additional experimental results supporting our claims in Section I.

B. Omitted proofs for concatenation

B.1. Proof of Proposition 3.1

By definition of φ and Pythagoras theorem,

$$\left\|\varphi(x)-\varphi(\tilde{x})\right\|^2 = \sum_{t\in\mathcal{S}\cap[T_{\max}]} \left\|u(x_t,t)-u(\tilde{x}_t,t)\right\|^2 \, .$$

By definition of u (Eq. (3)), one has

$$u(x_t,t)-u(\tilde{x}_t,t) = \left[u_e(x_t)-u(\tilde{x}_t);0\right] , \tag{16}$$

and therefore

$$\left\|u(x_t,t)-u(\tilde{x}_t,t)\right\|^2 = \left\|u_e(x_t)-u(\tilde{x}_t)\right\|^2 \, .$$

We deduce that

$$\left\|\varphi(x)-\varphi(\tilde{x})\right\|^2 \leq \left|\mathcal{S}\cap[T_{\max}]\right| \cdot \max_{j\neq k} \left\|u_e(j)-u_e(k)\right\|^2 \, .$$

□

Remark B.1 (Concatenation *v.s.* sum). Replacing the concatenation by a sum in the definition of u (Eq. (3)) does not change the proof. Indeed, the key step Eq. (16) remains unchanged in that case: the key idea here is that position tokens are the same for words in the same position, and cancel out when forming the difference.

C. Omitted proofs for TF-IDF vectorization

C.1. Proof of Proposition 4.1

By definition, we can write

$$\varphi(x) = \sum_{j=1}^{D} f_j v_j \mathbb{1}_j = \frac{1}{T}\sum_{j=1}^{D} m_j v_j \mathbb{1}_j \, .$$

Similarly, since $\tilde{x}$ has same length as x,

$$\varphi(\tilde{x}) = \frac{1}{T}\sum_{j=1}^{D} \tilde{m}_j v_j \mathbb{1}_j \, ,$$

where we let $\tilde{m}_j$ denote the multiplicity of word j in document $\tilde{x}$. We deduce that

$$\left\|\varphi(x)-\varphi(\tilde{x})\right\|^2 = \frac{1}{T^2}\sum_{j=1}^{D} (m_j-\tilde{m}_j)^2 v_j^2 \, .$$

By letting $v_{\max}$ be the maximal inverse document frequency on $\mathcal{D}$, we already see that

$$\|\varphi(x) - \varphi(\tilde{x})\|^2 \leq \frac{v_{\max}^2}{T^2} \sum_{j=1}^{D} (m_j - \tilde{m}_j)^2 \, .$$

In the previous display, only terms such that $m_j \neq \tilde{m}_j$ count. Using the inequality between p-norms, we have

$$\sum_{m_j \neq \tilde{m}_j} (m_j - \tilde{m}_j)^2 \leq \left(\sum_{m_j \neq \tilde{m}_j} |m_j - \tilde{m}_j| \right)^2 \, .$$

Now, by the triangle inequality,

$$\sum_{m_j \neq \tilde{m}_j} |m_j - \tilde{m}_j| \leq \sum_{m_j \neq \tilde{m}_j} m_j + \sum_{m_j \neq \tilde{m}_j} \tilde{m}_j \, .$$

We notice that these two sums are equal: every removed word has to appear somewhere. Moreover, $|\{j \text{ s.t. } m_j \neq \tilde{m}_j| \leq 2\,|\mathcal{S}|$, since modifying one word changes at most two multiplicities, and this happens at most $|\mathcal{S}|$ times. Therefore, we have proved that

$$\sum_{m_j \neq \tilde{m}_j} |m_j - \tilde{m}_j| \leq 4 m_{\max} |\mathcal{S}| \, , \tag{17}$$

where we recall that $m_{\max}$ is the maximal multiplicity of words of x. Backtracking, we have

$$\|\varphi(x) - \varphi(\tilde{x})\|^2 \leq \frac{v_{\max}^2}{T^2} \cdot 16 m_{\max}^2 |\mathcal{S}|^2 \, ,$$

and we can conclude by simply taking the square root of this last display. □

C.2. Proof of Proposition 4.2

We notice that

$$\|\phi(x) - \phi(\tilde{x})\|^2 = 1 + 1 - 2\phi(x)^\top \phi(\tilde{x}) = 2 - 2\frac{\varphi(x)^\top \varphi(\tilde{x})}{\|\varphi(x)\| \, \|\varphi(\tilde{x})\|} \, . \tag{18}$$

In this last term we recognize the *cosine similarity* between $\varphi(x)$ and $\varphi(\tilde{x})$. Since we are working under the assumptions of Lemma C.1, we have

$$\frac{\varphi(x)^\top \varphi(\tilde{x})}{\|\varphi(x)\| \, \|\varphi(\tilde{x})\|} \geq 1 - \frac{8 m_{\max} v_{\max} |\mathcal{S}|}{\|\varphi(x)\|} \, .$$

Coming back to Eq. (18), we see that

$$\|\phi(x) - \phi(\tilde{x})\|^2 \leq \frac{16 m_{\max} v_{\max} |\mathcal{S}|}{\|\varphi(x)\|} \, .$$

We conclude by using Lemma C.2 and taking the square root. □

C.3. Auxilliary results

We have the following result, key to the proof of Prop. 4.2, and of independent interest:

Lemma C.1 (Cosine similarity robustness). *Let x be a document. Let $\mathcal{S} \subseteq [T]$ such that $|\mathcal{S}| \leq \|\varphi(x)\| / (4 m_{\max} v_{\max})$ and $\tilde{x} \in B_{\mathcal{S}}(x)$. Then*

$$\frac{\varphi(x)^\top \varphi(\tilde{x})}{\|\varphi(x)\| \, \|\varphi(\tilde{x})\|} \geq 1 - \frac{8 m_{\max} v_{\max} |\mathcal{S}|}{\|\varphi(x)\|} \, . \tag{19}$$

Proof. By homogeneity, we can multiply numerator and denominator in Eq. (19) by T and deal with multiplicities instead of frequencies in this proof. We first focus on the numerator and write

$$\varphi(x)^\top \varphi(\tilde{x}) = \varphi(x)^\top (\varphi(x) + \varphi(\tilde{x}) - \varphi(x)) = \|\varphi(x)\|^2 + \sum_{j=1}^{D} m_j (\tilde{m}_j - m_j) v_j^2 \, , \tag{20}$$

by definition of φ. Using Cauchy-Schwarz inequality, we find that

$$\sum_{j=1}^{D} m_j(m_j - \tilde{m}_j)v_j^2 \leq \sqrt{\sum_j m_j v_j^2}\sqrt{\sum_j (m_j - \tilde{m}_j)^2 v_j^2}\,.$$

In the first part of the right-hand side we recognize $\|\varphi(x)\|$, and in the second part, the same quantity bounded in the proof of Proposition 4.1. We deduce that

$$\sum_j m_j(m_j - \tilde{m}_j)v_j^2 \leq \|\varphi(x)\| \cdot 4m_{\max}v_{\max}\left|\mathcal{S}\right| \,.$$

Coming back to Eq. (20), we have proved that

$$\varphi(x)^\top \varphi(\tilde{x}) \geq \|\varphi(x)\|^2 - 4m_{\max}v_{\max}\|\varphi(x)\|\left|\mathcal{S}\right| \,,$$

which is positive under our assumption. Let us now look into the denominator of Eq. (19). Using the triangle inequality and Proposition 4.1, we write

$$\|\varphi(\tilde{x})\| \leq \|\varphi(x)\| + 4m_{\max}v_{\max}\left|\mathcal{S}\right| \,.$$

Putting everything together, we have

$$\frac{\varphi(x)^\top \varphi(\tilde{x})}{\|\varphi(x)\|\,\|\varphi(\tilde{x})\|} \geq \frac{\|\varphi(x)\|^2 - 4m_{\max}v_{\max}\|\varphi(x)\|\left|\mathcal{S}\right|}{\|\varphi(x)\| \cdot (\|\varphi(x)\| + 4m_{\max}v_{\max}\left|\mathcal{S}\right|)} = \frac{1-u}{1+u}\,,$$

with $u := 4m_{\max}v_{\max}\left|\mathcal{S}\right| / \|\varphi(x)\|$. Again, by our assumption, $u \in (0,1)$. It is straightforward to show that $(1-u)/(1+u) \geq 1 - 2u$ for all $u \in (0,1)$, and we deduce the result. □

We also have the following:

Lemma C.2 (Lower bound on $\|\varphi(x)\|$). *Let x be a document. Let $v_{\min}$ be the minimum inverse document frequency for words contained in x and $D(x)$ the size of the local dictionary. Then*

$$\|\varphi(x)\| \geq \frac{Tv_{\min}}{\sqrt{D(x)}}\,.$$

Proof. Straightforward from the definitions and the comparison of p-norms. □

D. Dynamics of interpolation

Recall that we are considering, for all $\mu \in [0,1]$, the following minimization problem:

$$q(\mu) := \operatorname*{arg\,min}_{q \in \mathbb{R}^d} \left\{ \Psi^{\text{lin}}(\mu, q) + \frac{\alpha}{2}\|q\|^2 \right\} \,. \tag{21}$$

In this section, we show that under mild regularity assumptions on Ψ, q is the unique solution of the following ODE:

$$\left(\nabla^2 \Psi^{\text{lin}}(\mu, q(\mu)) + \alpha\,\mathrm{I}\right) q'(\mu) + \partial_\mu \nabla \Psi^{\text{lin}}(\mu, q(\mu)) = 0\,. \tag{22}$$

Notation. For any matrix $M \in \mathbb{R}^{A \times B}$, let us define the *operator norm* of M as

$$\|M\|_{\text{op}} := \sup\left\{ \frac{\|Mv\|}{\|v\|}, v \in \mathbb{R}^B \setminus \{0\} \right\} \,.$$

For any $\rho > 0$, we also define $B_d(\rho)$ the open Euclidean ball of center 0 and radius ρ. Finally, for $a_1, a_2 > 0$, define $a_1 \vee a_2 := \max(a_1, a_2)$.

We can now state the required assumptions on Ψ.

Assumption D.1 (Convexity). Let $d \geq 1$. We suppose that $\Psi \in \mathcal{C}^{1,2}([0,1] \times \mathbb{R}^d; \mathbb{R})$ and that, for all $(\mu, q) \in [0,1] \times \mathbb{R}^d$, $\nabla^2 \Psi(\mu, q)$ is a positive semi-definite matrix.

Since $\alpha > 0$, A.D.1 this guarantees that $q(\mu)$ is uniquely-defined for each μ. Next, we define some quantities related to the local Lipschitz continuity of Ψ and its derivatives.

Definition D.2 (Local Lipschitz semi-norms). Let $\Psi \in \mathcal{C}^{1,2}([0,1] \times \mathbb{R}^d; \mathbb{R})$. For all $\rho > 0$, let us define

$$L_1(\rho) := \sup_{\substack{\mu \in [0,1] \\ q \neq \tilde{q} \in B_d(0,\rho)}} \frac{\left\| \nabla^2 \Psi(\mu, q) - \nabla^2 \Psi(\mu, \tilde{q}) \right\|_{\mathrm{op}}}{\|q - \tilde{q}\|}, \quad L_2(\rho) := \sup_{\substack{\mu \in [0,1] \\ q \neq \tilde{q} \in B_d(0,\rho)}} \frac{\left\| \partial_\mu \nabla \Psi(\mu, q) - \partial_\mu \nabla \Psi(\mu, \tilde{q}) \right\|}{\|q - \tilde{q}\|}, \tag{23}$$

and

$$M(\rho) := \sup_{\substack{\mu \in [0,1] \\ q \in B_d(0,\rho)}} \left\| \partial_\mu \nabla \Psi(\mu, q) \right\| . \tag{24}$$

Our second assumption on Ψ at this stage is that these quantities are all finite.

Assumption D.3 (Global Lipschitz continuity). Let $\Psi \in \mathcal{C}^{1,2}([0,1] \times \mathbb{R}^d; \mathbb{R})$. Suppose that

$$\sup_{\rho > 0} \big(L_1(\rho) + L_2(\rho) \big) < +\infty \quad \text{and} \quad \sup_{\rho > 0} M(\rho) < +\infty,$$

where $L_1(\rho)$, $L_2(\rho)$, and $M(\rho)$ are defined in Eq. (23) and Eq. (24).

In this setting, we are able to prove the following result:

Theorem D.4 (Equivalence ODE/minimization problem). *Assume that Ψ satisfies A.D.1 and A.D.3. Then $\mu \mapsto q(\mu)$ is differentiable on $[0,1]$, and q is the unique solution of Eq. (22).*

Note that under assumption A.D.1 the matrix $\nabla^2 \Psi(\mu, q) + \alpha \, \mathrm{I}$ is invertible. One can then rewrite Eq. (22) in a more standard form, namely

$$q'(\mu) = - \left(\nabla^2 \Psi(\mu, q(\mu)) + \alpha \, \mathrm{I} \right)^{-1} \partial_\mu \nabla \Psi(\mu, q(\mu)) \, . \tag{25}$$

Thus, to study the ODE problem, one needs the regularity properties (local Lipschitz continuity, boundedness...) of the function

$$\Phi : (\mu, q) \in [0,1] \times \mathbb{R}^d \mapsto \Phi(\mu, q) := - \left(\nabla^2 \Psi(\mu, q) + \alpha \, \mathrm{I} \right)^{-1} \partial_\mu \nabla \Psi(\mu, q) \, . \tag{26}$$

The interplay between $\partial_\mu \nabla \Psi$ and $\nabla^2 \Psi$ here is crucial. Indeed, in Section F we will see that when specified in the doc2vec case, the term in ∂_μ gives the desired quantity $\frac{|S|}{T}$ whereas the term in $\nabla^2 \Psi$ has to be handled using precise properties on the softmax function. Theorem D.4 is standard in the ODE literature and holds as soon as the quantities appearing in Eq. (25) are well-behaved. More precisely, this is the case $c = 0$ of Theorem G.1 in Section G. We now simply check that the assumptions of Theorem G.1 are satisfied in the setting of Theorem D.4. This is achieved by Lemma D.5 and Lemma D.6. We start by a result upper bounding the norm of the inverse Hessian.

