How Smooth Is Attention?

Valérie Castin [1] **Pierre Ablin** [2] **Gabriel Peyré** [1 3]

Abstract

Self-attention and masked self-attention are at the heart of Transformers' outstanding success. Still, our mathematical understanding of attention, in particular of its Lipschitz properties – which are key when it comes to analyzing robustness and expressive power – is incomplete. We provide a detailed study of the Lipschitz constant of self-attention in several practical scenarios, discussing the impact of the sequence length n and layer normalization on the local Lipschitz constant of both unmasked and masked self-attention. In particular, we show that for inputs of length n in any compact set, the Lipschitz constant of self-attention is bounded by $\sqrt{n}$ up to a constant factor and that this bound is tight for reasonable sequence lengths. When the sequence length n is too large for the previous bound to be tight, which we refer to as the mean-field regime, we provide an upper bound and a matching lower bound which are independent of n. Our mean-field framework for masked self-attention is novel and of independent interest. Our experiments on pretrained and randomly initialized BERT and GPT-2 support our theoretical findings.

1. Introduction

Introduced by Vaswani et al. (2017), Transformers and their multi-head attention mechanism (Bahdanau et al., 2014) have significantly changed the machine learning landscape in just a few years, by becoming state-of-art models on a wide variety of tasks, from natural language processing (Brown et al., 2020; Radford et al., 2019; Wolf et al., 2019) to computer vision (Dosovitskiy et al., 2020; Zhao et al., 2020; Zhai et al., 2022; Lee et al., 2019). Despite this great empirical success, however, little is known from a theoretical perspective about the smoothness of Transformer architectures, particularly of self-attention, their main building block. We tackle this problem by focusing on the Lipschitz properties of self-attention, especially on its local Lipschitz constant, which controls how fast the output can change with respect to the input in the neighborhood of each point of the domain.

Studying the Lipschitz continuity of neural networks is particularly relevant for various questions (Rosca et al., 2020). It provides guarantees of adversarial robustness, in an attack-agnostic way (Szegedy et al., 2013; Cisse et al., 2017; Tsuzuku et al., 2018; Anil et al., 2019; Weng et al., 2018). Identifying inputs with a high local Lipschitz constant and understanding which local perturbations trigger the biggest change in the output also allows for robustifying the network, for example using adversarial training (Goodfellow et al., 2014; Miyato et al., 2015; Moosavi-Dezfooli et al., 2016; Kurakin et al., 2016). The Lipschitz constant is also involved in generalization bounds (Sokolić et al., 2017; Neyshabur et al., 2017; Bartlett et al., 2017; von Luxburg & Bousquet, 2004). From a different perspective, Lipschitz-constrained neural networks can be used to estimate Wasserstein distances (Peyré et al., 2019), enhance expressivity and improve the performance of deep models (Miyato et al., 2018; Dasoulas et al., 2021), and build invertible neural networks (Behrmann et al., 2019; Chen et al., 2019). Finally, bounding the Lipschitz constant of a neural network is an important step in the study of the associated neural ODE (Chen et al., 2018), in particular of its well-posedness (Lu et al., 2019; Geshkovski et al., 2023a;b).

Lipschitz continuity of feed-forward neural networks (FFNs) has been extensively studied and remains a hard problem. Estimating numerically the Lipschitz constant of a FFN is indeed NP-hard (Virmaux & Scaman, 2018), and theoretical bounds appear to be much larger than the actual Lipschitz constant (Virmaux & Scaman, 2018; Fazlyab et al., 2019; Latorre et al., 2020). The main theoretical difficulty here is to handle the composition of several layers more accurately than just bounding it by the product of spectral norms of weight matrices, as done by Szegedy et al. (2013). Still, taken independently, each linear map or activation function has a known tight Lipschitz constant. This is *not* the case for Transformers: the self-attention map has an involved non-linear structure, which makes the estimation of its local Lipschitz constant challenging and brings into play com-

[1]École Normale Supérieure PSL, Paris, France [2]Apple, Paris, France [3]CNRS. Correspondence to: Valérie Castin <valerie.castin@orange.fr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

pletely different approaches than for FFNs (Kim et al., 2021; Vuckovic et al., 2021).

1.1. Contributions

We make the following contributions.

- We derive a bound on the local Lipschitz constant of self-attention, which takes the form $C\sqrt{n}$ with n the sequence length of inputs and C a constant factor that depends on the parameters of self-attention and on an upper bound R on the magnitude of tokens (Theorem 3.3). We show that our bound is tight in n for reasonable (i.e. not too large) sequence lengths (Proposition 3.4). Moreover, in most Transformer architectures, the magnitude R only depends on the parameters of the network because of normalization layers (Subsection 2.3).
- We identify a *large radius regime* that is easier to analyze theoretically, with n fixed and R very large. In this regime and except for a measure-zero set of pathological configurations, we show that the local Lipschitz constant of self-attention is bounded by $C\sqrt{n}$ with C a constant that does not depend on R anymore (Theorem 3.7).
- We also study the mean-field regime, where self-attention is modeled as a map on probability measures, which corresponds to the limit $n \to +\infty$. In this framework, we show that the upper bound obtained by Geshkovski et al. (2023a), which is of the form $CR^2e^{CR^2}$, cannot be significantly improved (Proposition 3.6), by finding a R-indexed family of two-Dirac probability measures supported in the closed ball B_R of center 0 and of radius R whose local Lipschitz constant grows like $\frac{C'}{2}R^2e^{C'R^2}$ with $C' \geq C/16$.
- We are the first to study the Lipschitz constant of *masked* self-attention. We introduce a novel mean-field framework for masked self-attention, where the order of points in input measures is encoded in a supplementary coordinate, and show both in the general regime, the large radius regime and the mean-field regime that similar bounds hold for masked self-attention as for unmasked self-attention (Section 4).
- We compute numerically the local Lipschitz constant of unmasked and masked self-attention in a BERT model and a GPT-2 model, where inputs are text extracts, and observe a growth rate of $n^{1/4}$ up to a constant factor, with n the sequence length. Then, with the same networks, we build adversarial data in the input space of self-attention whose Lipschitz constant grows like $\sqrt{n}$, which evidentiates the tightness of our bounds (Section 5).

1.2. Related Work

Robustness and local Lipschitz constant estimation. Neural networks are vulnerable to adversarial attacks (Szegedy et al., 2013), and most of the methods proposed to measure and increase their robustness focus on specific attacks (Goodfellow et al., 2014; Papernot et al., 2016). It turns out, however, that such methods can be defeated by well-chosen unseen attacks (Carlini & Wagner, 2017). Measures of robustness that are agnostic to attack methods have therefore been proposed, often relying on the notion of Lipschitz constant of networks (Szegedy et al., 2013; Leino et al., 2021; Tsuzuku et al., 2018). As robustness lower bounds that rely on the (global) Lipschitz constant tend to be too loose, tighter constraints have been proposed involving the local Lipschitz constant (Hein & Andriushchenko, 2017; Weng et al., 2018). The problem of evaluating the local Lipschitz constant of deep networks is now at the heart of several recent articles (Tsuzuku et al., 2018; Leino et al., 2021), in particular for Transformers (Kim et al., 2021; Vuckovic et al., 2020; Geshkovski et al., 2023a; Catellier et al., 2023). From a more practical viewpoint, several Lipschitz-constrained variants of the Transformer architecture have been proposed, to increase robustness and reliability (Jia et al., 2023; Gupta & Verma, 2023; Ye et al., 2023; Qi et al., 2023).

Neural networks acting on measures. De Bie et al. (2019) and Pevny & Kovarik (2019) are the first to define neural networks whose inputs are probability measures, followed by several other articles (Vuckovic et al., 2020; Zweig & Bruna, 2021; Sander et al., 2022; Geshkovski et al., 2023a). Modeling neural networks as maps on probability measures has multiple applications, such as studying Wasserstein regularity (Vuckovic et al., 2020; Geshkovski et al., 2023a), proving generalization bounds (Zweig & Bruna, 2021) and doing a mean-field limit analysis of the dynamics of particles as they go through the network (Geshkovski et al., 2023a). The mean-field approach is particularly suited to the case of Encoder-only Transformers (Devlin et al., 2018), as the self-attention map is permutation equivariant, i.e., ignores the order of vectors in its input. This property can be leveraged to model any infinitely deep Encoder as a partial differential equation (PDE) on the space of measures (Sander et al., 2022), following the principle of neural ODEs (Chen et al., 2018). Analyzing this PDE then provides information about the dynamics of tokens as they go through the Transformer, showing for instance the emergence of clusters (Geshkovski et al., 2023a;b). In contrast, masked self-attention, which is crucial in Decoder-only (Liu et al., 2018) and Encoder-Decoder (Vaswani et al., 2017) architectures, is not permutation equivariant, so cannot be cast as naturally into a mean-field framework.

Regularity of self-attention and its variants. Kim et al. (2021) show that the self-attention map is not globally Lipschitz continuous by deriving a lower bound on its Lipschitz constant restricted to B_R^n. Their lower bound grows quadratically with R. To gain regularity, they define a new

self-attention map, called L2 self-attention, which is globally Lipschitz continuous on the set of inputs of length n, for all $n \geq 1$. Dasoulas et al. (2021) enforce the Lipschitz continuity of self-attention modules by normalizing the attention scores with a well-chosen normalization function. Geshkovski et al. (2023a) and Vuckovic et al. (2020) prove a mean-field upper bound on the Lipschitz constant of self-attention on B_R, by viewing self-attention as a map acting on probability measures. Their upper bound grows more than exponentially with R so that the quadratic lower bound and the exponential upper bound put together provide a very loose estimation of the Lipschitz constant of self-attention on compact sets. Finally, Sander et al. (2022) propose a modification of the attention kernel that builds on the Sinkhorn-Knopp algorithm, and provide empirical evidence of the better properties of this new choice of kernel with respect to the classical one.

1.3. Notations

The Euclidean norm on $\mathbb{R}^d$ is denoted $|\cdot|$. For any vector $w \in \mathbb{R}^n$, we denote $\mathrm{softmax}(w) := \left(\exp(w_i)/\sum_{j=1}^d \exp(w_j)\right)_{1\leq i\leq n}$ the Softmax operator, and $\mathrm{diag}(w)$ the diagonal matrix such that $\mathrm{diag}(w)_{ii} = w_i$. For any function $g\colon \mathcal{E} \to \mathcal{F}$ and any subset $\mathcal{X} \subset \mathcal{E}$, the restriction of g to $\mathcal{X}$ is denoted $g_{|\mathcal{X}}$. The closed ball centered at 0 and of radius $R > 0$ is denoted B_R. For $\varphi, \psi\colon \mathbb{R} \to \mathbb{R}$ and $a \in \mathbb{R} \cup \{+\infty\}$ we write $\varphi(x) \sim_{x\to a} \psi(x)$ if $\varphi(x)/\psi(x)$ is well-defined for x close enough to a, and $\varphi(x)/\psi(x) \to_{x\to a} 1$.

2. Standard and Mean-Field Self-Attention

2.1. Unmasked Self-Attention

Unmasked self-attention, usually just called self-attention, is central in the architecture of Transformer's Encoders (Vaswani et al., 2017), which are nowadays widely used for computer vision tasks (Dosovitskiy et al., 2020). It maps sequences of n vectors to sequences of n vectors, for any integer n.

Definition 2.1 (Single-head self-attention)**.** Let $k, d \in \mathbb{N}$. Let $Q, K, V \in \mathbb{R}^{k\times d}$ be three matrices. For any integer $n \in \mathbb{N}$ and any vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, self-attention with parameters (Q, K, V) maps the sequence $(x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ to

$$f(x_1, \ldots, x_n) := \left(V\textstyle\sum_{j=1}^n P_{ij}x_j\right)_{1\leq i\leq n} \in (\mathbb{R}^k)^n,$$

$$\text{with} \quad P_i := \mathrm{softmax}\left((x_i^\top Q^\top K x_j/\sqrt{k})_{1\leq j\leq n}\right).$$

To alleviate notations, we will denote $A := K^\top Q/\sqrt{k} \in \mathbb{R}^{d\times d}$ in what follows. In Encoders, several self-attention heads are usually combined to obtain multi-head self-attention.

Definition 2.2 (Multi-head self-attention)**.** Let $d \in \mathbb{N}$ and H a divisor of d. For $1 \leq h \leq H$, let $Q^{(h)}, K^{(h)}, V^{(h)} \in \mathbb{R}^{k\times d}$ with $k := d/H$, and $W^{(h)} \in \mathbb{R}^{d\times k}$. Multihead self-attention with parameters $(Q^{(h)}, K^{(h)}, V^{(h)}, W^{(h)})_{1\leq h\leq H}$ maps any sequence $(x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ to

$$f^{MH}(x_1, \ldots, x_n) := \sum_{h=1}^H W^{(h)} f^{(h)}(x_1, \ldots, x_n) \in (\mathbb{R}^d)^n,$$

where $f^{(h)}$ denotes single-head self-attention with parameters $(Q^{(h)}, K^{(h)}, V^{(h)})$.