Lemma D.5 (Norm of inverse Hessian). *Let $\Psi : [0,1] \times \mathbb{R}^d \to \mathbb{R}$. Assume that A.D.1 holds. Then,*

$$\forall \mu, q \in [0,1] \times \mathbb{R}^d, \qquad \left\| (\nabla^2 \Psi(\mu, q) + \alpha \, \mathrm{I})^{-1} \right\|_{\mathrm{op}} \leq \frac{1}{\alpha} \, . \tag{27}$$

The proof of Lemma D.5 exploits the fact that $\nabla^2 \Psi$ is a non-negative symmetric matrix and can be diagonalized in orthonormal basis with non-negative eigenvalues. The regularization of the minimization problem with the addition of the term $\frac{\alpha}{2} \|q\|$ can be translated with the addition of the term $\alpha \, \mathrm{I}$ to the previous Hessian matrix, which then becomes a positive definite symmetric matrix. One then only has to estimate the smallest eigenvalue of the matrix to conclude.

Proof. By A.D.1, for all $\mu \in [0,1]$, $q \mapsto \Psi(\mu, q)$ is convex and, for any $\mu, q \in [0,1] \times \mathbb{R}^d$, $\nabla^2 \Psi(\mu, q)$ is a positive semi-definite matrix with non-negative eigenvalues. From these, $N_0(\mu, q) = \mathrm{Rank}\left(\nabla^2 \Psi(\mu, q) \right)$ of them are non-zero, and they can be ranked as

$$0 < \lambda_1(\mu, q) \leq \cdots \leq \lambda_{N_0(\mu,q)}(\mu, q) \, .$$

Moreover, there exists an orthogonal matrix $P(\mu, q)$ (meaning that $P(\mu, q)P(\mu, q)^\top = \mathrm{I}$) such that

$$P(\mu, q)\nabla^2\Psi(\mu, q)P(\mu, q)^\top = \mathrm{diag}(0, \dots, 0, \lambda_1(\mu, q), \dots, \lambda_{N_0}(\mu, q))\,.$$

Furthermore since $\nabla^2\Psi(\mu, q)$ is a symmetric matrix, its range and its kernel are orthogonal complements, $\mathrm{Ker}\left(\nabla^2\Psi(\mu, q)\right) \oplus^\perp \mathrm{Im}(\nabla^2\Psi(\mu, q)) = \mathbb{R}^d$ and

$$h \in \mathrm{Im}(\nabla^2\Psi(\mu, q)) \quad \text{if, and only if,} \quad P(\mu, q)h = (0, \dots, 0, h_1, \cdots, h_{N_0}).$$

Hence

$$P(\mu, q)\left(\nabla^2\Psi(\mu, q) + \alpha\,\mathrm{I}\right)P(\mu, q)^\top = \mathrm{diag}(\alpha, \dots, \alpha, \lambda_1(\mu, q) + \alpha, \dots, \lambda_{N_0}(\mu, q) + \alpha)\,,$$

which implies that $\nabla^2\Psi(\mu, q) + \alpha\,\mathrm{I}$ is an invertible positive definite matrix such that

$$P(\mu, q)\left(\nabla^2\Psi(\mu, q) + \alpha\,\mathrm{I}\right)^{-1}P(\mu, q)^\top = \mathrm{diag}\left(\frac{1}{\alpha}, \dots, \frac{1}{\alpha}, \frac{1}{\lambda_1(\mu, q) + \alpha}, \dots, \frac{1}{\lambda_{N_0}(\mu, q) + \alpha}\right)\,.$$

From the last display, one readily sees that the maximum eigenvalue of $\left(\nabla^2\Psi(\mu, q) + \alpha\,\mathrm{I}\right)^{-1}$ is $1/\alpha$, proving our claim. □

The next lemma shows how regularity assumptions on Ψ translate into regularity conditions for Φ.

Lemma D.6 (Global-Lispchitz continuity of Φ). *Let Ψ such that A.D.1 and A.D.3 hold. Then Φ is globally Lipschitz continuous in q uniformly in $\mu \in [0, 1]$. Moreover, for all $\rho > 0$,*

$$\sup_{\substack{\mu\in[0,1]\\ q\neq\tilde{q}\in\mathbb{R}^d}} \frac{\|\Phi(\mu, q) - \Phi(\mu, \tilde{q})\|}{\|q - \tilde{q}\|} \leq \frac{1}{\alpha}\left(\sup_{\rho>0} L_2(\rho) + \frac{\left(\sup_{\rho>0} L_1(\rho)\right)\left(\sup_{\rho>0} M(\rho)\right)}{\alpha}\right)\,. \tag{28}$$

The proof of Lemma D.6 relies on the following identity, which is true for any non-negative symmetric matrices $A, B \in \mathbb{R}^{d\times d}$ and vectors $X, Y \in \mathbb{R}^d$:

$$(A + \alpha\,\mathrm{I})^{-1}X - (B + \alpha\,\mathrm{I})^{-1}Y = -(A + \alpha\,\mathrm{I})^{-1}(A - B)(B + \alpha\,\mathrm{I})^{-1}X + (B + \alpha\,\mathrm{I})^{-1}(X - Y)\,. \tag{29}$$

Lemma D.5 allows us to conclude.

Proof. Let $q, \tilde{q} \in B_d(0, \rho)$. Using Eq. (29), we have

$$\begin{aligned}\Phi(\mu, q) - \Phi(\mu, \tilde{q}) =& -\left((\nabla^2\Psi(\mu, q) + \alpha I)^{-1} - (\nabla^2\Psi(\mu, \tilde{q}) + \alpha I)^{-1}\right)\partial_\mu\nabla\Psi(\mu, q)\\ &- (\nabla^2\Psi(\mu, \tilde{q}) + \alpha I)^{-1}\left(\partial_\mu\nabla\Psi(\mu, q) - \partial_\mu\nabla\Psi(\mu, \tilde{q})\right)\\ =& -(\nabla^2\Psi(\mu, q) + \alpha I)^{-1}\left(\nabla^2\Psi(\mu, \tilde{q}) - \nabla^2\Psi(\mu, q)\right)(\nabla^2\Psi(\mu, \tilde{q}) + \alpha I)^{-1}\partial_\mu\nabla\Psi(\mu, q) \qquad (30)\\ &- (\nabla^2\Psi(\mu, \tilde{q}) + \alpha I)^{-1}\left(\partial_\mu\nabla\Psi(\mu, q) - \partial_\mu\nabla\Psi(\mu, \tilde{q})\right)\,. \qquad (31)\end{aligned}$$

Taking the norm and using Lemma D.5 (in particular Inequality (27)), we have for $\rho = \|q\| \vee \|\tilde{q}\|$,

$$\begin{aligned}\|\Phi(\mu, q) - \Phi(\mu, \tilde{q})\| \leq& \frac{1}{\alpha^2}\left\|\nabla^2\Psi(\mu, q) - \nabla^2\Psi(\mu, \tilde{q})\right\|_{\mathrm{op}}\|\partial_\mu\nabla\Psi(\mu, q)\|\\ &+ \frac{1}{\alpha}\|\partial_\mu\nabla\Psi(\mu, q) - \partial_\mu\nabla\Psi(\mu, \tilde{q})\|\\ \|\Phi(\mu, q) - \Phi(\mu, \tilde{q})\| \leq& \frac{1}{\alpha}\left(\frac{L_1(\rho)M(\rho)}{\alpha} + L_2(\rho)\right)\|q - \tilde{q}\|\,.\end{aligned}$$

Taking the supremum for $\mu \in [0, 1]$, $q \neq \tilde{q}$ belonging to $B_d(0, \rho)$ and $\rho > 0$ yields the claim. □

We now have all the tools to prove Theorem D.4.

Proof of Theorem D.4. Note that in that setting, using Lemma D.5, for all $\mu \in [0,1]$, $q \mapsto \Psi(\mu, q) + \frac{\alpha}{2}\left\|q\right\|^2$ is a strongly convex function and has a unique minimum, which is also the unique critical point of the gradient $q \mapsto \nabla\Psi(\mu, q) + \alpha q$. Let $q_0 \in \mathbb{R}^d$ be such that

$$\{q_0\} = \arg\min \Psi(0, q) + \frac{\alpha}{2}\left\|q\right\|^2 .$$

Thanks to Lemma D.6, Φ satisfies the hypothesis of Theorem G.1, with

$$a = \frac{1}{\alpha}\sup_{\rho>0} M(\rho), \quad b = \frac{1}{\alpha}\left(\frac{\sup_{\rho>0} L_1(\rho)\sup_{\rho>0} M(\rho)}{\alpha} + \sup_{\rho>0} L_2(\rho)\right), \quad \text{and} \quad c = 0. \tag{32}$$

Let $\mu \in [0,1] \to q(\mu)$ be the unique solution up to time 1 to the ODE

$$q'(\mu) = -\left(\nabla^2\Psi(\mu, q(\mu)) + \alpha\,\mathrm{I}\right)^{-1}\partial_\mu\nabla\Psi(\mu, q(\mu)) = \Phi(\mu, q(\mu)), \quad q(0) = q_0 .$$

According to Theorem G.1 applied to $\Lambda = \Phi$, it exists and is well-defined up until $\mu = 1$.

Remark that when differentiating in $\mu \in [0,1]$ the function $\mu \mapsto \nabla\Psi(\mu, q(\mu)) + \alpha q(\mu)$, we have

$$\left(\nabla^2\Psi(\mu, q(\mu)) + \alpha\,\mathrm{I}\right) q'(\mu) + \partial_\mu\nabla\Psi(\mu, q(\mu)) = \left(\nabla^2\Psi(\mu, q(\mu)) + \alpha\,\mathrm{I}\right)\left(q'(\mu) - \Phi(\mu, q(\mu))\right) = 0 .$$

Hence

$$\nabla\Psi(\mu, q(\mu)) + \alpha q(\mu) = \nabla\Psi(0, q(0)) + \alpha q(0) = 0 .$$

Thus, for any $\mu \in [0,1]$,

$$\{q(\mu)\} = \arg\min\left\{\Psi(\mu, q) + \frac{\alpha}{2}\left\|q\right\|^2\right\} ,$$

which is the promised result. □

Remark D.7 (Crude bounds under mild assumptions). Using the same standard result (condition $c = 0$ in Theorem G.1) could naturally give us some crude bounds on $\|q(\mu) - q(0)\|$, relying only on assumptions A.D.1 and A.D.3. More precisely, these bounds would strongly depend on α and improve as $\alpha \to \infty$. Namely, using Eq. (32) and Theorem D.4 one have for all $\mu \in [0,1]$,

$$\|q(\mu) - q(0)\| \le \frac{\mu}{\alpha}\cdot\sup_{\rho} M(\rho)\cdot\exp\left(\frac{1}{\alpha}\left(\frac{1}{\alpha}\big(\sup_{\rho>0} L_1(\rho)\big)\big(\sup_{\rho} M(\rho)\big) + \sup_{\rho>0} L_1(\rho)\big)\right)\mu\right) .$$

This is not the regime we aim at, since α is a small, fixed regularization constant whose role is simply to ensure that the minimization problem is well-posed.

E. Quantitative bounds on the trajectory

Let us recall that q is the minimizer of the interpolated problem (21). In the previous section, we have made two assumptions (A.D.1 and A.D.3), guaranteeing that q is well-defined and is the unique solution to the ODE (22). In this section, we show how to obtain quantitative bounds on $\|q(0) - q(\mu)\|$ by studying the ODE (22). To derive these bounds, we now make two additional assumptions on Ψ. The first one is an algebraic assumption which greatly improves the computations.

Assumption E.1 (Common kernel). We assume that there exists a fixed subspace $E \subset \mathbb{R}^d$ such that $\dim E = N_0$ and for all $(\mu, q) \in [0,1] \times \mathbb{R}^d$

$$\mathrm{Ker}\left(\nabla^2\Psi(\mu, q)\right) = E^{\perp}, \quad \mathrm{Im}(\nabla^2\Psi(\mu, q)) = E, \quad \text{and} \quad \partial_\mu\nabla\Psi(\mu, q) \in E .$$

The second one is a refined local-Lipschitz assumption (a quantitative version of A.D.3), which will allow us to use the case $c \neq 0$ in the Gronwall-Bahouri-Bellman type result Theorem F.7.

Assumption E.2 (quantitative (local)-Lipschitz continuity). Recall L_1 and L_2 from Definition D.2, and M from Eq. (24). For any μ, q, define $\lambda_1(\mu, q)$ the smallest positive eigenvalue of $\nabla^2\Psi(\mu, q)$. For any $\rho > 0$, define

$$w_{-1}(\rho) := \inf_{\substack{\mu\in[0,1] \\ q\in B_d(0,\rho)}} \lambda_1(\mu, q) .$$

We assume that there exist positive constants $(\Gamma_i)_{i\in -1,\ldots,2}$ and non negative constants $(\gamma_i)_{i\in -1,\ldots,2}$, such that for all $\rho > 0$,

$$L_1(\rho) \leq \Gamma_1 \,\mathrm{e}^{\gamma_1 \rho}\,, \quad L_1(\rho) \leq \Gamma_2 \,\mathrm{e}^{\gamma_2 \rho}\,, \quad M(\rho) \leq \Gamma_0 \,\mathrm{e}^{\gamma_0 \rho}\,,$$

and

$$w_{-1}(\rho) \geq \frac{1}{\Gamma{-1}} \,\mathrm{e}^{-\gamma_{-1}\rho}\ .$$

Under these stronger assumptions, we can obtain the following:

Theorem E.3 (Quantitative bounds on the trajectory). *Assume that Ψ satisfies A.D.1, A.E.1, and A.E.2. Suppose furthermore that*

$$4\Gamma_{-1}(\Gamma_0\Gamma_{-1}\Gamma_1 + \Gamma_2) < \exp\Big(-2\big((\gamma_{-1}+\gamma_0+\gamma_1)\vee\gamma_2\big)\Big(\|q_0\| + \Gamma_{-1}\Gamma_0 \,\mathrm{e}^{(\gamma_{-1}+\gamma_0+\gamma_1)\vee\gamma_2\|q_0\|}\Big)\Big)\ . \tag{33}$$

Then $\mu \mapsto q(\mu)$ is differentiable on $[0,1]$*, it is the unique solution of Eq.* (22) *and furthermore*

$$\forall \mu \in [0,1], \qquad \|q(\mu) - q_0\| \leq 2\mu\Gamma_{-1}\Gamma_0 \,\mathrm{e}^{(\sum_{i=-1}^{2}\gamma_i)\|q_0\|}\ .$$

The proof of Theorem E.3 follows the same path as the proof of Theorem D.4, with analogues of Lemmas D.5 and D.6. The crucial differences come from the fundamental use of A.E.1, which somehow allows us to diagonalize the Hessian $\nabla^2\Psi$ for all μ, q, and thus allows is to use estimates on the smallest positive eigenvalue of the Hessian. In practical cases, this assumption will not allow us to use global-Lipchitz estimates. We therefore introduce A.E.2 to deal with that. These two ingredients allow us to use the case $c > 0$ in the Grönwall-Bahouri-Bellman type lemma (Theorem G.1).

The following Lemma gives an improve bounds for the norm of the inverse of the Hessian, using the algebraic requirement on the Hessian. Its proof is similar to the proof of Lemma D.5, and we only point out how to modify it.

Lemma E.4 (Quantitative norm of inverse Hessian). *Let $\Psi : [0,1]\times\mathbb{R}^d \to \mathbb{R}$. Assume that A.D.1 and A.E.1 hold. Then*

$$\Big\|(\nabla^2\Psi(\mu,q) + \alpha\,\mathrm{I})^{-1}|_{\mathrm{Im}(\nabla^2\Psi(\mu,q))}\Big\|_{\mathrm{op}} \leq \frac{1}{\lambda_1(\mu,q)}\,, \tag{34}$$

where $f|_E$ denotes the restriction of f to the set E.

Proof. Remind that from the proof of Lemma D.5, for all $(q,\mu)\in\mathbb{R}^d\times[0,1]$, we have

$$P(\mu,q)\left(\nabla^2\Psi(\mu,q) + \alpha\,\mathrm{I}\right)^{-1} P(\mu,q)^\top = \mathrm{diag}\left(\frac{1}{\alpha},\ldots,\frac{1}{\alpha},\frac{1}{\lambda_1(\mu,q)+\alpha},\ldots,\frac{1}{\lambda_{N_0}(\mu,q)+\alpha}\right)\ .$$

Assuming that A.E.1 holds, we have for all $(\mu,q)\in[0,1]\times\mathbb{R}^d$, $N_0(\mu,q) = N_0$. Restricting to E, we see readily that the largest eigenvalue becomes $1/(\alpha + \lambda_1(\mu,q))$. □

Here again, by using the algebraic requirements on Ψ and the local-Lipshcitz bound we are able to derive a local-Lipschitz continuity result for Φ. Here again, the proof is quite similar to the one of Lemma E.5.

Lemma E.5 (Local-Lispchitz continuity of Φ). *Let Ψ such that A.D.1, and A.E.1 hold. Then Φ is locally-Lipschitz continuous in q uniformly in $\mu \in [0,1]$. More precisely, for all $q,\tilde{q}\in\mathbb{R}^d$ and all $\mu\in[0,1]$;*

$$\|\Phi(\mu,q) - \Phi(\mu,\tilde{q})\| \leq \frac{1}{w_{-1}(\|\tilde{q}\|)}\left(\frac{L_1(\|q\|\vee\|\tilde{q}\|)M(\|q\|)}{w_{-1}(\|q\|)} + L_2(\|q\|\vee\|\tilde{q}\|)\right)\|q-\tilde{q}\|\ .$$

If additionally A.E.2 holds, we get

$$\|\Phi(\mu,q) - \Phi(\mu,\tilde{q})\| \leq 2\Gamma_{-1}(\Gamma_0\Gamma_{-1}\Gamma_1 + \Gamma_2)\,\mathrm{e}^{\big((\gamma_{-1}+\gamma_0+\gamma_1)\vee\gamma_2\big)\|q\|\vee\|\tilde{q}\|}\,\|q-\tilde{q}\|\,, \tag{35}$$

and

$$\|\Phi(\mu,q)\| \leq \Gamma_{-1}\Gamma_0\,\mathrm{e}^{(\gamma_{-1}+\gamma_0)\|q\|}\ .$$

Proof. Since Ψ satisfies A.E.1, for all $(\mu, q) \in [0,1] \times \mathbb{R}^d$ and all $\tilde{q} \in \mathbb{R}^d$ $\partial_\mu \nabla \Psi(\mu, q) \in \operatorname{Im}(\nabla^2 \Psi(\mu, \tilde{q}))$, and we can use we can use the second part of Lemma D.5, namely Inequality (34). Indeed, Eq. (30), in norm, is upper bounded by

$$\frac{1}{\lambda_1(\mu, q) + \alpha} L_1(\|q\| \vee \|\tilde{q}\|) \frac{1}{\lambda_1(\mu, \tilde{q}) + \alpha} M(\|q\|) \|q - \tilde{q}\| ,$$

while (31) is bounded by

$$\frac{1}{\lambda_1(\mu, \tilde{q}) + \alpha} L_2(\|q\| \vee \|\tilde{q}\|) \|q - \tilde{q}\| .$$

Summing these last two displays and using the definition of w_{-1} and the bounds of A.E.2 allows us to conclude. □

Proof of Theorem E.3. Remark that thanks to Lemma E.5, Φ satisfies the condition of Theorem G.1 with

$$a = \Gamma_{-1}\Gamma_0, \quad b = 2\Gamma_{-1}(\Gamma_0\Gamma_{-1}\Gamma_1 + \Gamma_2) \quad \text{and} \quad c = (\gamma_{-1} + \gamma_0 + \gamma_1) \vee \gamma_2.$$

Furthermore, Eq. (35) can be translated into

$$2b < \exp\left(-2c\left(\|q_0\| + a\, \mathrm{e}^{c\|q_0\|}\right)\right) .$$

which is exactly the condition of application of Theorem G.1. It ensure that there exists a unique solution $\mu \mapsto q(\mu)$ to Eq. (22). Following the proof of Theorem D.4 we can conclude easily. □

F. Specializing our results for `doc2vec`

In the previous sections, we have seen that, under some technical assumptions on Ψ, the mapping q is solution to an ODE, and we proved some bounds on $\|q(\mu) - q(0)\|$ (by means of Theorem E.3). In this section, we check that these assumptions are satisfied for the Ψ occurring when considering `doc2vec` embeddings. That is, $\Psi = \Psi^{\text{lin}}$, where Ψ^{lin} is defined by Eq. (12). This is embodied as Theorem F.7, which is Theorem 5.1 with explicit constants. We first prove a useful bound on the norm of π_t:

Lemma F.1 (Bound on π_t). *Define*

$$\Pi := 2\nu\sigma_{\max}(R) \cdot \sup_i \|P_{:,i}\| .$$

Then, for any document x and any position $t \in x$, it holds that

$$\|\pi_t\| \leq \Pi .$$

We emphasize that Lemma F.1 is true regardless of the model used (PVDMmean, PVDMconcat, PVDBOW), even though this bound can be strengthened for specific models. Moreover, it only depends on the P and R matrices, which are fixed matrices after training.

Proof. Recall that we defined $\pi_t = RPh_t$. For PVDBOW, $h_t = 0$ and there is nothing to prove. Otherwise, let us first write

$$\|\pi_t\| = \|RPh_t\| \leq \sigma_{\max}(R) \cdot \|Ph_t\|$$

and focus on $\|Ph_t\|$. Let us assume that we work with PVDMconcat. Since, in that case, h_t is the concatenation of 2ν arbitrary one-hot vectors, Ph_t is the sum of 2ν arbitrary columns of P. Using the triangle inequality, we deduce that $\|Ph_t\|$ is smaller than 2ν times the largest norm of a column of P. When PVDMmean is used, the reasoning is similar. Ignoring the $1/(2\nu)$ factor (which we consider to be part of P), the bound is the same. □

Since the matrix R appears in all the definition of the embeddings, one needs some (mild) assumptions on R. The first one ensures that the condition number of R is not equal to $+\infty$.

Assumption F.2 (Condition number of R). Let us $R \in \mathbb{R}^{D\times d}$. We assume that $\operatorname{Im}(R) \subset \mathbb{1}^\perp$, and further that the smallest singular value of R is non-negative, that is,

$$\sigma_{\min}(R) > 0 .$$

The requirement for the range of R is needed here in order to work in the setting of Lemma H.6 and H.8, and then use the nice bounds for the (local)-Lipschitz constant of the softmax and its Jacobian.

Lemma F.3. *Suppose that A.F.2 hold. Then Ψ^{lin} satisfies A.D.1.*

Proof. Recall that $\mathcal{S}$ denotes the set of modified words. Coming back to the definition of F and G, we see that, when forming the difference $F-G$, many cancellations happen. To be more precise, replacing a word at position t only modifies π_s for s belonging to the neighborhood of t. Thus

$$G(q) - F(q) = \sum_{t\in\mathcal{E}} \Big(\psi_{\tilde{x}_t}(\tilde{\pi}_t + Rq) - \psi_{x_t}(\pi_t + Rq)\Big)\,, \tag{36}$$

where $\mathcal{E} \subseteq \{s \in [T], |s-t| \leq \nu \text{ with } t \in \mathcal{S}\}$. In particular, there is a numerical constant $\ell > 0$ such that $|\mathcal{E}| \leq \ell\nu\,|\mathcal{S}|$. From the definition of Ψ^{lin}, Eq. (36), and Lemma H.1, we deduce that

$$\begin{aligned}\nabla\Psi^{\text{lin}}(\mu,q) =& R^\top\left(\mu\frac{1}{T}\sum_{t\in\mathcal{E}}\big(\nabla\psi_{x_t}(\pi_t+Rq) - \nabla\psi_{\tilde{x}_t}(\tilde{\pi}_t+Rq)\big) + \frac{1}{T}\sum_{t\in x}\nabla\psi_{x_t}(\pi_t+Rq)\right)R\\ =& R^\top\left(-\mu\frac{1}{T}\sum_{t\in\mathcal{E}}\big(\sigma(\pi_t+Rq)-\sigma(\tilde{\pi}_t+Rq)\big) + \mu\frac{1}{T}\sum_{t\in\mathcal{E}}\big(\mathbb{1}_{x_t}-\mathbb{1}_{\tilde{x}_t}\big) - \frac{1}{T}\sum_{t\in x}\big(\sigma(\pi_t+Rq)-\mathbb{1}_{x_t}\big)\right),\end{aligned}$$

$$\begin{aligned}\partial_\mu\nabla\Psi^{\text{lin}}(\mu,q) =& R^\top\left(\frac{1}{T}\sum_{t\in\mathcal{E}}\big(\mathbb{1}_{x_t}-\mathbb{1}_{\tilde{x}_t}\big) - \frac{1}{T}\sum_{t\in\mathcal{E}}\big(\sigma(\pi_t+Rq)-\sigma(\tilde{\pi}_t+Rq)\big)\right)\\ =& R^\top\left(\frac{1}{T}\sum_{t\in\mathcal{E}}\int_0^1\Big(\big(\mathbb{1}_{x_t}-\mathbb{1}_{\tilde{x}_t}\big) - \nabla\sigma(u(\pi_t-\tilde{\pi}_t)+Rq)(\pi_t-\tilde{\pi}_t)\Big)\,\mathrm{d}u\right)\end{aligned}$$

and

$$\nabla^2\Psi^{\text{lin}}(\mu,q) = R^\top\left(\mu\frac{1}{T}\sum_{t\in\mathcal{E}}\left(\nabla\sigma(\pi_t+Rq)-\nabla\sigma(\tilde{\pi}_t+Rq)\right) + \frac{1}{T}\sum_{t\in x}\nabla\sigma(\pi_t+Rq)\right)R, \tag{37}$$

where we remind that $\nabla\sigma = \operatorname{diag}(\sigma) - \sigma\sigma^\top$. Hence, $\nabla^2\Psi^{\text{lin}}(\mu,\cdot)$ is a symmetric non-negative matrix and Ψ^{lin} satisfies A.D.1. □

Next, we show that Ψ^{lin} satisfies A.E.1.