When n is very large, it can be convenient to model self-attention as a map between probability measures (Sander et al., 2022; Geshkovski et al., 2023a). Indeed, the self-attention map f is permutation equivariant, which means that for all permutations σ of the set $\{1, \ldots, n\}$ and all inputs $X = (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$, it holds that $f(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) = (f(X)_{\sigma(1)}, \ldots, f(X)_{\sigma(n)})$. Informally, this means that self-attention is blind to the order of vectors so that it naturally induces a map between empirical measures, by replacing ordered sequences $X = (x_1, \ldots, x_n)$ with their associated empirical measure $m(X) := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, which does not encode the order of points anymore. To extend self-attention to more general probability measures, following Sander et al. (2022), let us introduce the notion of pushforward.

Definition 2.3 (Santambrogio (2015))**.** For a probability measure μ on $\mathbb{R}^d$ and a measurable map $\varphi\colon \mathbb{R}^d \to \mathbb{R}^d$, the pushforward of μ by φ, denoted $\varphi_\sharp\mu$, is the probability measure given by $(\varphi_\sharp\mu)(B) := \mu(\varphi^{-1}(B))$ for any Borel set $B \subset \mathbb{R}^d$, where $\varphi^{-1}(B) := \{x \in \mathbb{R}^d : \varphi(x) \in B\}$.

Intuitively, $\varphi_\sharp\mu$ is obtained by transporting each element of mass $\mu(\mathrm{d}x)$ from x to $\varphi(x)$. We are now ready to define mean-field self-attention.

Definition 2.4 (Mean-field self-attention)**.** Let $Q, K, V \in \mathbb{R}^{k\times d}$, and denote $A := K^\top Q/\sqrt{k}$. Mean-field self-attention with parameters (A, V) is defined as

$$F\colon \mu \in \mathcal{P}_c(\mathbb{R}^d) \mapsto (\Gamma_\mu)_\sharp\,\mu$$

$$\text{with} \quad \Gamma_\mu\colon x \in \mathbb{R}^d \mapsto \frac{\int \exp\left(x^\top A^\top y\right) Vy\,\mathrm{d}\mu(y)}{\int \exp\left(x^\top A^\top y\right)\mathrm{d}\mu(y)}.$$

Mean-field self-attention F generalizes discrete self-attention f in the sense that for any input $X \in (\mathbb{R}^d)^n$, we have $F(m(X)) = m(f(X))$ (see Appendix B.1).

2.2. Masked Self-Attention

In Decoder-only architectures, typically used for text generation (Liu et al., 2018; OpenAI, 2023), unmasked self-attention is replaced by masked self-attention.

Definition 2.5 (Masked self-attention)**.** Let $Q, K, V \in \mathbb{R}^{k\times d}$, and $A \coloneqq K^\top Q/\sqrt{k}$. For any integer $n \in \mathbb{N}$ and any vectors $x_1, \dots, x_n \in \mathbb{R}^d$, residual masked self-attention with parameters (A, V) maps the sequence $X = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$ to $(f^m(X)_1, \dots, f^m(X)_n) \in (\mathbb{R}^d)^n$, where

$$f^m(X)_i \coloneqq f(x_1, \dots, x_i)_i$$

with f unmasked self-attention (see Definition 2.1).

Masked self-attention processes inputs sequentially, so it is not permutation equivariant and the map f^m does not directly induce a map on empirical measures as for unmasked self-attention. To overcome this limitation and still give meaning to masked self-attention when the sequence length goes to infinity, we introduce the following novel mean-field framework. Instead of viewing inputs $(x_1, \dots, x_n) \in (\mathbb{R}^d)^n$ as empirical measures $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, we add a coordinate $s_i \in [0, 1]$ to each x_i, to encode its position in the sequence. We then define mean-field masked self-attention as a map between probability measures on the product space $[0, 1] \times \mathbb{R}^d$.

Definition 2.6 (Mean-field masked self-attention)**.** For any probability measure $\bar\mu \in \mathcal{P}_c([0, 1] \times \mathbb{R}^d)$, denote μ the second marginal of $\bar\mu$, i.e. $\mu(A) \coloneqq \int_{s=0}^1 \int_{x\in A} \mathrm{d}\bar\mu(s, x)$ for all Borel sets $A \subset \mathbb{R}^d$. We define mean-field masked self-attention on $\mathcal{P}_c([0, 1] \times \mathbb{R}^d)$ as

$$F^m \colon \bar\mu \mapsto (\Gamma_{\bar\mu})_\sharp\, \bar\mu \quad \text{where}$$

$$\Gamma_{\bar\mu}(s, x) \coloneqq \left(s, \frac{\int_{[0,1]\times\mathbb{R}^d} \exp(x^\top A^\top y) V y \mathbf{1}_{\tau\le s} \mathrm{d}\bar\mu(\tau, y)}{\int_{[0,1]\times\mathbb{R}^d} \exp(x^\top A^\top y) \mathbf{1}_{\tau\le s} \mathrm{d}\bar\mu(\tau, y)}\right).$$

This generalizes Definition 2.5 in the following sense: given any increasing sequence $0 \le s_1 < \cdots < s_n \le 1$, denoting ord the transformation $\mathrm{ord} \colon X = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n \mapsto \frac{1}{n}\sum_{i=1}^n \delta_{(s_i, x_i)} \in \mathcal{P}_c([0, 1] \times \mathbb{R}^d)$, we have $F^m(\mathrm{ord}(X)) = \mathrm{ord}(f^m(X))$ for all $X \in (\mathbb{R}^d)^n$.

Beyond the mean-field analysis of Lipschitz continuity of masked self-attention, the map F^m can be used in future work to model Decoders as partial differential equations on probability measures and study the dynamics of tokens as they go through the network, as Geshkovski et al. (2023a;b) do for Encoders.[1]

2.3. Normalization

Normalization is a key part of the Transformer architecture. The most common choice of normalization is LayerNorm (Ba et al., 2016), which has two learnable parameters $\gamma, \beta \in \mathbb{R}^d$. It acts on each input of the sequence individually with the formula $\mathrm{norm}(x) = \gamma \odot \frac{x-\mathrm{mean}(x)}{\mathrm{std}(x)} + \beta$, where $\mathrm{mean}(x) \coloneqq \frac{1}{d}\sum_{j=1}^d x_j$ and $\mathrm{std}(x) \coloneqq (\frac{1}{d}\sum_{j=1}^d (x_j - \mathrm{mean}(x))^2)^{1/2}$ are two scalars that depend on x. Each vector of the output of LayerNorm is on an ellipsis $\mathcal{S}$ of center β and of covariance $\mathrm{diag}(\gamma)^2 d$. Another popular and simpler normalization is RMSNorm (Zhang & Sennrich, 2019), which has one learnable parameter $\gamma \in \mathbb{R}^d$ and acts on each input of the sequence individually with the formula $\mathrm{norm}(x) = \gamma \odot \frac{x}{|x|}\sqrt{d}$. RMSNorm is used in recent large language (Jiang et al., 2023; Touvron et al., 2023) and vision (Dehghani et al., 2023) models. Each vector of the output of RMSNorm is on an ellipsis $\mathcal{S}$ centered at 0 and of covariance $\mathrm{diag}(\gamma)^2 d$. There are two ways to place the normalization layers in the transformer. The original transformer uses *post*-normalization: normalization is applied after each residual connection. Letting $X = (x_1, \dots, x_n)$, the output of residual attention is therefore $\mathrm{norm}(X + f(X))$. However, *pre*-normalization (Xiong et al., 2020), where normalization is applied before the attention layer: $X + f(\mathrm{norm}(X))$, is now more widespread. Although the two formulations are not equivalent, the input of self-attention f is normalized in both cases — by definition for pre-normalization, and because the previous layer was normalized for post-normalization. Hence, in practice, the input of self-attention is not any sequence in $(\mathbb{R}^d)^n$, but a sequence in B_R^n for R depending only on the parameters of norm.

It is also worth noticing that for RMSNorm, the parameter γ can be absorbed in the parameters $\theta \coloneqq (Q, K, V)$ of self-attention:

$$f_\theta \circ \mathrm{norm}(x_1, \dots, x_n) = f_{\theta'}\left(\frac{x_1}{|x_1|}, \dots, \frac{x_n}{|x_n|}\right)$$

with $\theta' \coloneqq (Q\mathrm{diag}(\lambda), K\mathrm{diag}(\lambda), V\mathrm{diag}(\lambda))$. In other words, RMSNorm followed by self-attention is equivalent to a projection on the unit sphere followed by self-attention with different parameters. This provides a simple way to bound the Lipschitz constant of the composition $f_\theta \circ \mathrm{norm}$, by directly applying our bounds on the Lipschitz constant of $f_{\theta'}$ for $R = 1$.

3. Tight Bounds on the Lipschitz Constant of Self-Attention

3.1. Lipschitz Constants and Self-Attention

Lipschitz constants provide a natural way of controlling the regularity of a function. Their definition depends on the structure that is chosen for the input and output spaces.

Euclidean framework. Let $d, n \in \mathbb{N}$ and $f \colon (\mathbb{R}^d)^n \to (\mathbb{R}^d)^n$. We equip the input and output spaces of f with the

[1] However, to do so one should rather set the first coordinate of $\Gamma_{\bar\mu}(s, x)$ to 0 instead of s so that the residual map $(\mathrm{Id} + \Gamma_{\bar\mu})_\sharp\bar\mu$ preserves the first marginal.

Frobenius norm

$$\|X\|_F \coloneqq (\textstyle\sum_{i=1}^n |x_i|^2)^{1/2}$$

for any sequence of vectors $X = (x_1, \dots, x_n)$, and assume that f is differentiable. The local Lipschitz constant of f at an input $X = (x_1, \dots, x_n)$ is defined as

$$\mathrm{Lip}_X f \coloneqq \|D_X f\|_2 ,$$

where $D_X f$ is the differential of the function f, and $\|\cdot\|_2$ denotes the operator norm induced by $\|\cdot\|_F$. We can also define, for any subset $\mathcal{X} \subset (\mathbb{R}^d)^n$, the Lipschitz constant of f on $\mathcal{X}$, as

$$\mathrm{Lip}\left(f_{|\mathcal{X}}\right) \coloneqq \sup_{\substack{X,Y \in \mathcal{X} \\ X \neq Y}} \frac{\|f(X) - f(Y)\|_F}{\|X - Y\|_F}.$$

The local Lipschitz constant tells us how fast the output of f can vary locally, while the Lipschitz constant on f controls how fast the output of f can vary on the whole set $\mathcal{X}$. We have the following connection between the two.

Lemma 3.1 (Federer (2014))**.** *Let $\mathcal{X}$ be an open and connected subset of $(\mathbb{R}^d)^n$. Then*

$$\mathrm{Lip}(f_{|\mathcal{X}}) = \sup_{X \in \mathcal{X}} \|D_X f\|_2 .$$

Mean-field framework. Let $d \in \mathbb{N}$ and denote $\mathcal{P}_c(\mathbb{R}^d)$ the set of compactly supported probability measures on $\mathbb{R}^d$. We equip this set with the 2-Wasserstein distance, defined as

$$W_2(\mu, \nu) \coloneqq \left(\inf_{\pi \in \Pi(\mu,\nu)} \int |x - y|^2 \mathrm{d}\pi(x, y) \right)^{1/2}$$

for $\mu, \nu \in \mathcal{P}_c(\mathbb{R}^d)$, where $\Pi(\mu, \nu)$ is the set of couplings between μ and ν, i.e. of probability measures $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ such that $\int \pi(\cdot, y)\mathrm{d}y = \mu$ and $\int \pi(x, \cdot)\mathrm{d}x = \nu$ (see for example Santambrogio (2015) for more details on the subject). Consider a map $F \colon \mathcal{P}_c(\mathbb{R}^d) \to \mathcal{P}_c(\mathbb{R}^d)$. For any subset $\mathcal{X} \subset \mathcal{P}_c(\mathbb{R}^d)$, the Lipschitz constant of F on $\mathcal{X}$ is defined as

$$\mathrm{Lip}(F_{\mathcal{X}}) \coloneqq \sup_{\substack{\mu,\nu \in \mathcal{X} \\ \mu \neq \nu}} \frac{W_2(F(\mu), F(\nu))}{W_2(\mu, \nu)}.$$

A notion of local Lipschitz constant can also be defined in the mean-field framework. We defer it to Appendix B.2 as it only appears in some of our proofs.