Lemma F.4. *Suppose that A.F.2 holds. For all $\mu \in [0,1]$ and all $q \in \mathbb{R}^d$,*

$$\operatorname{Ker}\left(\nabla^2\Psi^{lin}(\mu,q)\right) = \{0\}\,,$$

and Ψ^{lin} satisfies A.E.1 with $N_0 = d$. Let us recall that we defined λ_1 the smallest non-zero eigenvalue of the Hessian of Ψ^{lin}. Then, for all $(\mu,q) \in [0,1]\times\mathbb{R}^d$, it holds that

$$\lambda_1(\mu,q) \geq \mathrm{e}^{-2\sqrt{2}\Pi}\,\frac{1}{D}\sigma_{\min}(R)^2\,\mathrm{e}^{-2\sqrt{2}\sigma_{\max}(R)\|q\|}\,.$$

Proof. Let us remind from Lemma H.5 the definition of $\lambda_{\min}$, namely for $z \in \mathbb{R}^D$,

$$\lambda_{\min}(z) = \min\left(\operatorname{Spec}\left(\operatorname{diag}\left(\sigma(z)\right) - \sigma(z)\sigma(z)^\top\right)\setminus\{0\}\right)\,.$$

For $q, y \in \mathbb{R}^d$ and since $Ry \in \mathbb{1}^\perp$ (thanks to A.F.2), the minimax theorem allows us to write (using Eq. (37))

$$\begin{aligned}\langle\nabla^2\Psi^{\text{lin}}(\mu,q)y,y\rangle =& \mu\frac{1}{T}\sum_{t\in x}\langle\left(\nabla\sigma(\tilde{\pi}_t+Rq)\right)(Ry),(Ry)\rangle\\ &+ (1-\mu)\frac{1}{T}\sum_{t\in x}\langle\left(\nabla\sigma(\pi_t+Rq)\right)(Ry),(Ry)\rangle\\ \geq& \frac{1}{T}\sum_{t\in x}\left(\mu\lambda_{\min}(\tilde{\pi}_t+Rq)+(1-\mu)\lambda_{\min}(\pi_t+Rq)\right)\|Ry\|^2\,.\end{aligned}$$

Here we have crucialy used A.F.2 and in particular the fact that $\mathrm{Im}(R) \subset \mathbb{1}^{\perp}$ and that $\pi_t \in \mathbb{1}^{\perp}$ in order to make $\lambda_{\min}$ appears. Let us set

$$\sigma_{(1)}(z) = \min_{i \in [D]} \sigma_i(z)\,.$$

Thanks to Lemma H.5, one has

$$\langle \nabla^2 \Psi^{\mathrm{lin}}(\mu, q) y, y \rangle \geq \frac{1}{T} \sum_{t \in x} \left(\mu D \sigma_{(1)}(\tilde{\pi}_t + Rq)^2 + (1-\mu) D \sigma_{(1)}(\pi_t + Rq)^2 \right) D \left\| Ry \right\|^2 \,.$$

Furthermore, thanks to Theorem H.9,

$$\sigma_{(1)}(z) \geq \frac{1}{D} \, \mathrm{e}^{-\sqrt{2}\|q\|},$$

and we have

$$\begin{aligned}
\langle \nabla^2 \Psi^{\mathrm{lin}}(\mu, q) y, y \rangle \geq & \frac{1}{T} \sum_{t \in x} \left(\mu \exp\left(-2\sqrt{2} \left\| \tilde{\pi}_t + Rq \right\| \right) + (1-\mu) \exp\left(-2\sqrt{2} \left\| \pi_t + Rq \right\| \right) \right) \frac{1}{D} \left\| Ry \right\|^2 \\
\geq & \, \mathrm{e}^{-2\sqrt{2}\Pi} \, \mathrm{e}^{-2\sqrt{2}\|Rq\|} \, \frac{1}{D} \left\| Ry \right\|^2 \\
\geq & \, \mathrm{e}^{-2\sqrt{2}\Pi} \, \mathrm{e}^{-2\sqrt{2}\sigma_{\max}(R)\|q\|} \, \frac{1}{D} \sigma_{\min}(R)^2 \left\| y \right\|^2 ,
\end{aligned}$$

where we remind that Π is defined in Lemma F.1. This implies that $\mathrm{Ker}\left(\nabla^2 \Psi^{\mathrm{lin}}(\mu, q) \right) = \{0\}$, that Ψ^{lin} fulfills A.E.1 with $N_0 = d$, and that

$$\lambda_1(\mu, q) \geq \mathrm{e}^{-2\sqrt{2}\Pi} \, \frac{1}{D} \sigma_{\min}(R)^2 \, \mathrm{e}^{-2\sqrt{2}\sigma_{\max}(R)\|q\|} \,. \tag{38}$$

□

Next, we show that Ψ^{lin} satisfies A.E.2.

Lemma F.5 (Local Lipschitz continuity of Ψ^{lin}). *Suppose that A.F.2 holds. Then Ψ^{lin} satisfies A.E.2 with*

$$\Gamma_{-1} = D \, \mathrm{e}^{2\sqrt{2}\Pi} \, \frac{1}{\sigma_{\min}(R)^2}, \quad \Gamma_0 = 4\ell\nu\sigma_{\max}(R) \frac{|\mathcal{S}|}{T}\,, \quad \Gamma_1 = \frac{8\,\mathrm{e}^{6\sqrt{2}\Pi}}{(D-1)} \sigma_{\max}(R)^3\,, \quad \Gamma_2 = \frac{4\ell\nu\Pi\,\mathrm{e}^{4\sqrt{2}\Pi}}{D-1} \sigma_{\max}(R)^2 \frac{|\mathcal{S}|}{T}\,,$$

and

$$\gamma_{-1} = 2\sqrt{2}\sigma_{\max}(R)\,, \quad \gamma_0 = 0\,, \quad \gamma_1 = 3\sqrt{2}\sigma_{\max}(R)\,, \quad \textit{and} \quad \gamma_2 = 2\sqrt{2}\sigma_{\max}(R)\,.$$

Proof. We have for all $\mu \in [0,1]$ and all $q \in \mathbb{R}^d$,

$$\begin{aligned}
\left\| \partial_\mu \nabla \Psi^{\mathrm{lin}}(\mu, q) \right\| \leq & \left\| R^\top \left(\frac{1}{T} \sum_{t \in \mathcal{E}} \left(\mathbb{1}_{x_t} - \mathbb{1}_{\tilde{x}_t} \right) - \frac{1}{T} \sum_{t \in \mathcal{E}} \left(\sigma(\pi_t + Rq) - \sigma(\tilde{\pi}_t + Rq) \right) \right) \right\| \\
\leq & 4\sigma_{\max}(R) \frac{|\mathcal{E}|}{T} \\
\leq & 4\ell\nu\sigma_{\max}(R) \frac{|\mathcal{S}|}{T},
\end{aligned}$$

where we have used the fact that $\|\sigma\| \leq 1$ and the previous bound gives the value of Γ_0 and γ_0. Thanks to Lemma H.6 σ is locally-Lipschitz continuous and thanks to Lemma H.8, $\nabla\sigma$ is also locally-Lipschitz continuous, hence for $q, \tilde{q} \in \mathbb{R}^d$

$$\begin{aligned}
\left\| \partial_\mu \nabla \Psi^{\mathrm{lin}}(\mu, q) - \partial_\mu \nabla \Psi^{\mathrm{lin}}(\mu, \tilde{q}) \right\| \leq & \left\| R^\top \frac{1}{T} \sum_{t \in \mathcal{E}} \left(\int_0^1 \left(\sigma(u(\pi_t - \tilde{\pi}_t) + Rq) - \sigma(u(\pi_t - \tilde{\pi}_t) + R\tilde{q}) \right) (\pi_t - \tilde{\pi}_t) \, \mathrm{d}u \right) \right\| \\
\leq & 4 \frac{1}{D-1} \, \mathrm{e}^{2\sqrt{2}(2\Pi + \sigma_{\max}(R)(\|q\| \vee \|\tilde{q}\|))} \, \ell\nu\sigma_{\max}(R)^2 \frac{|\mathcal{S}|}{T} \Pi \left\| q - \tilde{q} \right\| \\
\leq & \frac{4\ell\nu\Pi\,\mathrm{e}^{4\sqrt{2}\Pi}}{D-1} \frac{|\mathcal{S}|}{T} \, \mathrm{e}^{2\sqrt{2}\sigma_{\max}(R)(\|q\| \vee \|\tilde{q}\|)} \, \sigma_{\max}(R)^2 \left\| q - \tilde{q} \right\| \,,
\end{aligned}$$

where we have used Lemma H.6 and the bound $\frac{D}{D-1} \leq 2$, which gives the value of Γ_2 and γ_2. Finally, let us remark that

$$\begin{aligned}\left\| \nabla^2 \Psi^{\text{lin}}(\mu, q) - \nabla^2 \Psi^{\text{lin}}(\mu, q) \right\|_{\text{op}} \leq & \mu \frac{1}{T} \sum_{t \in x} \left\| R^\top \left(\nabla \sigma(\tilde{\pi}_t + Rq) - \nabla \sigma(\tilde{\pi}_t + R\tilde{q}) \right) R \right\| \\ & + (1-\mu) \frac{1}{T} \sum_{t \in x} \left\| R^\top \left(\nabla \sigma(\pi_t + Rq) - \nabla \sigma(\pi_t + R\tilde{q}) \right) R \right\| \\ & \leq \frac{8\,\mathrm{e}^{6\sqrt{2}\Pi}}{(D-1)} \sigma_{\max}(R)^3\, \mathrm{e}^{3\sqrt{2}\sigma_{\max}(R)(\|q\| \vee \|\tilde{q}\|)} \left\| q - \tilde{q} \right\| ,\end{aligned}$$

which gives the value of Γ_1 and γ_1. Finally, Eq. 38 gives directly that

$$w_{-1}(\rho) = \mathrm{e}^{-2\sqrt{2}\Pi}\, \frac{1}{D} \sigma_{\min}(R)^2\, \mathrm{e}^{-2\sqrt{2}\sigma_{\max}(R)\|q\|} ,$$

which concludes the proof. □

Next, we show that $\|q_0\|$ is not too large.

Lemma F.6 (Bound on $\|q_0\|$). *Suppose that A.F.2 holds. Then*

$$\|q_0\| \leq \frac{\sqrt{2}\sigma_{\max}(R)}{\alpha} .$$

We demonstrate Lemma F.6 in practice in Section I.2. The key idea behind the proof is that the regularization term $\frac{\alpha}{2} \left\| q \right\|^2$ prevents q from escaping to infinity.

Proof. Let us recall that

$$q_0 = \operatorname*{arg\,min}_{q \in \mathbb{R}^d} \left\{ \frac{1}{T} \sum_{t \in x} \psi_{x_t}(\pi_t + Rq) + \frac{\alpha}{2} \left\| q \right\|^2 \right\} .$$

In view of Lemma F.3, q_0 is the unique solution of the following equation:

$$R^\top \left(\frac{1}{T} \sum_{t \in x} \Big(\sigma(\pi_t + Rq) - \mathbb{1}_{x_t} \Big) \right) + \alpha q = 0 . \tag{39}$$

From Eq. (39), we deduce that

$$\|q_0\| = \frac{1}{T\alpha} \left\| R^\top \left(\sum_{t \in x} \Big(\sigma(\pi_t + Rq_0) - \mathbb{1}_{x_t} \Big) \right) \right\| .$$

By definition of $\sigma_{\max}(R)$ and the triangle inequality, this is upper bounded by

$$\frac{\sigma_{\max}(R)}{T\alpha} \sum_{t \in x} \left\| \sigma(\pi_t + Rq_0) - \mathbb{1}_{x_t} \right\| . \tag{40}$$

But we notice that, for any $q \in \mathbb{R}^d$ and $t \in x$,

$$\begin{aligned}\left\| \sigma(\pi_t + Rq) - \mathbb{1}_{x_t} \right\|^2 = & \sum_{j \neq x_t} \sigma_j(\pi_t + Rq)^2 + (\sigma_{x_t}(\pi_t + Rq) - 1)^2 \\ = & \sum_j \sigma_j(\pi_t + Rq)^2 + 1 - 2\sigma_{x_t}(\pi_t + Rq) \\ \left\| \sigma(\pi_t + Rq) - \mathbb{1}_{x_t} \right\|^2 \leq & 2 ,\end{aligned}$$

where we have used the fact that $\|\sigma\| \leq 1$ and $\sigma_i \geq 0$. Hence each term in Eq. (40) is upper bounded by $\sqrt{2}$. Keeping in mind that the summation over $t \in x$ has at most T terms, we deduce the result. □

We are now ready to apply case $c > 0$ of Theorem G.1 to obtain the promised quantitative bounds.

Theorem F.7 (Quantitative bounds for `doc2vec` embeddings). *Let Ψ^{lin} defined in Eq.* (12)*, and suppose A.F.2 holds. Let us define*

$$A := 4\ell\nu D\, \mathrm{e}^{2\sqrt{2}\Pi}\, \frac{\sigma_{\max}(R)}{\sigma_{\min}(R)^2}\,,$$

$$B := 64\ell\nu D \frac{\sigma_{\max}(R)^2}{\sigma_{\min}(R)^2}\, \mathrm{e}^{10\sqrt{2}\Pi} \left(\frac{\sigma_{\max}(R)^2}{\sigma_{\min}(R)^2} + \frac{\Pi}{D-1} \right)\,,$$

and

$$C := 5\sqrt{2}\sigma_{\max}(R)\,.$$

Suppose that

$$\frac{|\mathcal{S}|}{T} \leq \frac{1}{2B}\, \mathrm{e}^{-2(AC+1)\, \mathrm{e}^{C \frac{\sqrt{2}\sigma_{\max}(R)}{\alpha}}}\,, \tag{41}$$

Then

$$\sup_{\mu \in [0,1]} \|q(\mu) - q_0\| \leq 2A\, \mathrm{e}^{C\|q_0\|}\, \frac{|\mathcal{S}|}{T}\,. \tag{42}$$

Proof. Remark that Ψ^{lin} satisfies A.(D.1) (Lemma F.3), A.(E.1) (Lemma F.4) and A.(E.2) (Lemma F.5). Therefore the assumptions of Theorem E.3 are satisfied. Let us note that

$$\gamma_0\gamma_{-1} = A\frac{|\mathcal{S}|}{T}\,,$$

$$2\Gamma_{-1}(\Gamma_{-1}\Gamma_1\Gamma_0 + \Gamma_2) \leq B\frac{|\mathcal{S}|}{T}$$

and

$$(\gamma_{-1} + \gamma_1) \vee \gamma_2 = C\,.$$

Remark that in that setting, thanks to Lemma F.6, $\|q_0\| \leq \frac{\sqrt{2}\sigma_{\max}(R)}{\alpha}$. We also have the following straightforward bounds:

$$\frac{|\mathcal{S}|}{T} \leq 1 \quad \text{and} \quad C\,\|q_0\| \leq \mathrm{e}^{C\|q_0\|} \leq \mathrm{e}^{C\frac{\sqrt{2}\sigma_{\max}(R)}{\alpha}}\,.$$

Using Eq. (41), one necessarily have

$$\begin{aligned} 2B\frac{|\mathcal{S}|}{T} &\leq \mathrm{e}^{-2(AC+1)\,\mathrm{e}^{C\frac{\sqrt{2}\sigma_{\max}(R)}{\alpha}}} \\ &\leq \mathrm{e}^{-2C\left(\|q_0\| + A\frac{|\mathcal{S}|}{T}\mathrm{e}^{C\|q_0\|}\right)}\,. \end{aligned}$$

This guarantees that Eq. (33) and one can apply Theorem E.3, which yields the desired result. □

G. Grönwall-Bahouri-Bellman type result

In this section, we collect all results related to ODEs. In our setting, as seen in Lemma D.6, and in view of A.E.2, the coefficients of Eq. 22 are not globally Lispchitz (although locally-Lipschitz). Thus, while local existence and uniqueness of solutions to Eq. (22) is a given (small μ regime), existence up to time 1 and non-explosion of the solutions is much more challenging to achieve (large μ regime). Unfortunately, this is the regime that we are interested into: the local behavior of the ODE at $\mu = 0$ does not tell us anything interesting, since what we aim at is the comparison between the starting point ($\mu = 0$) and final point ($\mu = 1$) of the dynamic. Our strategy is to use an *ad hoc* extension of the Grönwall-Bahouri-Bellman lemma to deal with our specific setting.

Our approach is inspired by proofs of Grönwall-Bellman-Bahouri type lemmas, see for example Dannan (1985); Agarwal et al. (2005); Kim (2009); Pachpatte (2004). It relies on an explicit integration of the integral inequality which will pop up in the computations. Note that, instead of generic local constants L, M, and w_{-1}, and in view of Section F, we will suppose that all those quantity are locally bounded by some exponential functions. Our derivation is very close to that of Pachpatte inequality (Pachpatte, 2004), but here we keep track of the constants. In doing, so **we gain an explicit criteria for non-explosion of the solutions** up to time $\mu = 1$. To view other applications of non-explosion on the time-one map, one could also consult (Bailleul & Catellier, 2020) and the references therein.

Theorem G.1 (Grönwall-Bahouri-Bellman type inequality). *Let $\Lambda : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ be a continuous function and $a, b, c > 0$ be numerical constants such that, for all $q, \tilde{q} \in \mathbb{R}^d$,*

$$\sup_{\mu \in [0,1]} \|\Lambda(\mu, q)\| \leq a\, \mathrm{e}^{c\|q\|}\,, \tag{43}$$

and

$$\sup_{\mu \in [0,1]} \|\Lambda(\mu, q) - \Lambda(\mu, \tilde{q})\| \leq b\, \mathrm{e}^{c(\|q\| \vee \|\tilde{q}\|)} \|q - \tilde{q}\|\ . \tag{44}$$

Let $q_0 \in \mathbb{R}^d$ such that either $c = 0$ or

$$2b < \exp\left(-2c\left(\|q_0\| + a\, \mathrm{e}^{c\|q_0\|}\right)\right). \tag{45}$$

Then, there exists a unique function $q : [0,1] \to \mathbb{R}^d$ such that $q(0) = 0$ and for all $\mu \in [0,1]$

$$q'(\mu) = \Lambda(\mu, q(\mu))\,. \tag{46}$$

Furthermore, for all $\mu \in [0,1]$,

$$\|q(\mu) - q_0\| \leq \begin{cases} 2\mu a\, \mathrm{e}^{c\|q_0\|} & \textit{if } c > 0, \\ \mu a\, \mathrm{e}^{b} & \textit{if } c = 0\,. \end{cases}$$

Proof. **Step 1: Existence of the map satisfying** (46)**.** Note that since Λ is locally Lipschitz continuous, thanks to the Cauchy-Lipschitz/Picard-Lindelöf theorem (see Arnold (1978, Chapter 2)), there exists an open interval $I^\star$ of $[0,1]$ and a unique function $q : I^\star \to \mathbb{R}^d$ such that q is the unique solution to Eq. (46). Note that an open interval of $[0,1]$ which contains 0 is necessarily of the form $[0,\tau)$ with $0 < \tau < 1$ or $[0,1]$. Remark also that for all $\mu \in I^\star$, thanks to the regularity assumption on Λ, on $I^\star$, $\mu \mapsto \Lambda(\mu, q(\mu))$ is continuous and for $\mu \in I^\star$ the following integral equation is satisfied:

$$q(\mu) = q_0 + \int_0^\mu \Lambda(\tilde{\mu}, q(\tilde{\mu}))\, \mathrm{d}\tilde{\mu}\,.$$

Step 2: . Taking the norm in the previous display and using the triangle inequality, we see that

$$\|q(\mu) - q_0\| \leq \int_0^\mu \|\Lambda(\tilde{\mu}, q_0)\|\, \mathrm{d}\tilde{\mu} + \int_0^\mu \|\Lambda(\tilde{\mu}, q(\mu)) - \Lambda(\tilde{\mu}, q_0)\|\, \mathrm{d}\tilde{\mu}\,.$$

Using our assumptions on Λ, namely Eqs. (43) and (44), we obtain

$$\|q(\mu) - q_0\| \leq \mu a\, \mathrm{e}^{c\|q_0\|} + b \int_0^\mu \mathrm{e}^{c(\|q(\tilde{\mu})\| + \|q_0\|)} \|q(\tilde{\mu}) - q_0\|\, \mathrm{d}\tilde{\mu}\,.$$

Since $\|q(\tilde{\mu})\| - \|q(q_0)\| \leq \|q(\tilde{\mu}) - q_0\|$, we deduce that

$$\|q(\mu) - q_0\| \leq \mu a\, \mathrm{e}^{c\|q_0\|} + b\, \mathrm{e}^{2c\|q_0\|} \int_0^\mu \mathrm{e}^{c\|q(\tilde{\mu}) - q_0\|} \|q(\tilde{\mu}) - q_0\|\, \mathrm{d}\tilde{\mu}\,.$$

Let us define for all $\mu \in I^\star$,

$$\mathcal{Q}(\mu) = \begin{cases} a\, \mathrm{e}^{c\|q_0\|} + b\, \mathrm{e}^{2c\|q_0\|}\, \frac{1}{\mu} \int_0^\mu \mathrm{e}^{c\|q(\tilde{\mu}) - q_0\|} \|q(\tilde{\mu}) - q_0\|\, \mathrm{d}\tilde{\mu} & \text{if } \mu > 0\,, \\ a\, \mathrm{e}^{c\|q_0\|} & \text{if } \mu = 0\,. \end{cases}$$

Note that $\frac{1}{\mu} \int_0^\mu \mathrm{e}^{c\|q(\tilde{\mu}) - q_0\|} \|q(\tilde{\mu}) - q_0\|\, \mathrm{d}\tilde{\mu} \to_{\mu \to 0} \mathrm{e}^{c\|q(0) - q_0\|} \|q(0) - q_0\| = 0$ and $\mathcal{Q}$ is continuous in $\mu = 0$. Furthermore, for $\mu \in I^\star \backslash \{0\}$,

$$\mathcal{Q}'(\mu) = -\frac{1}{\mu^2} b\, \mathrm{e}^{2c\|q_0\|} \int_0^\mu \mathrm{e}^{c\|q(\tilde{\mu}) - q_0\|} \|q(\tilde{\mu}) - q_0\|\, \mathrm{d}\tilde{\mu} + \frac{1}{\mu} b\, \mathrm{e}^{c\|q(\mu) - q_0\|} \|q(\mu) - q_0\|\ . \tag{47}$$

With this notation in hand, for any $\mu \in I^\star \backslash \{0\}$, $\|q(\mu) - q_0\| \leq \mu \mathcal{Q}(\mu)$. Since we restrict our attention to $\mu \leq 1$, we have

$$\|q(\mu) - q_0\| \leq \mathcal{Q}(\mu)\,. \tag{48}$$

Step 3: Differential inequality. Since $x \mapsto x\,\mathrm{e}^{cx}$ is a non-decreasing function, we have for $\mu \in I^\star \backslash \{0\}$

$$\begin{aligned} \mathcal{Q}'(\mu) &\leq b\,\mathrm{e}^{2c\|q_0\|}\,\frac{\|q(\mu) - q_0\|}{\mu}\,\mathrm{e}^{c\mu\frac{\|q(\mu)-q_0\|}{\mu}} \\ &\leq b\,\mathrm{e}^{2c\|q_0\|}\,\mathcal{Q}(\mu)\,\mathrm{e}^{c\mathcal{Q}(\mu)}\,, \end{aligned} \tag{49}$$

where we used Eq. (47) for a direct bound one the derivative and Eq. (48) to obtain the last display.

Step 4: Cauchy-Lipschitz setting ($c = 0$). Suppose for a moment that $c = 0$ so that we are in the standard setting of global Cauchy-Lipschitz/Picard Lindelöf Theorem and standard Grönwall Lemma. We have for $\mu \in I^\star$

$$\log\left(\frac{\mathcal{Q}(\mu)}{a}\right) \leq b\mu\,,$$

$I^\star = [0, 1]$ and

$$\|q(\mu) - q_0\| \leq \mu\mathcal{Q}(\mu) \leq a\mu\,\mathrm{e}^{b\mu}\ .$$

Step 5: Grönwall-Bahouri-Bellman integration ($c > 0$). Suppose now that $c > 0$. Let $\beta > 0$. Let us remark that for all $x \geq 0$, $x\,\mathrm{e}^{cx} \leq \frac{1}{c}\,\mathrm{e}^{2cx}$, and we have

$$\mathcal{Q}'(\mu) \leq \frac{b}{c}\,\mathrm{e}^{2c\|q_0\|}\,\mathrm{e}^{2c\mathcal{Q}(\mu)}\ . \tag{50}$$

Multiplying both sides of Eq. (50) by $\mathrm{e}^{-2c\mathcal{Q}(\mu)}$, one recognize (up to constants) the derivative of $\mathrm{e}^{-2c\mathcal{Q}}$. Integrating from 0 to μ, we have proved that

$$\frac{\mathrm{e}^{-2ca\,\mathrm{e}^{c\|q_0\|}} - \mathrm{e}^{-2c\mathcal{Q}(\mu)}}{2} \leq b\,\mathrm{e}^{2c\|q_0\|}\,\mu\,. \tag{51}$$

When

$$\mathrm{e}^{-2ca\,\mathrm{e}^{c\|q_0\|}} - 2b\,\mathrm{e}^{2c\|q_0\|}\,\mu > 0\,, \tag{52}$$

we have

$$\mathrm{e}^{c\mathcal{Q}(\mu)} \leq \left(\mathrm{e}^{-2ca\,\mathrm{e}^{c\|q_0\|}} - 2b\,\mathrm{e}^{2c\|q_0\|}\,\mu\right)^{-\frac{1}{2}}\ . \tag{53}$$

Furthermore, whenever Eq (45) holds (namely Eq. (52) is true for all $\mu \in [0, 1]$) we can take $I^\star = [0, 1]$, since Eq. (53) guaranty that $\mathcal{Q}$ does not explode.