Measuring the regularity of self-attention in Wasserstein distance is the natural generalization to the mean-field case of the Euclidean regularity in the case of a finite sequence length. Indeed, when f is self-attention and F is its mean-field generalization, we can connect the two frameworks as follows (see Appendix B.3).

Lemma 3.2. *Let $R > 0$. We have*

$$\mathrm{Lip}^{\|\cdot\|_F}(f_{|B_R^n}) \leq \mathrm{Lip}^{W_2}(F_{|\mathcal{P}(B_R)}).$$

3.2. Lipschitz Bounds for Self-Attention in Different Regimes

General upper bound. Kim et al. (2021) show that self-attention is not globally Lipschitz continuous. Let us therefore restrict to sequences $(x_1, \dots, x_n) \in B_R^n$, where $B_R \subset \mathbb{R}^d$ is the closed ball centered at 0 and of radius R. We have the following general bound (see Appendix C.2).

Theorem 3.3. *Let $Q, K, V \in \mathbb{R}^{k \times d}$ and $A \coloneqq K^\top Q/\sqrt{k}$. Let $R > 0$ and $n \in \mathbb{N}$. Unmasked self-attention f with parameters (A, V) is Lipschitz continuous on B_R^n, with*

$$\mathrm{Lip}\Big(f_{|B_R^n}\Big) \leq \sqrt{3}\,\|V\|_2 \left(\|A\|_2^2\, R^4(4n+1) + n\right)^{1/2}.$$

Theorem 3.3 shows that when tokens are restricted to the compact set B_R, the Lipschitz constant of self-attention grows at most like $\sqrt{n}$ with the sequence length n, up to a constant factor. On the other side, the growth rate $\sqrt{n}$ in Theorem 3.3 is tight as long as n is not too large, a statement made rigorous by the following result (see Appendix C.4).

Proposition 3.4. *Let $Q, K \in \mathbb{R}^{k \times d}$ and $V \coloneqq I_d$. Let $A \coloneqq K^\top Q/\sqrt{k}$. Denote f unmasked self-attention with parameters (A, V). Let $\gamma_1 \geq \cdots \geq \gamma_\delta$ be the real eigenvalues of A. Then, for any $R > 0$, and denoting $\gamma \coloneqq \max(-\gamma_\delta, \gamma_1/8)$, it holds*

$$\mathrm{Lip}(f_{|B_R^n}) \geq \frac{1}{1 + (n-1)e^{-2R^2\gamma}} \sqrt{n-1}.$$

Proposition 3.4 shows that for any radius $R > 0$, the sequence of Lipschitz constants $(\mathrm{Lip} f_{|B_R^n})_{n \in \mathbb{N}}$ grows faster than $\sqrt{n}$ up to a constant factor in a certain range of sequence lengths n. For example, if $n \leq 1 + e^{2R^2\gamma}$, then

$$\mathrm{Lip}(f_{|B_R^n}) \geq \frac{\sqrt{n-1}}{2},$$

and we check that for real data and pretrained GPT-2 and BERT, the factor $R^2\gamma$ is of the order of 10^2 to 10^3 (see Appendix C.5) so that the inequality $n \leq 1 + e^{2R^2\gamma}$ is always satisfied in practice. Note that Proposition 3.4 gives a worst-case lower bound: there are inputs with a large sequence length and a small local Lipschitz constant, such as $X \coloneqq (x, \dots, x) \in (\mathbb{R}^d)^n$ for any $x \in \mathbb{R}^d$ and n, which satisfies $\|D_X f\|_2 = 1$.

Mean-field regime. What happens when the sequence length n is extremely large? As explained above, the bound provided by Theorem 3.3 becomes too loose – it even goes to $+\infty$ when $n \to +\infty$ with fixed R and r. For very large sequence lengths, this bound can be refined by leveraging the mean-field framework, as follows.

Theorem 3.5. *Let $R > 0$ and $A, V \in \mathbb{R}^{k \times d}$. The mean-field self-attention map F with parameters (A, V) is W_2-Lipschitz continuous on the set $\mathcal{P}(B_R)$, and its Lipschitz*

constant is bounded by

$$\mathrm{Lip}^{W_2}(F_{|\mathcal{P}(B_R)}) \leq \|V\|_2 (1 + 3\|A\|_2 R^2) e^{2\|A\|_2 R^2}.$$

Theorem 3.5 follows from computations made by Geshkovski et al. (2023a). We state it to draw a full picture of the regularity of self-attention on compact sets. Let us highlight the following connection between Theorem 3.5 and Theorem 3.3. For any radius $R > 0$, sequence length $n \in \mathbb{N}$ and input $X \in B_R^n$, it holds, according to Lemma 3.2 and Theorem 3.5:

$$\|D_X f\|_2 \leq \|V\|_2 (1 + 3\|A\|_2 R^2) e^{2\|A\|_2 R^2}. \quad (1)$$

On the other hand, Theorem 3.3 tells us that

$$\|D_X f\|_2 \leq \sqrt{3}\|V\|_2 \left(\|A\|_2^2 R^4 (4n+1) + n\right)^{1/2}.$$

When n is relatively small, Theorem 3.3 is more relevant than Equation (1), and vice versa for n very large. In the following Proposition, we identify an edge regime where both bounds have a similar growth in R, which corresponds to $n \sim_{R\to+\infty} e^{cR^2}$ for some constant factor $c > 0$. In this edge regime, the bound of Theorem 3.3 appears to be tight both in n and R, and the bound of Theorem 3.5 is almost tight in R up to a constant factor in the exponential.

Proposition 3.6. *Let $R > 0$. Assume that $k = d$ and $V = I_d$, and denote $\gamma_1 \geq \cdots \geq \gamma_\delta$ the real eigenvalues of A. Let $\gamma := \max(-\gamma_\delta, \gamma_1/8)$. Then, if $n \sim_{R\to+\infty} \exp(2\gamma R^2)$, there exists a function $\theta\colon [0,+\infty) \to [0,+\infty)$ such that $\theta(R) \to_{R\to+\infty} 1$ and:*

$$\mathrm{Lip}(f_{|B_R^n}) \geq \theta(R)\frac{\gamma}{2} R^2 e^{\gamma R^2}.$$

One sees that the dependency in R of the lower bound in Proposition 3.6 is catastrophic. Fortunately, in practical cases, n is significantly smaller than $e^{2\gamma R^2}$, and the mean-field regime bound is over-pessimistic: one should rather use Theorem 3.3. It is also interesting to note that configurations that lead to the explosion of the right-hand side in Proposition 3.6 are made of two extremely unbalanced clusters, one with 1 vector, and the other with the other $n-1$ vectors of the sequence (see Appendix C.6).

Large radius regime. Let us now analyze a third regime: the large radius regime, where R goes to infinity with a fixed n. This complements the mean-field analysis, where R is fixed and n goes to infinity. Let $n \in \mathbb{N}$ be a fixed sequence length. We show, drawing inspiration from Kim et al. (2021), that there exist configurations with n particles supported in B_R whose local Lipschitz constant grows like R^2 up to constant factors (see Appendix C.7). However, if we rule out a measure-zero set of pathological configurations, we get in the large radius regime that the Lipschitz constant grows at most like $\sqrt{n}$ up to a *constant* factor.

Theorem 3.7. *Let $A \in \mathbb{R}^{d\times d}$ and $V \in \mathbb{R}^{k\times d}$ two non-zero matrices. Denote $\mathcal{E}_A \subset \overline{B}(0,1)^n$ the set of sequences $(x_1, \ldots, x_n)$ such that for all $i \in \{1, \ldots, n\}$, the maximum $\max_{1\leq j\leq n} x_i^\top A^\top x_j$ is reached at a unique index j, denoted m_i. The complement of $\mathcal{E}_A$ in $\overline{B}(0,1)^n$ has zero Lebesgue measure. Moreover, for any $X \in \mathcal{E}_A$, there exists a function $\theta\colon [0,+\infty) \to [0,+\infty)$ such that $\theta(R)$ converges exponentially fast to 1 when $R \to +\infty$ and*

$$\|D_{RX} f\|_2 \leq \theta(R)\|V\|_2 \sqrt{n}.$$

The proof of this result is interesting, as it provides a better understanding of the Jacobian of self-attention in this limiting regime.

Proof. The sequences $(x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ that are not in $\mathcal{E}_A$ are such that either there exists an index $i \in \{1, \ldots, n\}$ for which $Ax_i = 0$, or the x_i are not distinct. Both cases are measure-zero situations, as $A \neq 0$. Now let $X \in \mathcal{E}_A$. For all perturbations $\epsilon := (\epsilon_1, \ldots, \epsilon_n) \in (\mathbb{R}^d)^n$ and all $i \in \{1, \ldots, n\}$, we have (see Appendix C.1)

$$(D_{RX} f)(\epsilon)_i = VR^2 \sum_{j=1}^n P_{ij}(x_j - \sum_{k=1}^n P_{ik} x_k) x_i^\top A^\top \epsilon_j$$
$$+ V \sum_{j=1}^n P_{ij}\epsilon_j + VR^2 \sum_{j=1}^n P_{ij}(x_j - \sum_{k=1}^n P_{ik} x_k) x_j^\top A \epsilon_i,$$

with $P_{ij} := e^{R^2 x_i^\top A^\top x_j} / \sum_{k=1}^n e^{R^2 x_i^\top A^\top x_k}$. By definition of m_i, we have $P_{ij} \to_{R\to+\infty} \mathbf{1}_{j=m_i}$, and the convergence is exponentially fast, so $R^2 P_{ij}$ has the same limit as P_{ij}. As a consequence, in the large radius regime, the Jacobian of self-attention has a much simpler form: $(D_{RX} f)(\epsilon)_i \to_{R\to+\infty} V \sum_{j=1}^n P_{ij}\epsilon_j$, and the operator norm of the limit is bounded by $\|V\|_2 \sqrt{n}$. Moreover, when $V = I_d$ for example, this bound is reached up to a constant factor if there exists an index j such that $j = m_i$ for a constant fraction of the indices i, i.e. if a token grasps the attention of a constant fraction of all tokens. □

In practice, the large radius regime on general configurations (i.e. that belong to the set $\mathcal{E}_A$) provides an oversimplified model for self-attention. Indeed, in this regime, attention matrices are 1-sparse row-wise, i.e. have exactly one non-zero coefficient – equal to 1, on each row $i \in \{1, \ldots, n\}$, which corresponds to the index m_i. If we look at real data, attention matrices indeed tend to have sparse rows, but with more than one non-zero coefficient on each row (Vaswani et al., 2017; Vyas et al., 2020; Likhosherstov et al., 2021) – which is expected, otherwise the representation given by self-attention would be too rough. Still, Theorem 3.7 gives some nice intuition about the growth rate of $\sqrt{n}$ obtained in Theorem 3.3 and observed in practice (see Figure 1 in the experiments).

Multi-head self-attention. Bounding the Lipschitz constant of single-head self-attention provides the following bound on the Lipschitz constant of multi-head self-attention.

Lemma 3.8 (Kim et al. (2021))**.** *Let $R > 0$. With the notations of Definition 2.2, it holds*

$$\mathrm{Lip}(f^{MH}_{|B_R^n}) \leq \sum_{h=1}^{H} \|W^{(h)}\|_2 \,\mathrm{Lip}(f^{(h)}_{|B_R^n}).$$

In the whole paper, we focus on single-head self-attention and avoid tackling the possibly tedious dependencies between the matrices $W^{(1)}V^{(1)}, \dots, W^{(H)}V^{(H)}$. Deriving a finer estimation of the Lipschitz constant of multi-head attention than what Lemma 3.8 gives us is left for future work.

4. Tight Bounds on the Lipschitz Constant of Masked Self-Attention

4.1. Measuring Distances With Conditional Optimal Transport

To study the Lipschitz properties of masked self-attention as defined in Definition 2.5, the Euclidean framework introduced in Section 3.1 still applies. In contrast, the Wasserstein framework used for mean-field unmasked self-attention is not suited to measuring the regularity of mean-field masked self-attention. Indeed, in the standard case, the distance between the outputs $f^m(X)$ and $f^m(Y)$ for two inputs $X, Y \in (\mathbb{R}^d)^n$ is measured by $(\sum_{i=1}^n |f(x_1, \dots, x_i)_i - f(y_1, \dots, y_i)_i|^2)^{1/2}$ so that i-th coordinates are compared to each other, separately for each index i. On the contrary, when looking at $W_2(F^m(\bar{\mu}), F^m(\bar{\nu}))$, the optimal transport plan may transport mass from a point (s, x) to a point (s', y) with $s \neq s'$, which contradicts the sequential nature of masked self-attention. It is therefore natural to introduce the following distance on $\mathcal{P}_c([0,1] \times \mathbb{R}^d)$, which comes from the theory of conditional optimal transport (Hosseini et al., 2023).