For $\mu \in I^\star \backslash \{0\}$ which satisfies Eq. (52), when using the previous bound and Eq. (49), we have the following inequality :

$$\mathcal{Q}'(\mu) \leq b\,\mathrm{e}^{2c\|q_0\|}\left(\mathrm{e}^{-2ca\,\mathrm{e}^{c\|q_0\|}} - 2b\,\mathrm{e}^{2c\|q_0\|}\,\mu\right)^{-\frac{1}{2}}\mathcal{Q}(\mu)\,.$$

When dividing by $\mathcal{Q}(\mu)$ and integrating, we get

$$\log(\mathcal{Q}(\mu)) - \log(\mathcal{Q}(0)) \leq \exp\left(b\,\mathrm{e}^{2c\|q_0\|}\int_0^\mu \left(\mathrm{e}^{-2ca\,\mathrm{e}^{c\|q_0\|}} - 2b\,\mathrm{e}^{2c\|q_0\|}\,\tilde{\mu}\right)^{-\frac{1}{2}}\mathrm{d}\tilde{\mu}\right)\ .$$

Therefore, for all μ which satisfies Eq. (52),

$$\begin{aligned} \|q(\mu) - q_0\| \leq \mu\mathcal{Q}(\mu) &\leq \mu a\,\mathrm{e}^{c\|q_0\|}\exp\left(\int_0^\mu b\,\mathrm{e}^{2c\|q_0\|}\left(\mathrm{e}^{-2ca\,\mathrm{e}^{c\|q_0\|}} - 2b\,\mathrm{e}^{2c\|q_0\|}\,\tilde{\mu}\right)^{-\frac{1}{2}}\mathrm{d}\tilde{\mu}\right) \\ &\leq \mu a\,\mathrm{e}^{c\|q_0\|}\exp\left(\left(\mathrm{e}^{-2ca\,\mathrm{e}^{c\|q_0\|}}\right)^{\frac{1}{2}} - \left(\mathrm{e}^{-2ca\,\mathrm{e}^{c\|q_0\|}} - 2b\,\mathrm{e}^{2c\|q_0\|}\,\mu\right)^{\frac{1}{2}}\right) \\ &\leq \mu a\,\mathrm{e}^{c\|q_0\|}\exp\left(\left(2b\,\mathrm{e}^{2c\|q_0\|}\,\mu\right)^{\frac{1}{2}}\right)\,, \end{aligned} \tag{54}$$

where we have use that for $0 \leq v < \tilde{v}$, $\sqrt{\tilde{v}} - \sqrt{v} \leq \sqrt{\tilde{v} - v}$. Note that Eq. (54) makes sense since Eq. (52) is satisfied. Finally, one can use Eq. (52) and write

$$2b\,\mathrm{e}^{2c\|q_0\|}\,\mu \leq 2b\,\mathrm{e}^{2c\|q_0\|} \leq \mathrm{e}^{-2ac\,\mathrm{e}^{2\|q_0\|}} \leq 1\,,$$

which gives

$$\|q(\mu) - q_0\| \leq \mu 2a \, \mathrm{e}^{c\|q_0\|} \,,$$

which is the wanted result. □

Remark G.2 (Improving Theorem F.7). There are several open avenues to improve Theorem F.7. One possibility is to keep a finer the dependency in μ when bounding $\|q(\mu) - q_0\|$ (namely keeping the μ factor when deriving Eq. (48)). A second possible improvement is to use a finer inequality than $y \leq \mathrm{e}^y$ when deriving Eq. (50). Unfortunately, in both cases, we were unsuccessful in integrating these more complicated expressions in a tractable form (derivation leading to Eq. (51)).

H. Technical results related to the softmax function

In this section, we collect all technical facts related to the softmax function used throughout the proofs. Let us recall that we defined the softmax function from $\mathbb{R}^D$ to $\mathbb{R}^D$ as $\sigma(x) = (\sigma_1(x), \ldots, \sigma_D(x))^\top$, where for all $i \in [D]$,

$$\sigma_i(x) = \frac{\mathrm{e}^{x_i}}{\sum_{j=1}^D \mathrm{e}^{x_j}} .$$

We also defined, for all $x \in \mathbb{R}^D$ and all $i \in [D]$,

$$\psi_i(x) = -\log(\sigma_i(x)) \,.$$

H.1. Basics on the softmax function

We start by recalling elementary properties of the softmax function.

Lemma H.1 (Softmax derivatives). *We have*

$$\frac{\partial}{\partial x_k}\sigma_\ell(x) = \begin{cases} \sigma_k(x)(1 - \sigma_k(x)) & \text{if } k = \ell \\ -\sigma_k(x)\sigma_\ell(x) & \text{otherwise.} \end{cases}$$

In a more concise way,

$$\nabla\sigma(x) = \operatorname{diag}(\sigma(x)) - \sigma(x)\sigma(x)^\top \,.$$

A straightforward consequence of Lemma H.1 is the computation of the first derivatives of ψ_i (these are very standard computations, see for instance Proposition 1 and 2 in Gao & Pavel (2017)).

Lemma H.2 (Gradient of ψ_i). *We have*

$$\frac{\partial}{\partial x_k}\psi_i(x) = \begin{cases} -1 + \sigma_k(x) & \text{if } k = i \\ \sigma_k(x) & \text{otherwise.} \end{cases}$$

In more concise notation, $\nabla\psi_i = \sigma - \mathbb{1}_i$.

Similarly, we have:

Lemma H.3 (Hessian of ψ_i). *We have*

$$\frac{\partial^2}{\partial x_k \partial x_\ell}\psi_i(x) = \begin{cases} \sigma\left(x\right)_k \left(1 - \sigma\left(x\right)_k\right) \text{ if } k = \ell \\ -\sigma\left(x\right)_k \sigma\left(x\right)_\ell \text{ otherwise.} \end{cases}$$

In more concise notation,

$$\nabla^2\psi_i = \nabla\sigma = \operatorname{diag}(\sigma) - \sigma\sigma^\top . \tag{55}$$

Corollary H.4 (Convexity of log-softmax). *For any $i \in [D]$, the function ψ_i is convex.*

The proof of the previous fact relies on the Courant minimax theorem, which gives the value of the eigenvalue of a real symmetric matrix. Furthermore, we also use that fact that a function such that its Hessian is a non-negative symmetric matrix is convex.

Proof. Let $x, v \in \mathbb{R}^D$. Since $\sum_i \sigma_i(x) = 1$, we have

$$\begin{aligned}\langle \nabla^2 \psi_i(x) v, v\rangle &= \sum_k \sigma_k(x) v_k^2 - \sum_k \sigma_k(x) \sum_\ell \sigma_\ell(x) v_\ell v_k \\ &= \sum_k \sigma_k(x) v_k^2 - \left(\sum_k \sigma_k(x) v_k\right)^2 = \sum_k \sigma_k(x) \left(v_k - \sum_\ell \sigma_\ell(x) v_\ell\right)^2 \geq 0\,,\end{aligned}$$

Hence, thanks to the Courant minimax principle, all the eigenvalues of $\nabla^2\psi_i$ are non-negative. Hence $\nabla^2\psi_i$ is a non-negative symmetric matrix, and ψ_i is a convex function. □

The following proposition controls the spectrum of the Hessian of ψ_i, that is the gradient of the softmax, in function of the minimal and maximal values of the softmax function.

Lemma H.5 (Spectrum of the softmax Jacobian). *Let $\rho > 0$. For $x \in \mathbb{R}^D$, let us define*

$$\sigma_{(1)}(x) := \min_{i \in [D]} \sigma_i(x)\,,$$

and

$$\sigma_{(D)}(x) := \max_{i \in [D]} \sigma_i(x)\,.$$

Let us define

$$\lambda_{\min}(x) := \min \left\{\operatorname{Spec}\left(\nabla\sigma(x)\right) \backslash \{0\}\right\}\,,$$

and

$$\lambda_{\max}(x) := \max \left\{\operatorname{Spec}\left(\nabla\sigma(x)\right)\right\}\,.$$

Then

$$D\sigma_{(1)}^2(x) \leq \lambda_{\min}(x) \leq \lambda_{\max}(x) \leq D\sigma_{(D)}^2(x)\,.$$

Proof. According to Lemma H.3,

$$\nabla\sigma(x) = \operatorname{diag}(\sigma(x)) - \sigma(x)\sigma(x)^\top\,.$$

This matrix is symmetric, and according to Corollary H.4, its eigenvalues are non-negative real numbers. Since $\sum_i \sigma_i(x) = 1$, one has

$$\nabla\sigma(x)\mathbb{1} = 0\,,$$

where, as before, $\mathbb{1} = (1, \ldots, 1)^\top$. Since for all $i \in [D]$, $\sigma_i(x) \neq 0$, if $v \in \operatorname{Ker}\left(\nabla\sigma(x)\right)$ then necessarily for all $i \in [D]$, $v_i\sigma_i(x) - \sigma_i(x)\sum_j \sigma_j(x) v_j = 0$, and $v = v_1\mathbb{1}$. Hence, $\operatorname{Im}(\nabla\sigma(x)) = \mathbb{1}^\perp$ and $\operatorname{Ker}\left(\nabla\sigma(x)\right) = \operatorname{Vec}\left(\mathbb{1}\right)$. Using the Courant minimax characterization of eigenvalues, we have

$$\lambda_{\min}(x) = \min_{\substack{v \in \mathbb{1}^\perp \\ \|v\|=1}} \langle \nabla\sigma(x)v, v\rangle = \min_{\substack{v \in \mathbb{1}^\perp \\ \|v\|=1}} v^\top \left(\nabla\sigma(x)\right) v \quad \text{and} \quad \lambda_{\max}(x) = \max_{\substack{v \in \mathbb{1}^\perp \\ \|v\|=1}} \langle \nabla\sigma(x)v, v\rangle = \max_{\substack{v \in \mathbb{1}^\perp \\ \|v\|=1}} v^\top \left(\nabla\sigma(x)\right) v\,. \tag{56}$$

Note then that for $v \in \mathbb{1}^\perp$ (and dropping the x dependency),

$$v^\top \left(\nabla\sigma(x)\right) v = \sum_{i=1}^D \sigma_i v_i^2 - \left(\sum_{i=1}^D \sigma_j v_j\right)^2\,.$$

Now, the Cauchy-Schwarz inequality guarantees that the previous display is non-negative, but this is not enough to conclude. We resort to the *four-letter identity* (Steele (2004, Exercise 3.7), see also Garreau & Mardaoui (2021, Proposition 13)) to write

$$v^\top \left(\nabla\sigma(x)\right) v = \sum_{i=1}^D \sigma_i v_i^2 - \left(\sum_{i=1}^D \sigma_i v_i\right)^2 = \sum_{j<k} \sigma_j\sigma_k (v_k - v_j)^2\,. \tag{57}$$

Keeping in mind that the σ_is are non-negative, this last identity gives

$$\sigma_{(1)}^2 \sum_{j<k} \sigma_j \sigma_k (v_k - v_j)^2 \leq v^\top \left(\nabla \sigma(x)\right) v \leq \sigma_{(D)}^2 \sum_{j<k} \sigma_j \sigma_k (v_k - v_j)^2 \, .$$

In the term

$$\left(\sigma_{(1)}\right)^2 \cdot \sum_{j<k} (v_k - v_j)^2 \, .$$

we recognize (D^2 times) the *variance* of the v_is. More precisely,

$$\frac{1}{D^2} \sum_{j<k} (v_k - v_j)^2 = \frac{1}{D} \sum_{i=1}^{D} \left(v_i - \frac{1}{D} \sum_j v_j \right)^2 \, .$$

Since $v \in \mathrm{Vec}\left(\mathbb{1}\right)^\perp$, we know that $\sum_j v_j = 0$, and the previous display reduces to ($1/D$ times) the norm of v. Whenever $\|v\| = 1$ and $v \in \mathbb{1}^\perp$ we have shown

$$D \sigma_{(1)}(x)^2 \leq v^\top \left(\nabla \sigma(x)\right) v \leq D \sigma_{(D)}(x)^2 .$$

Coming back to the characterization of the eigenvalues given by Eq. (56), we deduce the result. □

The previous bound, associated with estimates on the infimum and supremum of the softmax function on balls gives estimates on the (local)-Lipschitz constant of the softmax.

Lemma H.6 (local-Lipschitz continuity of the softmax). *For all $x, y \in \mathbb{R}^D$ such that $x, y \in \mathbb{1}^\perp$,*

$$\|\sigma(x) - \sigma(y)\| \leq \frac{D}{(D-1)^2} \exp\left(2 \sqrt{\frac{D}{D-1}} (\|x\| \vee \|y\|) \right) \|x - y\| \, .$$

In order to prove the previous lemma, one only has to remember that the operator norm for real symmetric matrices is the greatest eigenvalue, and use the fundamental theorem of analysis.