Definition 4.1. Let $\bar{\mu}$ and $\bar{\nu}$ be two probability measures in $\mathcal{P}_c([0,1] \times \mathbb{R}^d)$, and $p \geq 1$. If $\bar{\mu}$ and $\bar{\nu}$ have the same marginal with respect to s, i.e.

$$\int_{s_1}^{s_2} \int_{\mathbb{R}^d} \mathrm{d}\bar{\mu}(s, x) = \int_{s_1}^{s_2} \int_{\mathbb{R}^d} \mathrm{d}\bar{\nu}(s, x)$$

for all $0 \leq s_1 < s_2 \leq 1$, denote θ this marginal distribution, and write with the disintegration theorem (Bogachev & Ruas, 2007) $\mathrm{d}\bar{\mu}(\tau, x) =: \mathrm{d}\theta(\tau)\mathrm{d}\mu^\tau(x)$ and $\mathrm{d}\bar{\nu}(\tau, x) =: \mathrm{d}\theta(\tau)\mathrm{d}\nu^\tau(x)$. The measures μ^τ and ν^τ can be seen intuitively as the restriction of μ and ν to the mass located at position τ, rescaled to obtain probability measures. We then measure the distance between $\bar{\mu}$ and $\bar{\nu}$ with

$$d_p(\bar{\mu}, \bar{\nu}) := \left(\int_0^1 W_p(\mu^\tau, \nu^\tau)^p \mathrm{d}\theta(\tau) \right)^{1/p}.$$

If $\bar{\mu}$ and $\bar{\nu}$ do not have the same first marginal, we set $d_p(\bar{\mu}, \bar{\nu}) := +\infty$.

Considering $d_p(\bar{\mu}, \bar{\nu})$ amounts to minimizing the transport cost between $\bar{\mu}$ and $\bar{\nu}$ under the constraint that points must keep the first coordinate constant along their trajectory. Equivalently, allowed transport plans must be the identity on the first marginal.

As for unmasked self-attention, we have the following connection between the Euclidean framework and the mean-field framework for residual masked self-attention.

Lemma 4.2. *Let $R > 0$. We have*

$$\mathrm{Lip}^{\|\cdot\|_F}(f^m_{|B_R^n}) \leq \mathrm{Lip}^{d_2}(F^m_{|\mathcal{P}([0,1] \times B_R)}).$$

We do not detail the proof, which follows the same steps as for Lemma 3.2.

4.2. Lipschitz Bounds for Masked Self-Attention in Different Regimes

General upper bound. We show in Appendix D.1 that the bound provided by Theorem 3.3 still holds for masked self-attention.

Theorem 4.3. *Let $Q, K, V \in \mathbb{R}^{k \times d}$ and $A := K^\top Q/\sqrt{k}$. Let $R > 0$ and $n \in \mathbb{N}$. Masked self-attention f^m with parameters (A, V) is Lipschitz continuous on the set B_R^n, and*

$$\mathrm{Lip}\left(f^m_{|B_R^n}\right) \leq \sqrt{3}\,\|V\|_2 \left(\|A\|_2^2 R^4 (n+1) + n\right)^{1/2}.$$

Mean-field regime. Let us now bound from above the d_p Lipschitz constant of mean-field masked self-attention.

Theorem 4.4. *Let $R > 0$ and $p \geq 1$. The mean-field masked self-attention map F^m is Lipschitz continuous on the space of measures supported in $[0,1] \times B_R$, with a Lipschitz constant upper-bounded by*

$$\|V\|_2 \,(1 + 3\,\|A\|_2 R^2) e^{2\|A\|_2 R^2}.$$

Note that in Theorem 4.4, we consider that two measures $\bar{\mu}$ and $\bar{\nu}$ with different first marginals induce an infinite Lipschitz ratio $d_p(F^m(\bar{\mu}), F^m(\bar{\nu}))/d_p(\bar{\mu}, \bar{\nu})$. The proof can be found in Appendix D.2.

Large radius regime. We have a similar result as for unmasked attention in the large radius regime. Except for a measure-zero set of pathological configurations, when R is sufficiently large, the local Lipschitz constant of f^m at the input RX does not depend on R anymore and grows at most like $\sqrt{n}$ up to a constant factor.

Theorem 4.5. *Let $A \in \mathbb{R}^{d\times d}$ and $V \in \mathbb{R}^{k\times d}$ two non-zero matrices, and denote f^m the masked self-attention map with parameters (A, V). Denote $\mathcal{E}_A^m \subset \overline{B}(0,1)^n$ the set of sequences $(x_1, \dots, x_n)$ such that for all $i \in \{1, \dots, n\}$, the maximum $\max_{1\le j\le i} x_i^\top A^\top x_j$ is reached at a unique index j, denoted m_i. The complement of $\mathcal{E}_A^m$ in $\overline{B}(0,1)^n$ has zero Lebesgue measure. Moreover, for any $X \in \mathcal{E}_A$, there exists a function $\theta\colon [0,+\infty) \to [0,+\infty)$ such that $\theta(R)$ converges exponentially fast to 1 when $R \to +\infty$ and*

$$\|D_{RX} f^m\|_2 \le \theta(R)\, \|V\|_2 \sqrt{n}.$$

5. Experiments

The bound stated in Theorem 3.3 corresponds to a worst-case scenario. In practice, does it reflect the evolution of the Lipschitz constant of a self-attention layer of a Transformer on real data? We perform numerical experiments on BERT (Devlin et al., 2018), using the pretrained Huggingface model 'bert-base-uncased' first as an Encoder and then in its Decoder version, and on GPT-2 (Radford et al., 2019) both pretrained and randomly initialized. Both models have 12 multi-head attention layers, and 12 attention heads per layer, with an embedding dimension $d = 768$. We perform two different experiments, first with real data, and then with synthetic adversarial data.

5.1. Experiments With Real Data

We take our data from two test datasets, Alice in Wonderland from the NLTK corpus Gutenberg (Bird et al., 2009), and AG_NEWS from the PyTorch package torchtext (Zhang et al., 2015). The aim is, for various multi-head self-attention layers f^{model} of BERT and GPT-2, and for a batch of inputs of varying length taken from the two datasets mentioned above, to get a scatter plot of the local Lipschitz constant of f^{model} at each input $(x_1, \dots, x_n)$ as a function of the sequence length n.

Construction of the datasets. Given some raw text from Alice in Wonderland or AG_NEWS, we first tokenize it and then split the resulting sequence of tokens into subsequences with a fixed sequence length. For each even integer n in $\{2, \dots, 100\}$, we build 10 sequences with n tokens, so that none of the constructed sequences $(s_1, \dots, s_n)$ overlap. Then, for each input sequence $(s_1, \dots, s_n)$, we do a forward pass of the model, and get with a forward hook the intermediate activations just before the attention layer of interest f^{model}. This gives us a batch of sequences $(x_1, \dots, x_n) \in (\mathbb{R}^d)^n$ that are fed to f^{model} when $(s_1, \dots, s_n)$ goes through the model. Note that except for the inputs of the first attention layer, all the x_i are the result of normalization with LayerNorm, and therefore belong to an ellipsis, which depends on the parameters of LayerNorm.

Computation of the local Lipschitz constants. The local Lipschitz constant of f^{model} at an input sequence $X = (x_1, \dots, x_n)$ is equal to $\left\|D_X f^{\text{model}}\right\|_2$. As $D_X f^{\text{model}}$, which we denote J_X to alleviate notations, is of shape $nd \times nd$ with $d = 768$, we do not compute it explicitly but use a power method on the matrix $J_X^\top J_X$ by alternating Jacobian-vector products and vector-Jacobian products (see Appendix E.1). The power method converges to the largest eigenvalue of $J_X^\top J_X$, which is equal to $\|J_X\|_2^2$.

5.2. Experiments With Adversarial Data

Adversarial data. To check numerically the tightness in n of the bound provided by Theorem 3.3, we build adversarial data in the input space of each self-attention layer f^{model}. More precisely, for each sequence length n, and for a radius $R > 0$ to be discussed later, we look for an input $X \in B_R^n$ where f^{model} has a large (ideally maximal) local Lipschitz constant. Unfortunately, performing a gradient ascent on the local Lipschitz constant f^{model} gives poor results, as the optimization landscape is highly non-convex. We, therefore, build X as follows, using the heuristics provided by Proposition 3.6. For $h \in \{1, \dots, 12\}$, denote $A^{(h)} \coloneqq {K^{(h)}}^\top Q^{(h)}/\sqrt{k}$, with the notations of Definition 2.2 applied to the multi-head self-attention layer f^{model}. Choose any head h, and denote γ_1 (resp. γ_δ) the largest (resp. the smallest) real eigenvalue of A, and u_1 (resp. u_δ) an associated unit eigenvector. If $\gamma_1 \ge -8\gamma_\delta$, define $X = (x_1, \dots, x_n)$ with $x_1 \coloneqq Ru_1$ and $x_2 = \cdots = x_n \coloneqq R/2u_1$. If $\gamma_1 < -8\gamma_\delta$, define $X = (x_1, \dots, x_n)$ with $x_1 \coloneqq Ru_\delta$ and $x_2 = \cdots = x_n \coloneqq -Ru_\delta$. This way of defining adversarial inputs does not exactly maximize the local Lipschitz constant for each choice of n and R, but leads, for R large enough, to a growth rate of $\sqrt{n}$ for the local Lipschitz constant of f^{model} (see Figure 1), which is exactly what we need to recover tightness.

Influence of the scaling factor. In Figure 1, the scaling factor R is equal to 15.5 and 21.5 for the first two columns, which corresponds to an approximation of the mean radius of real data used with the same models in the first row. In other words, our adversarial data for the first two models have tokens with a magnitude similar to tokens obtained with real data. In contrast, for the third column, corresponding to GPT-2 randomly initialized, we take $R = 100$, which is much larger than the magnitude of tokens generated with real data (which is 27.7). Indeed, we observed that smaller scaling factors induce a growth rate that is slower than $n^{1/2}$. Studying this aspect more in-depth is an interesting perspective for future work.

5.3. Discussion

Figure 1 gives the following insights.

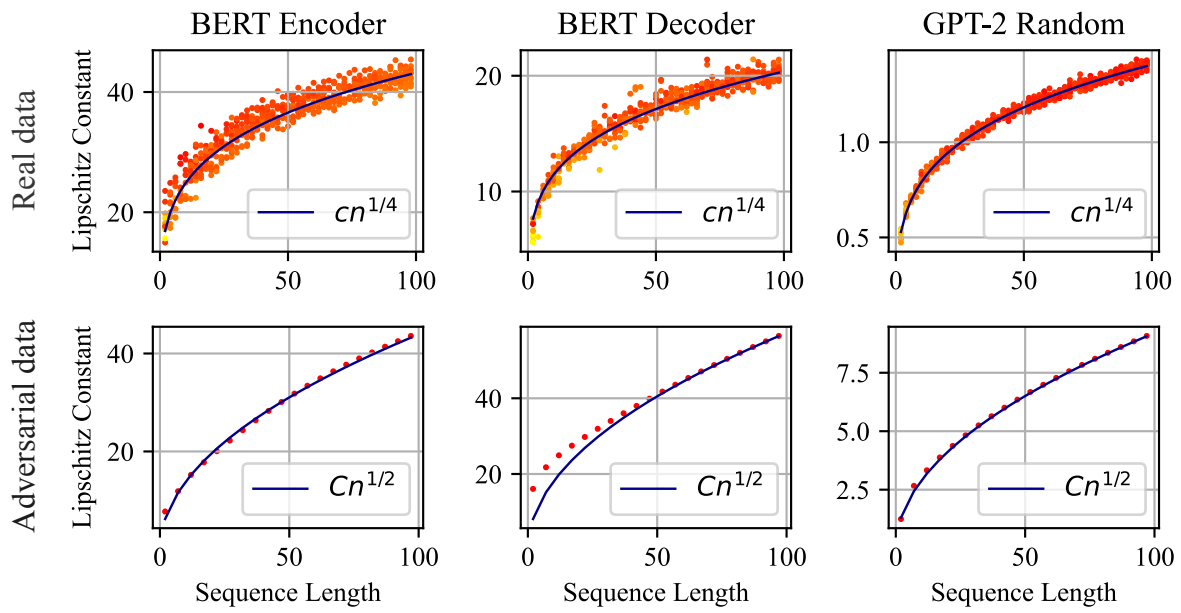

Figure 1. Scatter plots of the local Lipschitz constant of self-attention (column 1) and masked self-attention (columns 2 and 3) on text data (upper row) and adversarial data (lower row) as a function of the sequence length n. In the upper row, the color encodes the mean radius of inputs $X = (x_1, \dots, x_n)$, defined as $R := \sqrt{1/n \sum_{i=1}^n |x_i|^2}$. Lighter points have a smaller mean radius. The first two columns correspond to two different pretrained BERT models: an Encoder-only and a Decoder-only, on the same dataset Alice in Wonderland, respectively for attention layers 0 and 6. The third column is obtained with the masked self-attention layer 6 of GPT-2 randomly initialized, on the dataset AG_NEWS. We see that the Lipschitz constant of self-attention on real data grows approximately like $n^{1/4}$ with the sequence length n and that the growth rate is $\sqrt{n}$ for adversarial data, which shows the tightness of Theorems 3.3, 3.7, 4.3 and 4.5.

- The Lipschitz constant of self-attention on real data grows significantly with the sequence length, in all considered cases, independently of the dataset, the depth of the attention layer, and of whether self-attention is masked or not. The observed growth rate is close to $n^{1/4}$, which is smaller than the worst-case rate $\sqrt{n}$.
- The Lipschitz constant of self-attention on our adversarial data grows like $\sqrt{n}$, which is the worst-case rate according to Theorem 3.3. This evidentiates tightness of the bound with respect to the sequence length.

Let us make a few remarks. First, the architecture of BERT adds biases to the traditional formula for self-attention. This does not affect too much the theoretical analysis (see Appendix E.3). Second, the same experiments as in Figure 1 performed with GPT-2 pre-trained (Radford et al., 2019) lead to a different behavior of the Lipschitz constant. In particular, the growth rate of the Lipschitz constant can be faster than $\sqrt{n}$, which seems to come from a correlation between the sequence length of inputs and the magnitude of their tokens after going through LayerNorm (see Appendix E.2). Finally, our results point out the difficulty of designing Lispchitz-constrained self-attention layers independently of the sequence length. Indeed, dividing a self-attention layer by the mean-field bound of Theorem 3.5 to enforce its 1-Lipschitz continuity would induce a dramatic loss of expressive power on smaller sequence lengths. However, when the sequence length is fixed – for example with Vision Transformers (Dosovitskiy et al., 2020), dividing the output of the self-attention layer by the bound in Theorem 3.3 is a promising option.

Conclusion

In this thorough study of the Lipschitz constant of self-attention, we have identified sharp bounds in different regimes, the most relevant from a practical viewpoint being the general bounds stated in Theorems 3.3 and 4.3. Our theoretical and numerical analyses show that the Lipschitz constant of self-attention grows with the sequence length, the worst-case rate being $\sqrt{n}$, and the rate on real data being at least $n^{1/4}$, and possibly larger for learned positional encoding. This insight is new and represents an obstruction to designing robust Transformers without modifying the architecture or fixing the sequence length of inputs, which opens interesting avenues for future work. We have also introduced a novel mean-field framework for masked self-attention, which overcomes the lack of permutation equivariance and paves the way for a study of neural PDEs on Decoders, as Sander et al. (2022) and Geshkovski et al. (2023a;b) do for Encoders.

Impact Statement

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

References

Anil, C., Lucas, J., and Grosse, R. Sorting out lipschitz function approximation. In *International Conference on Machine Learning*, pp. 291–301. PMLR, 2019.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Behrmann, J., Grathwohl, W., Chen, R. T., Duvenaud, D., and Jacobsen, J.-H. Invertible residual networks. In *International conference on machine learning*, pp. 573–582. PMLR, 2019.

Bird, S., Klein, E., and Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Bogachev, V. I. and Ruas, M. A. S. *Measure theory*, volume 1. Springer, 2007.

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. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Catellier, R., Vaiter, S., and Garreau, D. On the robustness of text vectorizers. *arXiv preprint arXiv:2303.07203*, 2023.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Chen, R. T., Behrmann, J., Duvenaud, D. K., and Jacobsen, J.-H. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In *International conference on machine learning*, pp. 854–863. PMLR, 2017.

Dasoulas, G., Scaman, K., and Virmaux, A. Lipschitz normalization for self-attention layers with application to graph neural networks. In *International Conference on Machine Learning*, pp. 2456–2466. PMLR, 2021.

De Bie, G., Peyré, G., and Cuturi, M. Stochastic deep networks. In *International Conference on Machine Learning*, pp. 1556–1565. PMLR, 2019.

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A. P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pp. 7480–7512. PMLR, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

Fazlyab, M., Robey, A., Hassani, H., Morari, M., and Pappas, G. Efficient and accurate estimation of lipschitz constants for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Federer, H. *Geometric Measure Theory*. Classics in Mathematics. Springer Berlin Heidelberg, 2014. ISBN 9783642620102. URL https://books.google.fr/books?id=jld-BgAAQBAJ.

Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. The emergence of clusters in self-attention dynamics. *arXiv preprint arXiv:2305.05465*, 2023a.

Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. A mathematical perspective on transformers. 2023b.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gupta, K. and Verma, S. Certvit: Certified robustness of pre-trained vision transformers. *arXiv preprint arXiv:2302.10287*, 2023.

Hein, M. and Andriushchenko, M. Formal guarantees on the robustness of a classifier against adversarial manipulation. *Advances in neural information processing systems*, 30, 2017.

Hosseini, B., Hsu, A. W., and Taghvaei, A. Conditional optimal transport on function spaces. *arXiv preprint arXiv:2311.05672*, 2023.

Jia, X., Chen, Y., Mao, X., Duan, R., Gu, J., Zhang, R., Xue, H., and Cao, X. Revisiting and exploring efficient fast adversarial training via law: Lipschitz regularization and auto weight averaging. *arXiv preprint arXiv:2308.11443*, 2023.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Kim, H., Papamakarios, G., and Mnih, A. The lipschitz constant of self-attention, 2021.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Latorre, F., Rolland, P., and Cevher, V. Lipschitz constant estimation of neural networks via sparse polynomial optimization. *arXiv preprint arXiv:2004.08688*, 2020.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. W. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Leino, K., Wang, Z., and Fredrikson, M. Globally-robust neural networks. In *International Conference on Machine Learning*, pp. 6212–6222. PMLR, 2021.

Likhosherstov, V., Choromanski, K., and Weller, A. On the expressive power of self-attention matrices. *arXiv preprint arXiv:2106.03764*, 2021.

Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., and Shazeer, N. Generating wikipedia by summarizing long sequences. *arXiv:1801.10198*, 2018.

Lu, Y., Li, Z., He, D., Sun, Z., Dong, B., Qin, T., Wang, L., and Liu, T.-Y. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*, 2019.

Miyato, T., Maeda, S.-i., Koyama, M., Nakae, K., and Ishii, S. Distributional smoothing with virtual adversarial training. *arXiv preprint arXiv:1507.00677*, 2015.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

OpenAI. Gpt-4 technical report, 2023.

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pp. 582–597. IEEE, 2016.

Pevny, T. and Kovarik, V. Approximation capability of neural networks on spaces of probability measures and tree-structured domains. *arXiv preprint arXiv:1906.00764*, 2019.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Qi, X., Wang, J., Chen, Y., Shi, Y., and Zhang, L. Lipsformer: Introducing lipschitz continuity to vision transformers. *arXiv preprint arXiv:2304.09856*, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rosca, M., Weber, T., Gretton, A., and Mohamed, S. A case for new neural network smoothness constraints. *PMLR*, 2020.

Sander, M. E., Ablin, P., Blondel, M., and Peyré, G. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pp. 3515–3530. PMLR, 2022.

Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Sokolić, J., Giryes, R., Sapiro, G., and Rodrigues, M. R. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017.

Sra, S., Nowozin, S., and Wright, S. J. *Optimization for machine learning*. Mit Press, 2012.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tsuzuku, Y., Sato, I., and Sugiyama, M. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in neural information processing systems*, 31, 2018.

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.

Virmaux, A. and Scaman, K. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

von Luxburg, U. and Bousquet, O. Distance-based classification with lipschitz functions. *J. Mach. Learn. Res.*, 5 (Jun):669–695, 2004.

Vuckovic, J., Baratin, A., and des Combes, R. T. A mathematical theory of attention. *ArXiv*, abs/2007.02876, 2020.

Vuckovic, J., Baratin, A., and Combes, R. T. d. On the regularity of attention. *arXiv preprint arXiv:2102.05628*, 2021.

Vyas, A., Katharopoulos, A., and Fleuret, F. Fast transformers with clustered attention. *Advances in Neural Information Processing Systems*, 33:21665–21674, 2020.

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. *arXiv preprint arXiv:1801.10578*, 2018.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pp. 10524–10533. PMLR, 2020.

Ye, W., Ma, Y., Cao, X., and Tang, K. Mitigating transformer overconfidence via lipschitz regularization. *arXiv preprint arXiv:2306.06849*, 2023.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Zhang, B. and Sennrich, R. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhang, X., Zhao, J. J., and LeCun, Y. Character-level convolutional networks for text classification. In *NIPS*, 2015.

Zhao, H., Jiang, L., Jia, J., Torr, P., and Koltun, V. Point transformer. arxiv. *arXiv preprint arXiv:2012.09164*, 2020.

Zweig, A. and Bruna, J. A functional perspective on learning symmetric functions with neural networks. In *International Conference on Machine Learning*, pp. 13023–13032. PMLR, 2021.

A. Optimal Transport Toolbox

This section gathers some useful definitions and lemmas from optimal transport. In what follows, $\mathcal{X}$ is a Borel subset of $\mathbb{R}^d$.

A.1. Pushforward, Wasserstein Distance

Let us start with the notion of pushforward.

Definition A.1 (Pushforward)**.** Set μ a probability measure on $\mathcal{X}$ and $\varphi\colon \mathcal{X} \to \mathcal{X}$ a measurable function. The pushforward of μ by φ, denoted $\varphi_\sharp\mu$, is the probability measure given by

$$(\varphi_\sharp\mu)(B) = \mu(\varphi^{-1}(B))$$

for any Borel set $B \subset \mathcal{X}$, where $\varphi^{-1}(B) \coloneqq \{x \in \mathbb{R}^d : \varphi(x) \in B\}$.

The pushforward measure $\varphi_\sharp\mu$ can be seen as the result of a transportation of the mass of μ by φ. Intuitively, $\varphi_\sharp\mu$ is obtained by transporting each element of mass $\mu(\mathrm{d}x)$ from x to $\varphi(x)$.

Another crucial tool is the notion of Wasserstein distance.

Definition A.2 (Wasserstein space, Wasserstein distance)**.** Let $p \geq 1$. Denote

$$\mathcal{P}_p(\mathcal{X}) \coloneqq \{\mu \in \mathcal{P}(\mathcal{X}) : \int_{\mathcal{X}} |x|^p \mathrm{d}\mu(x) < \infty\}$$

the p-Wasserstein space. Then, the p-Wasserstein distance between two probability measures $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ is defined as

$$W_p(\mu,\nu) \coloneqq \left(\inf_{\pi \in \Pi(\mu,\nu)} \int |x-y|^p \mathrm{d}\pi(x,y)\right)^{1/p}$$

with $\Pi(\mu,\nu)$ the set of all couplings between μ and ν, i.e. of all probability measures $\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{X})$ such that $\int \pi(\cdot, y)\mathrm{d}y = \mu$ and $\int \pi(x,\cdot)\mathrm{d}x = \nu$.

Wasserstein distances have the following nice property, which is a direct consequence of Jensen inequality.

Lemma A.3. *For every $p \geq 1$, it holds*

$$W_1 \leq W_p.$$

The distance W_1 has also a simple dual formulation.

Lemma A.4 (W_1 duality formula)**.** *The distance W_1 can be rewritten with the so-called duality formula: for all $\mu, \nu \in \mathcal{P}_1(\mathcal{X})$, it holds*

$$W_1(\mu,\nu) = \sup_{\mathrm{Lip}(\varphi)\leq 1} \int \varphi \,\mathrm{d}(\mu-\nu), \tag{2}$$

where the supremum is taken over all functions $\varphi\colon \mathcal{X} \to \mathbb{R}$ with a Lipschitz constant bounded by one.

The following result is useful to bound the Wasserstein distance between two probability measures that are pushed forward by the same map.