Proof. Let $x, y \in \mathbb{1}^\perp$. We write

$$\begin{aligned} \|\sigma(x) - \sigma(y)\| &= \left\| \int_0^1 \nabla \sigma(u(x-y) + y)(x - y) \, \mathrm{d}u \right\| \\ &\leq \int_0^1 \|\nabla \sigma(u(x-y) + y)\|_{\mathrm{op}} \|x - y\| \, \mathrm{d}u. \end{aligned}$$

One can then use Theorem H.9 and Lemma H.5, and we have for all $u \in [0, 1]$,

$$\begin{aligned} \|\nabla \sigma(u(x-y) + y)\|_{\mathrm{op}} &= \lambda_{\max}(u(x-y) + y) \\ &\leq D \sigma_{(D)}(u(x-y) + y)^2 \\ &\leq D \left(\frac{1}{1 + (D-1) \, \mathrm{e}^{-\sqrt{\frac{D}{D-1}} \|u(x-y) + y\|}} \right)^2 \\ &\leq \frac{D \, \mathrm{e}^{2\sqrt{\frac{D}{D-1}} \|u(x-y) + y\|}}{(D-1)^2} \\ &\leq \frac{D}{(D-1)^2} \, \mathrm{e}^{2\sqrt{\frac{D}{D-1}} (\|x\| \vee \|y\|)} \, . \end{aligned}$$

Putting everything together, we have

$$\|\sigma(x) - \sigma(y)\| \leq \frac{D}{(D-1)^2} \, \mathrm{e}^{2\sqrt{\frac{D}{D-1}} \|x\| \vee \|y\|} \|x - y\| \, ,$$

which is the desired result. □

Remark H.7 (Lipschitz continuity of the softmax). Note that usually, the Lipschitz continuity of the softmax is considered, but with respect to the Frobenius norm. One can obtain a crude bound starting from the squared Frobenius norm of the Jacobian, namely

$$\sum_i \sigma_i^2 (1-\sigma_i)^2 + \sum_{i\neq j} \sigma_i^2 \sigma_j^2 \,. \tag{58}$$

Since the Frobenius norm is always greater than the operator norm, this implies the result for a (global) Lispchitz constant equal to 1. A finer study of Eq. (58) yields a better Lipschitz constant for σ. This is what Alghamdi et al. (2022) do, proving $1/2$-Lipschitz continuity for the softmax function (Proposition 1 in Appendix A.4).

In view of the specific form of the gradient of the softmax, this implies that we have (almost) the same local-Lipschitz constant for the gradient of the softmax.

Corollary H.8 (local-Lispchitz continuity of the softmax Jacobian). *For all $x, y \in \mathbb{1}^\perp$,*

$$\left\|\nabla\sigma(x) - \nabla\sigma(y)\right\|_{\mathrm{op}} \leq \frac{2D^2}{(D-1)^3}\, \mathrm{e}^{3\sqrt{\frac{D}{D-1}}(\|x\|\vee\|y\|)} \left\|x-y\right\| \,.$$

The proof is a direct consequence of the particular form of the Jacobian (see Lemma H.1) and of the fact that

$$\left|\sigma_i(x) - \sigma_i(y)\right| \leq \left\|\sigma(x) - \sigma(y)\right\| \,.$$

Proof. Let $x, y \in \mathbb{1}^\perp$. We have

$$\left\|\nabla\sigma(x) - \nabla\sigma(y)\right\|_{\mathrm{op}} = \sup_{\substack{v\in\mathbb{R}^D \\ \|v\|=1}} v^\top \left(\nabla\sigma(x) - \nabla\sigma(y)\right) v \,.$$

Furthermore, using the same argument as in the proof of Lemma H.5, one can only consider $v \in \mathbb{1}^\perp$ with $\|v\| = 1$. Applying Eq. (57) to x and y and forming the difference, we obtain

$$\begin{aligned} v^\top \left(\nabla\sigma(x) - \nabla\sigma(y)\right) v =& \sum_{i<k} \Big(\sigma_i(x)\sigma_k(x) - \sigma_i(y)\sigma_k(y)\Big)(v_i - v_k)^2 \\ =& \sum_{i<k} \Big(\sigma_i(x) - \sigma_i(y)\Big)\sigma_k(x)(v_i - v_k)^2 + \sum_{i<k} \sigma_i(y)\Big(\sigma_k(x) - \sigma_k(y)\Big)(v_i - v_k)^2 \end{aligned}$$

Each of these terms can be bounded, using successively the local Lipschitz continuity of the softmax (Lemma H.6) and the definition of $\sigma_{(D)}$. The last display is upper bounded by

$$\left(\frac{D}{(D-1)^2}\, \mathrm{e}^{2\sqrt{\frac{D}{D-1}}(\|x\|\vee\|y\|)}\, \sigma_{(D)}(x) + \frac{D}{(D-1)^2}\, \mathrm{e}^{2\sqrt{\frac{D}{D-1}}(\|x\|\vee\|y\|)}\, \sigma_{(D)}(y)\right) \sum_{i<k} (v_i - v_k)^2 \left\|x-y\right\| \,,$$

which, in turn, is smaller than

$$\left(\sigma_{(D)}(x) + \sigma_{(D)}(y)\right) \frac{D}{(D-1)^2}\, \mathrm{e}^{2\sqrt{\frac{D}{D-1}}(\|x\|\vee\|y\|)} \sum_{i<k} (v_i - v_k)^2 \left\|x-y\right\| \,.$$

Using the bound on $\sigma_{(D)}$ given by Theorem H.9, we have

$$v^\top \left(\nabla\sigma(x) - \nabla\sigma(y)\right) v \leq \frac{2D}{(D-1)^3}\, \mathrm{e}^{3\sqrt{\frac{D}{D-1}}(\|x\|\vee\|y\|)} \left\|x-y\right\| \sum_{i<k} (v_i - v_k)^2.$$

Using again the same argument as in the proof of Lemma H.6, we have $\sum_{i<k}(v_i - v_k)^2 = D$, and finally for $v \in \mathbb{1}^\perp$ with $\|v\| = 1$, we have

$$v^\top (\nabla\sigma(x) - \nabla\sigma(y)) v \leq \frac{2D^2}{(D-1)^3}\, \mathrm{e}^{2\sqrt{\frac{D}{D-1}}(\|x\|\vee\|y\|)} \left\|x-y\right\| \,,$$

which gives the wanted result by taking the supremum on v. □

H.2. Minimization of the softmax function

In this section, we study the extremal values of the softmax function. The reason of this study is the close connection of these extremal values with the spectrum of the softmax and log-softmax function. Intuitively, the trivial bound $\sigma_i(x) \leq 1$ can be greatly strengthened when the norm of x is constrained: $\sigma_i(x) = 1$ is achieved only when $x_i \to +\infty$, which can not be if x lives in a ball of radius ρ.

Theorem H.9 (Bounding the softmax function). *Let $\rho > 0$ and $D \geq 2$. We have*

$$\min_{i\in[D]} \inf_{\substack{\|x\|\leq\rho \\ \sum_j x_j = 0}} \sigma_i(x) = \frac{1}{1+(D-1)\,\mathrm{e}^{\sqrt{\frac{D}{D-1}}\rho}}\,.$$

and

$$\max_{i\in[D]} \sup_{\substack{\|x\|\leq\rho \\ \sum_j x_j = 0}} \sigma_i(x) = \frac{1}{1+(D-1)\,\mathrm{e}^{-\sqrt{\frac{D}{D-1}}\rho}}\,.$$

Remark H.10 (Bounding the softmax). The softmax function is ubiquitous in machine learning, and many bounds can be found in the literature (Wei et al., 2023). Generally, these bounds are pointwise, and not applicable in our case since we need a global bound on the ball of radius ρ (with the additional constraint $\sum_j x_j = 0$ coming from our algebraic assumption).

Proof. **Step 1: the infimum is achieved and invariant by permutation.** For any $x \in \mathbb{R}^D$ such that $\|x\| \leq \rho$, $\sigma_i(x) \in (0,1)$ for all $i \in [D]$. Furthermore,

$$\nabla\sigma_i(x) = \sigma_i(x)\mathbb{1}_i - \sigma_i(x)\sigma(x)\,,$$

where we remind that $(\mathbb{1}_1, \ldots, \mathbb{1}_D)$ is the canonical basis of $\mathbb{R}^D$. Hence $\nabla\sigma_i(x) \neq 0$ and the supremum is achieved on the sphere. Note that $B^0(\rho) := \left\{x \in \mathbb{R}^D : \|x\| = \rho, \sum_j x_j = 0\right\}$ is a compact set, and the infimum is a minimum. Consider $i_0 \in [D]$ and $y \in B^0(\rho)$ a joint minimizer such that

$$\sigma_{i_0}(y) = \min_{i\in[D]} \min_{x\in B^0(\rho)} \sigma_i(x)\,. \tag{59}$$

Remark that Eq. (59) is invariant by permutation, *i.e.,* for any permutation $\tau : [D] \to [D]$, we have

$$\sigma_{\tau(i_0)}(\tau\cdot y) = \sigma_{i_0}(y) = \min_{i\in[D]} \min_{x\in B^0(\rho)} \sigma_i(x)\,,$$

where $\tau \cdot y = (y_{\tau(i)})_{i\in\{1,\ldots,D\}}$. Hence, one can suppose without loss of generality that $i_0 = 1$.

Step 2: the coordinates of a minimizer are equal under $z \mapsto z\,\mathrm{e}^{-z}$ except at i_0. In this setting, since we have for all $i \in \{2, \ldots, D\}$, $\sigma_1(y) \leq \sigma_i(y)$ this implies that $y_1 \leq y_i$. Using the fact that $\sum_j y_j = 0$, when summing the previous inequality for all $i \in [D]$, one gets $y_1 \leq 0$. Note that in fact $y_1 < 0$. Indeed, if $y_1 = 0$, we have $y_i = 0$ for all $i \in [D]$ and $\|y\| = 0 \neq \rho$.

We are in the setting of a minimization problem under constrains, namely y solves

$$\text{minimize } \sigma(x) \qquad \text{subject to} \quad \|x\|^2 = \rho^2, \quad \langle x, \mathbb{1}\rangle = 0\,.$$

Using the Lagrange-Multiplier Theorem, there exist $\alpha, \beta \in \mathbb{R}$ such that for the aforementioned solution y we have

$$\nabla\sigma(y) + \alpha\nabla\left(\|\cdot\|^2 - \rho^2\right)(y) + \beta\nabla\left(\langle\cdot, \mathbb{1}\rangle\right)(y) = 0\,,$$

which translate into

$$\begin{aligned} \sigma_1(y) - \sigma_1(y)^2 + 2\alpha y_1 + \beta &= 0 \\ -\sigma_1(y)\sigma_i(y) + 2\alpha y_i + \beta &= 0 \qquad\qquad \text{for } i \in \{2, \ldots, D\}\,. \end{aligned}$$

Remark that $\beta = 0$ and $\alpha \neq 0$. Indeed, by summing all these previous equality, and using that $\sum y_i = 0$ and $\sum \sigma_i = 1$, one gets $D\beta = 0$ and $\beta = 0$. Remind that $y_1 < 0$, and since

$$\sigma_1(y) - \sigma_1(y)^2 + 2\alpha y_1 = 0\,,$$

if $\alpha = 0$ then $\sigma_1(y)(1-\sigma_1(y)) = 1$, which is not possible. Hence $\alpha \neq 0$.

We also have that for all $i, j \in \{2, \ldots, D\}$,

$$\sigma_1(y) = \frac{2\alpha y_i}{\sigma_i(y)} = \frac{2\alpha y_j}{\sigma_j(y)} .$$

Using that fact that $\alpha \neq 0$, this implies that $\frac{y_i}{\mathrm{e}^{y_i}} = \frac{y_2}{\mathrm{e}^{y_2}}$ for all $i \in \{2, \ldots, D\}$ and that

$$0 = y_1 + \sum_{i=2}^{D} y_i = y_1 + \left(\sum_{i=2}^{D} y_2 \,\mathrm{e}^{-y_2} \mathrm{e}^{y_i}\right) = y_1 + \left(\sum_{i=1}^{D} \mathrm{e}^{y_i} - \mathrm{e}^{y_1}\right) y_2 \,\mathrm{e}^{-y_2} = y_1 + \mathrm{e}^{y_1} \frac{1-\sigma_1(y)}{\sigma_1(y)} y_2 \,\mathrm{e}^{-y_2} .$$

As a consequence, for all $i \in \{2, \ldots, D\}$,

$$y_i \,\mathrm{e}^{-y_i} = y_2 \,\mathrm{e}^{-y_2} = -y_1 \,\mathrm{e}^{-y_1} \frac{\sigma_1(y)}{1-\sigma_1(y)} . \tag{60}$$

Step 3: expression of the minimum in function of the solution of $z\,\mathrm{e}^{-z} = c$. Since $y_1 < 0$, the previous equality (60) implies that $y_i > 0$ for $i \in \{2, \ldots, D\}$. For any $0 < c < \mathrm{e}^{-1}$, the equation $x\,\mathrm{e}^{-x} = c$ has exactly two solutions, which we call $0 < y_-(c) < 1 < y_+(c)$. Let us define

$$n = \left|\left\{2 \leq i \leq D, y_i = y_-\left(-y_1 \,\mathrm{e}^{-y_1} \frac{\sigma_1(y)}{1-\sigma_1(y)}\right)\right\}\right|$$

the number of "negative" solutions. By definition of n, we necessarily have

$$\sigma_1(y) = \frac{\mathrm{e}^{y_1}}{\mathrm{e}^{y_1} + n\,\mathrm{e}^{y_-} + (D-1-n)\,\mathrm{e}^{y_+}} . \tag{61}$$

Recall that $\sum_j y_j = 0$ and $\|y\| = \rho$, hence

$$y_1 + n y_- + (D-1-n) y_+ = 0 \tag{62}$$

$$y_1^2 + n y_-^2 + (D-1-n) y_+^2 = \rho^2 . \tag{63}$$

When $n = D-1$, one can solve the previous equations and we have $y_1 = \rho\sqrt{\frac{D}{D-1}}$ and $y_j = \rho\sqrt{\frac{1}{D(D-1)}}$ for all $j \in \{2, \cdots, D\}$, and $\sigma_1(y) = \frac{1}{1+(D-1)\,\mathrm{e}^{\sqrt{\frac{D}{D-1}}\rho}}$.