Lemma A.5. *Let $p \geq 1$. Consider a measurable function $\varphi\colon \mathcal{X} \to \mathcal{X}$, and probability measures $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ such that $\varphi_\sharp\mu \in \mathcal{P}_p(\mathcal{X})$ and $\varphi_\sharp\nu \in \mathcal{P}_p(\mathcal{X})$. Then, it holds*

$$W_p(\varphi_\sharp\mu, \varphi_\sharp\nu) \leq \mathrm{Lip}(\varphi) W_p(\mu,\nu).$$

Proof. We have

$$\begin{aligned}
W_p(\varphi_\sharp\mu, \varphi_\sharp\nu)^p &= \inf_{\pi' \in \Pi(\varphi_\sharp\mu, \varphi_\sharp\nu)} \int \|x-y\|^p \,\mathrm{d}\pi'(x,y) \\
&\leq \inf_{\pi \in \Pi(\mu,\nu)} \int \|\varphi(x)-\varphi(y)\|^p \,\mathrm{d}\pi(x,y) \\
&\leq \mathrm{Lip}(\varphi)^p \inf_{\pi \in \Pi(\mu,\nu)} \int \|x-y\|^p \,\mathrm{d}\pi(x,y) \\
&= \mathrm{Lip}(\varphi)^p W_p(\mu,\nu)^p,
\end{aligned}$$

where the first inequality derives from the fact that every $\pi \in \Pi(\mu, \nu)$ induces a coupling $\pi' \in \Pi(\varphi_\sharp \mu, \varphi_\sharp \nu)$ by setting

$$\pi'(B_1 \times B_2) \coloneqq \pi(\varphi^{-1}(B_1) \times \varphi^{-1}(B_2))$$

for all Borel sets $B_1, B_2 \subset \mathcal{X}$, and that with this choice of π' it holds

$$\int \|x - y\|^p \, \mathrm{d}\pi'(x, y) = \int \|\varphi(x) - \varphi(y)\|^p \, \mathrm{d}\pi(x, y).$$

□

Let us now bound the Wasserstein distance between two different pushforwards of the same probability measure.

Lemma A.6. *Let $p \geq 1$. Consider two measurable functions $\varphi, \psi \colon \mathcal{X} \to \mathcal{X}$, and a probability measure $\mu \in \mathcal{P}_p(\mathcal{X})$ such that $\varphi_\sharp \mu \in \mathcal{P}_p(\mathcal{X})$ and $\psi_\sharp \mu \in \mathcal{P}_p(\mathcal{X})$. Then, it holds*

$$W_p(\varphi_\sharp \mu, \psi_\sharp \mu) \leq \|\varphi - \psi\|_{L^p(\mu)}.$$

Proof. Recall that

$$W_p(\varphi_\sharp \mu, \psi_\sharp \mu)^p = \inf_{\pi' \in \Pi(\varphi_\sharp \mu, \psi_\sharp \mu)} \int \|x - y\|^p \, \mathrm{d}\pi'(x, y).$$

Now consider the following coupling between $\varphi_\sharp \mu$ and $\psi_\sharp \mu$, defined by the relation

$$\pi'(B \times C) \coloneqq \mu(\varphi^{-1}(B) \cap \psi^{-1}(C))$$

for every Borel sets $B, C \subset \mathcal{X}$. In other words, we set $\mathrm{d}\pi'(y, z) \coloneqq \int_{\varphi^{-1}(y) \cap \psi^{-1}(z)} \mathrm{d}\mu$, and $\mathrm{d}\pi'(y, z) = 0$ if $\varphi^{-1}(y) \cap \psi^{-1}(z) = \emptyset$. With this definition of π', we have

$$W_p(\varphi_\sharp \mu, \psi_\sharp \mu)^p \leq \int \|x - y\|^p \, \mathrm{d}\pi'(x, y) = \int \|\varphi(x) - \psi(x)\|^p \, \mathrm{d}\mu(x).$$

□

A.2. Geodesics

The notion of a geodesic is useful for the following section of the Appendix.

Definition A.7 (Geodesic)**.** Let $(\mathcal{E}, d_\mathcal{E})$ be a metric space. A curve $\gamma \colon [0, 1] \to \mathcal{E}$ is called a geodesic if there exists a constant $v \geq 0$ such that for all $t_1, t_2 \in [0, 1]$ we have

$$d_\mathcal{E}(\gamma(t_1), \gamma(t_2)) = v|t_2 - t_1|.$$

We say that the space $\mathcal{E}$ is geodesic if for any $x, y \in \mathcal{E}$, there exists a geodesic between x and y.

One important example of geodesic space is the 2-Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$.

Lemma A.8 (Santambrogio (2015))**.** *The space $\mathcal{P}_2(\mathbb{R}^d)$ is a geodesic space.*

B. Local Lipschitz Constant in the Mean-Field Setting

B.1. Mean-Field Self-Attention Generalizes Self-Attention

For any input $X \in (\mathbb{R}^d)^n$, we have $F(m(X)) = m(f(X))$. Indeed, we can rewrite f as

$$f(X) = (\Gamma_X(x_1), \ldots, \Gamma_X(x_n)), \tag{3}$$

with

$$\Gamma_X \colon x \mapsto \frac{\sum_{i=1}^n \exp\left(\frac{1}{\sqrt{k}} x^\top Q^\top K x_i\right) V x_i}{\sum_{i=1}^n \exp\left(\frac{1}{\sqrt{k}} x^\top Q^\top K x_i\right)}.$$

Seeing the ratio of sums as a ratio of integrals against the empirical measure $m(X)$, and Equation (3) as a pushforward leads precisely to the formula of mean-field self-attention.

B.2. Local Lipschitz Constant in a General Metric Framework

The local Lipschitz constant can be defined for any function between two metric spaces, as follows.

Definition B.1 (Local Lipschitz constant)**.** Let $\varphi\colon \mathcal{E} \to \mathcal{F}$ be a map between two metric spaces. We define the local Lipschitz constant of φ at point $x \in \mathcal{E}$ as

$$\mathrm{Lip}_x(\varphi) \coloneqq \lim_{\varepsilon \to 0^+} \mathrm{Lip}\varphi_{|B(x,\varepsilon)}.$$

The limit exists, as $\mathrm{Lip}\varphi_{|B(x,\varepsilon)}$ decreases with ε.

Definition B.1 is interesting, as it captures more information than the global Lipschitz constant. More precisely, we have the following connection between the two notions.

Lemma B.2. *Let $\varphi\colon \mathcal{E} \to \mathcal{F}$ be a map between two metric spaces. We have*

$$\mathrm{Lip}(\varphi) \geq \sup_{x \in \mathcal{E}} \mathrm{Lip}_x(\varphi).$$

Assume moreover that the space $\mathcal{E}$ admits geodesics, which is the case for $\mathbb{R}^d$ and $\mathcal{P}_2(\mathbb{R}^d)$ equipped with W_2 (see A). Then, this inequality becomes an equality.

B.3. Proof of Lemma 3.2

Let us prove the following slightly stronger result.

Lemma B.3. *Let $p \geq 1$. For any matrix $X \in \mathbb{R}^{n \times d}$, we have*

$$\mathrm{Lip}_X^{\|\cdot\|_{F,p}}(f) = \mathrm{Lip}_{m(X)}^{W_p}(F_{|\mathcal{M}_n(\mathbb{R}^d)}) \leq \mathrm{Lip}_{m(X)}^{W_p}(F).$$

Proof. Set $X \in \mathbb{R}^{n \times d}$. One can choose $\varepsilon_1 > 0$ small enough (for example $\varepsilon_1 < \min_{x_i \neq x_j} \|x_i - x_j\|/2$) to have

$$\|X - Y\|_{F,p} \leq \varepsilon_1 \Rightarrow \|X - Y\|_{F,p} = W_p(m(X), m(Y)).$$

Indeed, if

$$\|X - Y\|_{F,p} \leq \varepsilon_1 < \min_{x_i \neq x_j} \|x_i - x_j\|/2,$$

then for all $i \in \{1, \dots, n\}$, we have $\|x_i - y_i\| \leq \varepsilon_1$, and thus x_i is the nearest neighbor (or one of the nearest neighbors) of y_i among the x_j. Similarly, one can choose $\varepsilon_2 > 0$ small enough to have

$$\|f(X) - f(Y)\|_{F,p} \leq \varepsilon_2 \Rightarrow \|f(X) - f(Y)\|_{F,p} = W_p(m(f(X)), m(f(Y))).$$

Then, we can set $\varepsilon \leq \varepsilon_1$ small enough to have

$$\|X - Y\|_{F,p} \leq \varepsilon \Rightarrow \|f(X) - f(Y)\|_{F,p} \leq \varepsilon_2,$$

and it holds, for all $\eta \leq \varepsilon$ and all Y such that $\|X - Y\|_{F,p} \leq \eta$, that

$$\|X - Y\|_{F,p} = W_p(m(X), m(Y))$$

and

$$\|f(X) - f(Y)\|_{F,p} = W_p(m(f(X)), m(f(Y))).$$

Now for all $\eta \leq \varepsilon$, we have

$$\begin{aligned}
\mathrm{Lip}^{\|\cdot\|_{F,p}} f_{|B_{\|\cdot\|_{F,p}}(X,\eta)} &= \sup_{Y \in B_{\|\cdot\|_{F,p}}(X,\eta)} \frac{\|f(X) - f(Y)\|_{F,p}}{\|X - Y\|_{F,p}} \\
&= \sup_{Y \in B_{\|\cdot\|_{F,p}}(X,\eta)} \frac{W_p(m(f(X)), m(f(Y)))}{W_p(m(X), m(Y))} \\
&= \sup_{Y \in B_{\|\cdot\|_{F,p}}(X,\eta)} \frac{W_p(F(m(X)), F(m(Y)))}{W_p(m(X), m(Y))},
\end{aligned}$$

by definition of ε and F. We conclude the proof by noticing that:

- $Y \in B_{\|\cdot\|_{F,p}}(X, \eta)$ implies $m(Y) \in B_{W_p}(m(X), \eta)$, which shows that $\mathrm{Lip}^{\|\cdot\|_{F,p}} f_{|B_{\|\cdot\|_{F,p}}(X,\eta)} \leq \mathrm{Lip}^{W_p} F_{|\mathcal{M}_n(\mathbb{R}^d)}$,
- $\mu \in B_{W_p}(m(X), \eta)$ with $\mu \in \mathcal{M}_n(\mathbb{R}^d)$ implies the existence of $Y \in \mathbb{R}^{n\times d}$ such that $\mu = m(Y)$ and $\|X - Y\|_{F,p} = W_p(m(X), m(Y))$, so that $Y \in B_{\|\cdot\|_{F,p}}(X, \eta)$. Indeed, take Y such that $\mu_n = m(Y)$ and then permute its coordinates so that x_i becomes the nearest neighbor (or one of the nearest neighbors) of y_i. This shows the reverse inequality and concludes the proof.

□

C. Proofs of Section 3

We have the following useful Lemma.

Lemma C.1. *Let μ be a probability measure supported in $\overline{B}(x_0, R) \subset \mathbb{R}^d$, with any $x_0 \in \mathbb{R}^d$ and $R > 0$. Then, denoting $\mathrm{Var}\mu \coloneqq \mathbb{E}[(Z - \mathbb{E}Z)(Z - \mathbb{E}Z)^\top]$ with Z a random variable distributed according to μ, we have*

$$\|\mathrm{Var}\mu\|_2 \leq R^2,$$

with equality when $\mu = \frac{1}{2}(\delta_{x_0+x} + \delta_{x_0-x})$ for any $x \in \mathbb{R}^d$ such that $|x| = R$.

Proof. Let us assume without loss of generality that $x_0 = 0$. It is straightforward to check that if $\mu = \frac{1}{2}(\delta_x + \delta_{-x})$ then $\|\mathrm{Var}\mu\|_2 = R^2$. To show that this is the maximal value the variance can take, we use the triangle inequality:

$$\begin{aligned}\left\|\mathbb{E}[(Z - \mathbb{E}Z)(Z - \mathbb{E}Z)^\top]\right\|_2 &\leq \mathbb{E}\left\|(Z - \mathbb{E}Z)(Z - \mathbb{E}Z)^\top\right\|_2 \\ &= \mathbb{E}[|Z - \mathbb{E}Z|^2] \\ &= \mathbb{E}[|Z|^2] - |\mathbb{E}Z|^2.\end{aligned}$$

Now let us pick any $x \in B_R \setminus B(0, r)$. We have $\mathbb{E}(Z - x)^\top(Z + x) \leq 0$, as the angle between the vectors $Z - x$ and $Z + x$ is at least $\pi/2$ for Z in $B(0, R)$. By expanding this relationship we get

$$E[|Z|^2] - |x|^2 \leq 0,$$

which yields the result. □

C.1. The Jacobian of Self-Attention

Lemma C.2. *Let $X = (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$. For all perturbations $\epsilon \coloneqq (\epsilon_1, \ldots, \epsilon_n) \in (\mathbb{R}^d)^n$ and all $i \in \{1, \ldots, n\}$, we have*

$$(D_X f)(\epsilon)_i = V \sum_{j=1}^n P_{ij}(x_j - \sum_{k=1}^n P_{ik}x_k)x_i^\top A^\top \epsilon_j + V \sum_{j=1}^n P_{ij}\epsilon_j + V \sum_{j=1}^n P_{ij}(x_j - \sum_{k=1}^n P_{ik}x_k)x_j^\top A\epsilon_i,$$

with $P_{ij} \coloneqq e^{x_i^\top A^\top x_j} / \sum_{k=1}^n e^{x_i^\top A^\top x_k}$.

C.2. Proof of Theorem 3.3

Let $X \in B_R^n$ and $\epsilon \coloneqq (\epsilon_1, \ldots, \epsilon_n) \in (\mathbb{R}^d)^n$ such that $\|\epsilon\|_F = 1$. According to Lemma C.2, for all $i \in \{1, \ldots, n\}$ we have

$$(D_X f)(\epsilon)_i = V \sum_{j=1}^n P_{ij}(x_j - \sum_{k=1}^n P_{ik}x_k)x_i^\top A^\top \epsilon_j + V \sum_{j=1}^n P_{ij}\epsilon_j + V \sum_{j=1}^n P_{ij}(x_j - \sum_{k=1}^n P_{ik}x_k)x_j^\top A\epsilon_i,$$

with $P_{ij} \coloneqq e^{x_i^\top A^\top x_j} / \sum_{k=1}^n e^{x_i^\top A^\top x_k}$. The triangle inequality gives

$$|(D_X f)(\epsilon)_i| \leq \|V\|_2 \left(\|A\|_2 R \left| \sum_{j=1}^n P_{ij}(x_j - \sum_{k=1}^n P_{ik}x_k)\epsilon_j \right| + \left| \sum_{j=1}^n P_{ij}\epsilon_j \right| + \|A\|_2 \left\| \mathrm{Var}^{(i)} \right\|_2 |\epsilon_i| \right)$$

where $\mathrm{Var}^{(i)} \coloneqq \sum_{j=1}^n P_{ij}(x_j - \sum_{k=1}^n P_{ik}x_k)x_j^\top$ is the variance of the probability measure $\sum_{j=1}^n P_{ij}\delta_{x_j}$. Lemma C.1 gives us that $\left\|\mathrm{Var}^{(i)}\right\|_2 \leq R^2$. We can also apply Cauchy-Schwarz inequality to get

$$\left|\sum_{j=1}^n P_{ij}(x_j - \sum_{k=1}^n P_{ik}x_k)\epsilon_j\right| \leq \left(\sum_{j=1}^n P_{ij}\left|x_j - \sum_{k=1}^n P_{ik}x_k\right|^2\right)^{1/2}\left(\sum_{j=1}^n P_{ij}|\epsilon_j|^2\right)^{1/2}.$$

The proof of Lemma C.1 allows us to bound

$$\sum_{j=1}^n P_{ij}\left|x_j - \sum_{k=1}^n P_{ik}x_k\right|^2 \leq R^2.$$

Collecting terms, we get

$$|(D_X f)(\epsilon)_i| \leq \|V\|_2 \left(\|A\|_2 R^2 \left(\sum_{j=1}^n P_{ij}|\epsilon_j|^2\right)^{1/2} + \left|\sum_{j=1}^n P_{ij}\epsilon_j\right| + \|A\|_2 R^2|\epsilon_i|\right).$$

Then, using the inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ for any $a, b, c \in \mathbb{R}$, we obtain

$$\sum_{i=1}^n |(D_X f)(\epsilon)_i|^2 \leq 3\|V\|_2^2 \left(\|A\|_2^2 R^4(n+1) + n\right),$$

where we have used that with the triangle inequality and Cauchy-Schwarz inequality

$$\left|\sum_{j=1}^n P_{ij}\epsilon_j\right|^2 \leq \left(\sum_{j=1}^n P_{ij}|\epsilon_j|\right)^2 \leq \sum_{j=1}^n P_{ij}^2 \|\epsilon\|_F^2 \leq 1$$

as $\|\epsilon\|_F^2 = 1$ and $\sum_{j=1}^n P_{ij}^2 \leq \sum_{j=1}^n P_{ij} = 1$, and that

$$\sum_{j=1}^n P_{ij}|\epsilon_j|^2 \leq \sum_{j=1}^n |\epsilon_j|^2 = 1.$$

The proof above allows us to recover the following tighter bound but with maybe less natural assumptions on the tokens.

Theorem C.3. *Let $Q, K, V \in \mathbb{R}^{k\times d}$ and $A \coloneqq K^\top Q/\sqrt{k}$. Denote f unmasked self-attention with parameters (A, V). Let $X \in (\mathbb{R}^d)^n$ and $\rho, r > 0$ such that*

(i) $\max_{1\leq i,j\leq n}|x_i - x_j| \leq r$,

(ii) $\max_{1\leq i\leq n}|Ax_i| \leq \rho$.

Then, the local Lipschitz constant of f at X is bounded by

$$\|D_X f\|_2 \leq \sqrt{3}\|V\|_2 \left((\rho^2 r^2 + 1)n + \rho^2 r^2\right)^{1/2}.$$

Moreover, in practical Transformers architectures, inputs X of self-attention can be written as $\mathrm{norm}(\tilde{X})$ for some $\tilde{X} \in (\mathbb{R}^d)^n$, where norm stands for LayerNorm of RMSNorm (see Subsection 2.3). In both cases, all inputs are on an ellipsis whose shape depends on the parameters of LayerNorm or RMSNorm, so that there is a choice of parameters (r, ρ) such that for all $\tilde{X} \in (\mathbb{R}^d)^n$, the input $X = \mathrm{norm}(\tilde{X})$ satisfies assumptions (i) and (ii) in Theorem C.3.

C.3. Weighted Self-Attention

In the whole subsection, $\Sigma_n \coloneqq \{a \in [0,1]^n : \sum_{i=1}^n a_i = 1\}$ is the simplex.

In view of proving Propositions 3.4 and 3.6, let us introduce a framework for self-attention that extends the Euclidean framework, where the local Lipschitz constant is given by the operator norm of the differential, to probability measures with a finite number of diracs. We call this framework weighted self-attention.

Definition C.4 (Weighted self-attention)**.** For any vector $a \in \Sigma_n$, denote $\mathcal{P}_a(\mathbb{R}^d) \coloneqq \{\sum_{i=1}^n a_i \delta_{x_i} : x_1, \dots, x_n \in \mathbb{R}^d\}$. We define the Euclidean version of the restriction of self-attention with parameters (A, V) to $\mathcal{P}_a(\mathbb{R}^d)$ in the following way:

$$f_a \colon (x_1, \dots, x_n) \in (\mathbb{R}^d)^n \mapsto \left(V \sum_{j=1}^n P_{ij}^a x_j \right)_{1 \le i \le n} \in (\mathbb{R}^k)^n$$

with $P_i^a \coloneqq \mathrm{softmax}_a\left((x_i^\top A^\top x_j)_{1 \le j \le n}\right)$, where $\mathrm{softmax}_a(w_1, \dots, w_n) \coloneqq \left(\frac{a_i e^{w_i}}{\sum_j a_j e^{w_j}}\right)_{1 \le i \le n}$. The function f_a is called weighted self-attention associated to the coefficients a.

Definition C.4 is designed so that for any sequence $X \in (\mathbb{R}^d)^n$, it holds

$$F(m_a(X)) = m_a(f_a(X)),$$

with $m_a(X) \coloneqq \sum_{i=1}^n a_i \delta_{X_i}$. Weighted self-attention provides a representation that is very convenient to study both theoretically and numerically the local Lipschitz constant of self-attention at measures that have a finite number of Diracs but with weights such that they require a massive sequence length to be approximated well by an empirical measure. This is for example the case of measures of the form $e^{-2R^2}\delta_R + (1 - e^{-2R^2})\delta_{-R}$, which are at stake in the proof of Proposition 3.6. To study the Lipschitz continuity of f_a, we equip the space $(\mathbb{R}^d)^n$ with the norm

$$\|X\|_a \coloneqq \left(\sum_{i=1}^n a_i |x_i|^2 \right)^{1/2},$$

so that we have the following connection with the local Lipschitz constant of mean-field self-attention in the Wasserstein 2 sense.

Lemma C.5. *Let $X \in (\mathbb{R}^d)^n$ be an input sequence, and $a \in \Sigma_n$ a vector of coefficients. Then, we have*

$$\mathrm{Lip}^{W_2}_{m_a(X)} F_{|\mathcal{P}_a(\mathbb{R}^d)} = \|D^a_X f_a\|_{2,a},$$

where D^a is the Jacobian in the space $((\mathbb{R}^d)^n, \|\cdot\|_a)$ and $\|\cdot\|_{2,a}$ is the corresponding operator norm.[2]

This is a nice property, as it provides an optimized way to compute numerically the local Lipschitz constant of f_a, just as for Euclidean self-attention. Lemma C.5 can be proven with the same steps as for Lemma 3.2. Moreover, the differential of f_a has the following expression.

Lemma C.6. *Let $X = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$. For all perturbations $\epsilon \coloneqq (\epsilon_1, \dots, \epsilon_n) \in (\mathbb{R}^d)^n$ and all $i \in \{1, \dots, n\}$, we have*

$$(D^a_X f_a)(\epsilon)_i = V \sum_{j=1}^n P_{ij}^a (x_j - \sum_{k=1}^n P_{ik}^a x_k) x_i^\top A^\top \epsilon_j + V \sum_{j=1}^n P_{ij}^a \epsilon_j + V \sum_{j=1}^n P_{ij}^a (x_j - \sum_{k=1}^n P_{ik}^a x_k) x_j^\top A \epsilon_i,$$

with $P_{ij}^a \coloneqq a_j e^{x_i^\top A^\top x_j} / \sum_{k=1}^n a_k e^{x_i^\top A^\top x_k}$.

C.4. Proof of Proposition 3.4

Proposition 3.4 follows from the following Lemma.

Lemma C.7. *Let γ be a real eigenvalue of A and $u \in \mathbb{R}^d$ an associated unit eigenvector.*

[2] As we are in finite dimension, all norms are equivalent so the linear operator D^a is just the usual notion of differential.

1. *If $\gamma \geq 0$, denote $X \coloneqq (u, \frac{u}{2}, \dots, \frac{u}{2}) \in (\mathbb{R}^d)^n$. Then, for any scaling factor $R > 0$, it holds*

$$\|D_{RX} f\|_2 \geq \frac{1}{1 + (n-1)e^{-R^2\gamma/4}} \sqrt{n-1}.$$

2. *If $\gamma < 0$, denote $X \coloneqq (u, -u, \dots, -u) \in (\mathbb{R}^d)^n$. Then, for any scaling factor $R > 0$, it holds*

$$\|D_{RX} f\|_2 \geq \frac{1}{1 + (n-1)e^{-2R^2|\gamma|}} \sqrt{n-1}.$$

Proof. Let us start with the case $\gamma < 0$. Let $\tilde{X} \coloneqq (u, -u)$. Using weighted self-attention f_a introduced in Subsection C.3, and $a \coloneqq (1/n, 1 - 1/n)$, we have for any $\varepsilon \in \mathbb{R}^d$ and $1 \leq i \leq n$ that

$$D_{R\tilde{X}} f_a(\varepsilon)_i = \sum_{j=1}^n P^a_{ij} \varepsilon_j + \sum_{j=1}^n P^a_{ij} x_j (x_j - \sum_{k=1}^n P^a_{ik} x_k)^\top A \varepsilon_i + \sum_{j=1}^n P^a_{ij} (x_j - \sum_{k=1}^n P^a_{ik} x_k) x_i^\top A^\top \varepsilon_j.$$

Setting $\varepsilon_1 \coloneqq -u$ and $\varepsilon_2 \coloneqq 0$, we get

$$D_{R\tilde{X}} f_a(\varepsilon)_2 = P^a_{21}(1 + 2P^a_{22} R^2 |\gamma|) \geq P^a_{21}.$$

Moreover

$$P^a_{21} = \frac{1}{1 + (n-1)e^{-2R^2|\gamma|}},$$

so that

$$\|D_{RX} f\|_2 \geq \|D_{R\tilde{X}} f_a\|_{2,a} \geq P^a_{21} \sqrt{n-1},$$

which proves the result. When $\gamma \geq 0$, the same method applies with $\varepsilon_1 = u$ and $\varepsilon_2 = 0$. Denoting $\tilde{X} \coloneqq (u, u/2)$ and $a \coloneqq (1/n, 1 - 1/n)$ one obtains

$$\|D_{RX} f\|_2 \geq \|D_{R\tilde{X}} f_a\|_2 \geq P^a_{21} \sqrt{n-1}$$

and

$$P^a_{21} = \frac{1}{1 + (n-1)e^{-R^2\gamma/4}},$$

which proves the result. □

C.5. About the scaling factor in Proposition 3.4

Let us investigate numerically the scaling factor $2R^2\gamma$ that appears in Proposition 3.4. With the model BERT pretrained, and with a batch of five text extracts from the dataset Alice in Wonderland, we plot $\gamma^{(h)} R^2$ for each layer $0 \leq \ell \leq 11$ and for $h \in \{0, 5, 10\}$, where $\gamma^{(h)}$ is the parameter γ defined in Proposition 3.4 associated to the $A^{(h)}$ the weight matrix of head h, and R is the mean magnitude of tokens as they enter the attention block of layer ℓ, defined as $R\colon \sqrt{1/n \sum_{i=1}^n |x_i|^2}$. We obtain Figure 2.