Since the problem here is symmetric in y_- and y_+, one can suppose that $1 \leq n \leq D-2$. Hence rewriting Eq. (62), we obtain

$$y_+ = -\left(\frac{n}{D-1-n} y_- + \frac{1}{D-1-n} y_1\right) .$$

Replacing the value of y_+ by the right-hand side of the previous display in Eq. (63), we obtain

$$\left(n + \frac{n^2}{D-1-n}\right) y_-^2 + 2\frac{n}{D-1-n} y_1 y_- - \left(\rho^2 - y_1^2\left(1 + \frac{1}{D-1-n}\right)\right) = 0 .$$

Dividing by $\left(n + \frac{n^2}{D-1-n}\right)$, we get

$$y_-^2 + \frac{2}{D-1} y_1 y_- - \left(\frac{D-1-n}{n(D-1)}\rho^2 - \frac{D-n}{n(D-1)} y_1^2\right) = 0 .$$

We can see the previous display as a quadratic equation in y_-, which we now solve. There exists $\varepsilon \in \{-1, 1\}$ such that

$$y_- = \frac{-y_1 - \varepsilon\sqrt{\frac{D-1-n}{n}}\Delta(y_1)}{D-1},$$

where

$$\Delta(y_1) = \sqrt{(D-1)\rho^2 - Dy_1^2}\,.$$

Note that in this setting one necessarily have

$$-\sqrt{\frac{D}{D-1}}\rho \leq y_1 \leq 0\,, \tag{64}$$

since we have already seen that the minimization problem under constrains has a solution, y_- and y_+ exist and $1 \leq n \leq D-2$. When the previous condition is not satisfied, necessarily in the case $n = D-1$ or $n = 0$ holds which has already been treated.

Finally, when using the fact that $y_+ = -\frac{1}{D-1-n}(y_1 + ny_-)$,

$$y_+ = \frac{-y_1 + \varepsilon\sqrt{\frac{n}{D-1-n}}\Delta(y_1)}{D-1}\,.$$

And since $y_+ > y_-$, we have

$$\varepsilon\sqrt{\frac{n}{D-1-n}} > -\varepsilon\sqrt{\frac{D-1-n}{n}}\,,$$

and we conclude that $\varepsilon = 1$, *i.e.,*

$$y_- = \frac{-y_1 - \sqrt{\frac{D-1-n}{n}}\Delta(y_1)}{D-1} \tag{65}$$

$$y_+ = \frac{-y_1 + \sqrt{\frac{n}{D-1-n}}\Delta(y_1)}{D-1}\,. \tag{66}$$

Taking a step back, we have managed to express all coordinates as an explicit function of y_1.

Step 4: closed-form expression of the minimum. Replacing y_- and y_+ in Eq. (61) by the expression obtained in Eqs. (65) and (66), we have to minimize the function of y_1 defined

$$\begin{aligned}
g(y_1) =& \frac{\mathrm{e}^{y_1}}{\mathrm{e}^{y_1} + n\,\mathrm{e}^{\frac{-y_1 - \sqrt{\frac{D-1-n}{n}}\Delta(y_1)}{D-1}} + (D-1-n)\mathrm{e}^{\frac{-y_1 + \sqrt{\frac{n}{D-1-n}}\Delta(y_1)}{D-1}}} \\
=& \left(1 + \mathrm{e}^{-\frac{D}{D-1}y_1}\left(n\,\mathrm{e}^{-\frac{\sqrt{\frac{D-1-n}{n}}\Delta(y_1)}{D-1}} + (D-1-n)\,\mathrm{e}^{\frac{\sqrt{\frac{n}{D-1-n}}\Delta(y_1)}{D-1}}\right)\right)^{-1} \\
=& \Bigg(1 + \mathrm{e}^{-\frac{D}{D-1}y_1}\sqrt{n(D-1-n)} \\
& \qquad \times \left(\sqrt{\frac{n}{D-1-n}}\,\mathrm{e}^{-\frac{\sqrt{\frac{D-1-n}{n}}\Delta(y_1)}{D-1}} + \sqrt{\frac{D-1-n}{n}}\,\mathrm{e}^{\frac{\sqrt{\frac{n}{D-1-n}}\Delta(y_1)}{D-1}}\right)\Bigg)^{-1}
\end{aligned}$$

Note that for y_1 satisfying Eq. (64) y_1 is non-positive. It is elementary to show that $y \mapsto \Delta(y)$ is an increasing function on $\mathbb{R}_-$. Moreover, for all $a > 0$, $h : x \mapsto a\,\mathrm{e}^{-\frac{x}{a}} + \frac{1}{a}\,\mathrm{e}^{ax}$ is an increasing function on $\mathbb{R}_+$. Thus $y \mapsto h(\Delta(y)/(D-1))$ is a decreasing mapping on $\mathbb{R}_-$. Hence, by taking $a = \sqrt{\frac{D-1-n}{n}}$, we have

$$h\left(\frac{\Delta(y_1)}{D-1}\right) \leq h(0) = \sqrt{\frac{D-1-n}{n}} + \sqrt{\frac{n}{D-1-n}}\,,$$

and

$$\sqrt{(D-1-n)n}h\left(\frac{\Delta(y)}{D-1}\right) \leq \sqrt{(D-1-n)n}\left(\sqrt{\frac{D-1-n}{n}} + \sqrt{\frac{n}{D-1-n}}\right) = D-1\,.$$

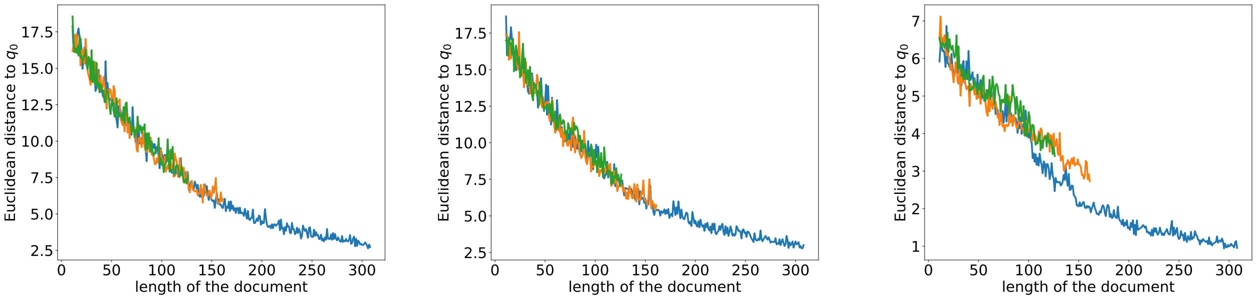

Figure 7. Influence of the length of the document with `gensim` implementation of `doc2vec`. Increasing the length of a document by considering the first words of 3 IMDB examples and replacing 5 words at random several times for each document lengTheorem Dimension of the embedding is $d = 50$, size of the dictionary is $D = 23,048$.

Using this last display, we write

$$g(y_1) \geq \frac{1}{1 + (D-1)\,\mathrm{e}^{-\frac{D}{D-1}y_1}} \,.$$

The right-hand side is an increasing function of y_1, whose minimum value is $-\sqrt{\frac{D-1}{D}}\rho$, and this gives

$$g(y_1) = \frac{1}{1 + (D-1)\,\mathrm{e}^{\sqrt{\frac{D}{D-1}}\rho}} \,. \tag{67}$$

Thus equality in the key bound is reached for

$$y = \left(-\sqrt{\frac{D-1}{D}}\rho, \sqrt{\frac{1}{D(D-1)}}\rho, \ldots, \sqrt{\frac{1}{D(D-1)}}\rho\right)^\top ,$$

with value given by Eq. (67).

Step 5: Proof for the maximum. Following the same reasoning as in the proof of Theorem H.9, we show that the maximum is reached for the point

$$\left(\sqrt{\frac{D-1}{D}}\rho, -\sqrt{\frac{1}{D(D-1)}}\rho, \ldots, -\sqrt{\frac{1}{D(D-1)}}\rho\right)^\top ,$$

and the coordinate σ_1, and we get the wanted result. □

I. Additional experimental results

In this section we collect additional experimental results.

I.1. Illustration of Theorem 5.1 with another implementation

In Figure 7 and 8, we present a replication of the experiment presented in Section 5.2 of the main paper. This time, we used the `gensim` implementation of the `doc2vec` model. The main difference is the use of *hierarchical softmax* instead of softmax. Despite this difference, the empirical results remain consistent with our theoretical claims and experimental results with an *ad hoc* implementation. We conjecture that the hierarchical softmax has similar algebraic properties to the softmax, in particular kernel stability, which would justify conducting the same analysis.

I.2. Illustration of Lemma F.6

In Figure 9, we illustrate the bound provided by Lemma F.6. We consider the 5 longest examples of the IMDB dataset and create artificial documents of increasing length as before. We observe no asymptotic dependency in T, as predicted by the theoretical result.

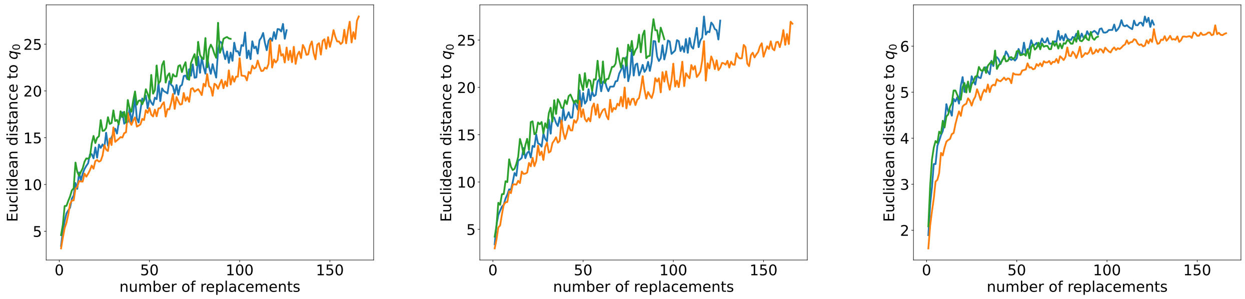

Figure 8. Influence of number of replacements with `gensim` implementation of `doc2vec`. Considering 3 examples from the IMDB dataset. Dimension of the embedding is $d = 50$, size of the dictionary is $D = 23,048$.

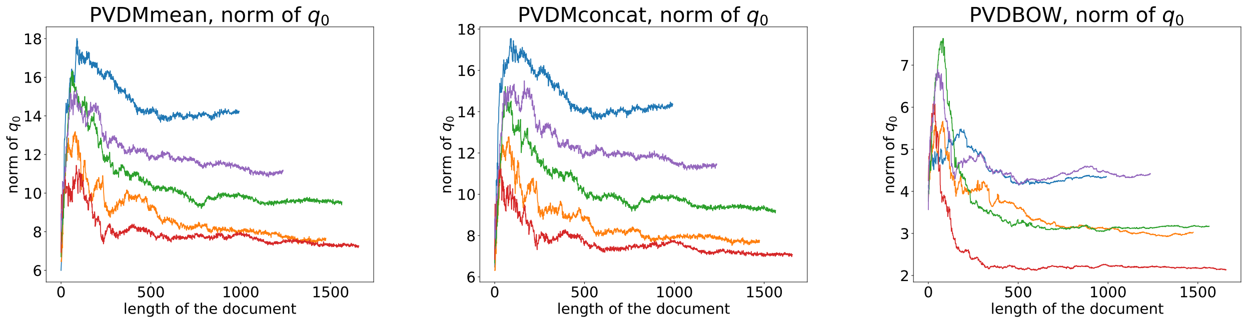

Figure 9. Norm of the original embedding as a function of T.

I.3. Singular values of R

In Figure 10, we empirically check that the singular values of the (learned) R are well-behaved. We considered the matrices from our local model and report the histogram of their singular values in log scale in Figure 10.

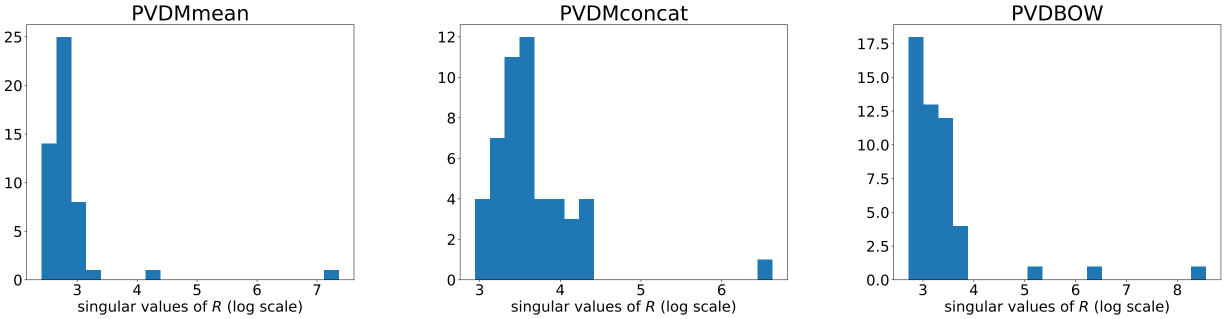

Figure 10. Singular values of R, in log scale. We observe that $\sigma_{\min}(R) > 0$.