C.6. Proof of Proposition 3.6

Let us prove the following result, which implies Proposition 3.6.

Proposition C.8. *Let $R > 0$. Assume that $k = d$ and $V = I_d$, and denote $\gamma_1 \geq \cdots \geq \gamma_\nu$ the real eigenvalues of A. Then, the following claims hold.*

1. *If $\gamma_1 \geq -8\gamma_\nu$, denoting $C \coloneqq \frac{\gamma_1}{8}$, there exists a function $\theta\colon [0, +\infty) \to [0, +\infty)$ such that $\theta(R) \to_{R \to +\infty} 1$ and:*

$$\mathrm{Lip}^{W_2}(F_{|\mathcal{P}(B_R)}) \geq \theta(R) \frac{C}{2} R^2 e^{CR^2}.$$

Moreover, the right-hand side is equivalent to the Lipschitz constant of mean-field self-attention at the probability measure $e^{-2CR^2} \delta_{Ru_1} + (1 - e^{-2CR^2}) \delta_{(R/2)u_1}$.

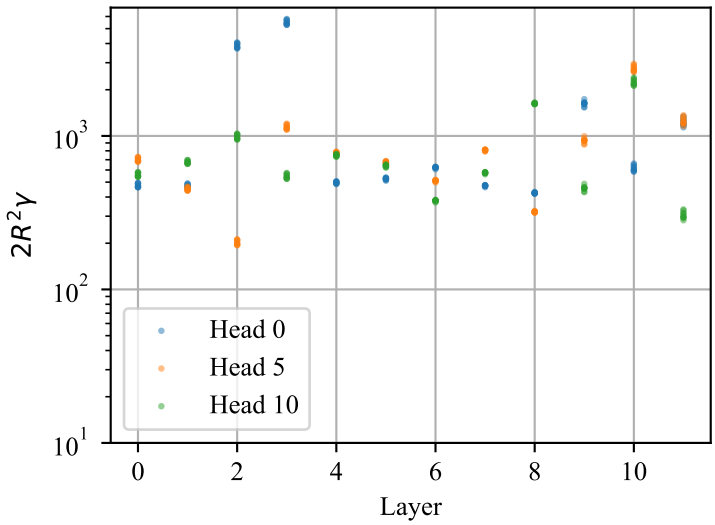

Figure 2. Plot of the scaling factor $2R^2\gamma$ across layers of BERT pretrained for three different heads and 5 text extracts of Alice in Wonderland (50 tokens for each extract).

2. *If $\gamma_1 < -8\gamma_\nu$, denoting $C' \coloneqq |\gamma_\nu|$, there exists a function $\theta\colon [0,+\infty) \to [0,+\infty)$ such that $\theta(R) \to_{R\to+\infty} 1$ and:*

$$\mathrm{Lip}^{W_2}(F_{|\mathcal{P}(B_R)}) \geq \theta(R)\frac{C'}{2}R^2 e^{C'R^2}.$$

Moreover, the right-hand side is equivalent to the Lipschitz constant of mean-field self-attention at the probability measure $e^{-2C'R^2}\delta_{Ru_\nu} + (1-e^{-2C'R^2})\delta_{-Ru_\nu}$.

Proof. Let us detail the proof in the second case only, as the first case is very similar. According to Lemma C.6 with $V = I_d$, for any $X = (x_1, x_2) \in (\mathbb{R}^d)^2$ and $a = (a_1, a_2) \in \Sigma_2$, for any $\epsilon_2 \in \mathbb{R}^d$, denoting $\epsilon \coloneqq (0, \epsilon_2) \in (\mathbb{R}^d)^2$ we have

$$D_X^a f_a(\epsilon)_1 = (P_{12}^a I_d + 2P_{11}^a P_{12}^a x_2 x_1^\top A^\top)\epsilon_2.$$

Let $u_1, \dots, u_\nu$ a family of unit eigenvectors of A, associated to the eigenvalues $\gamma_1 \geq \cdots \geq \gamma_\nu$. Let $R > 0$. Set $a_2 \coloneqq e^{-2|\gamma_\nu|R^2}$, so that $a_1 = 1 - e^{-2|\gamma_\nu|R^2}$, and choose $x_1 \coloneqq Ru_\nu$ and $x_2 \coloneqq -Ru_\nu$. Then

$$\|D_X^a f_a(\epsilon)\|_a \geq \sqrt{a_1}|P_{12}^a(I_d + 2P_{11}^a R^2|\gamma_\nu|u_\nu u_\nu^\top)\epsilon_2|. \tag{4}$$

Now let us notice that

$$P_{12}^a = \frac{a_2 e^{|\gamma_\nu|R^2}}{a_1 e^{-|\gamma_\nu|R^2} + a_2 e^{|\gamma_\nu|R^2}} = \frac{e^{-|\gamma_\nu|R^2}}{(1-e^{-2|\gamma_\nu|R^2})e^{-|\gamma_\nu|R^2} + e^{-|\gamma_\nu|R^2}} \to_{R\to+\infty} \frac{1}{2},$$

and

$$P_{11}^a = 1 - P_{12}^a \to_{R\to+\infty} \frac{1}{2}$$

as well. Going back to Equation (4) and setting $\epsilon_2 \coloneqq u_\nu$, we get

$$\|D_X^a f_a(\epsilon)\|_a \gtrsim \frac{1}{2}(1 + R^2|\gamma_\nu|)|u_\nu| \gtrsim \frac{1}{2}R^2|\gamma_\nu|,$$

where $f(R) \gtrsim g(R)$ means that there exists a function $\theta\colon \mathbb{R} \to \mathbb{R}_+$ such that $\theta(R) \to_{R\to+\infty} 1$ and $f(R) \geq \theta(R)g(R)$. Finally, we get

$$\frac{\|D_X^a f_a(\epsilon)\|_a}{\|\epsilon\|_a} \gtrsim \frac{1}{2}|\gamma_\nu|R^2 {a_1}^{-1/2} = \frac{1}{2}|\gamma_\nu|R^2 e^{|\gamma_\nu|R^2},$$

which proves the result as $\|D_X^a f_a\|_{2,a} = \sup_{\epsilon\in(\mathbb{R}^d)^2\setminus\{0\}} \frac{\|D_X^a f_a(\epsilon)\|_a}{\|\epsilon\|_a}$. □

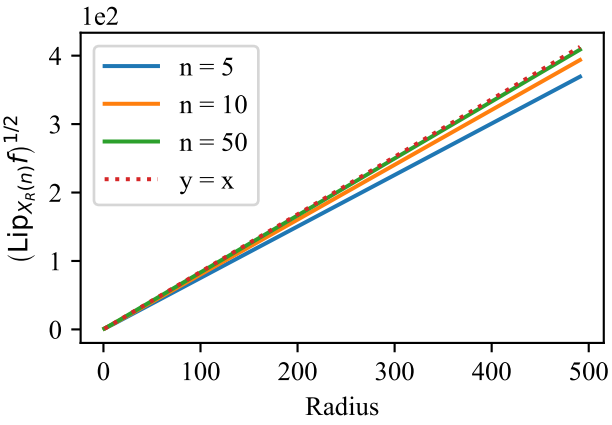

Figure 3. Linear growth of the square root of the Lipschitz constant of self-attention in the configuration $X_R(n)$.

C.7. An Example of Quadratic Growth with the Radius

For simplicity, assume that $A = I_d$ and $V = I_d$. Let $X = (0, x_2 \dots, x_n) \in B_R^n$. Kim et al. (2021) show that, as all norms are equivalent in finite dimension, the local Lipschitz constant of f at X grows like the empirical variance of X. Taking $x_2 = \cdots = x_j = Re_1$ and $x_{j+1} = \cdots = x_n = -Re_1$ with $e_1 = (1, 0, \dots, 0) \in \mathbb{R}^d$ and j such that $|2j - 1 - n| \leq 1$ and denoting X_R the resulting configuration, we obtain that the empirical variance of X_R grows like R^2 up to a constant factor that is close to 1.

This behavior can be easily checked numerically (see Figure 3).

D. Proofs of Section 4

We have the following formula for the Jacobian of masked self-attention.

Lemma D.1. *Let $X = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$. For all perturbations $\epsilon := (\epsilon_1, \dots, \epsilon_n) \in (\mathbb{R}^d)^n$ and all $i \in \{1, \dots, n\}$, we have*

$$(D_X f^m)(\epsilon)_i = V \sum_{j=1}^{i} P_{ij}(x_j - \sum_{k=1}^{i} P_{ik} x_k) x_i^\top A^\top \epsilon_j + V \sum_{j=1}^{i} P_{ij} \epsilon_j + V \sum_{j=1}^{i} P_{ij}(x_j - \sum_{k=1}^{i} P_{ik} x_k) x_j^\top A \epsilon_i,$$

with $P_{ij} := e^{x_i^\top A^\top x_j} / \sum_{k=1}^{i} e^{x_i^\top A^\top x_k}$.

D.1. Proofs of Theorem 4.3 and Theorem 4.5

In view of Theorem D.1, following the same steps as in the proof of Theorem 3.3 leads to Theorem 4.3. Likewise, the proof of Theorem 4.5 is the same as for Theorem 3.7.

D.2. Proof of Theorem 4.4

Let $\bar{\mu}$ and $\bar{\nu}$ be two distinct measures in $\mathcal{P}([0, 1] \times B_R)$. Assume that $\bar{\mu}$ and $\bar{\nu}$ have the same first marginal, denoted θ (otherwise, we consider that they are associated with an infinite Lipschitz ratio). We have

$$\begin{aligned} d_p(F^m(\bar{\mu}), F^m(\bar{\nu})) &\leq d_p((\Gamma_{\bar{\mu}})_\sharp \bar{\mu}, (\Gamma_{\bar{\mu}})_\sharp \bar{\nu}) + d_p((\Gamma_{\bar{\mu}})_\sharp \bar{\nu}, (\Gamma_{\bar{\nu}})_\sharp \bar{\nu}) \\ &\leq \left(\int_0^1 W_p\left((\Gamma_{\bar{\mu}}^s)_\sharp \mu^s, (\Gamma_{\bar{\mu}}^s)_\sharp \nu^s \right)^p \right)^{1/p} \\ &\quad + \left(\int_0^1 W_p\left(\Gamma_{\bar{\mu}}^s)_\sharp \nu^s, (\Gamma_{\bar{\nu}}^s)_\sharp \nu^s \right) \right)^{1/p}, \end{aligned}$$

where we denote

$$\Gamma^s_{\bar{\mu}}(x) := \frac{\int VyG(x,y)\mathbf{1}_{\tau\leq s}\mathrm{d}\bar{\mu}(\tau,y)}{\int G(x,y)\mathbf{1}_{\tau\leq s}\mathrm{d}\bar{\mu}(\tau,y)}.$$

Using Lemma A.5, we get

$$W_p\left((\Gamma^s_{\bar{\mu}})_\sharp\mu^s, (\Gamma^s_{\bar{\mu}})_\sharp\nu^s\right) \leq \mathrm{Lip}(\Gamma^s_{\bar{\mu}})W_p(\mu^s,\nu^s)$$

where the Lipschitz constant is taken on B_R. A similar computation as for traditional self-attention shows that for all $0\leq s\leq 1$ and $x\in B_R$ we have

$$\|D_x\Gamma^s\|_2 \leq \|V\|_2 \|A\|_2 R^2,$$

so that

$$\mathrm{Lip}(\Gamma^s_{\bar{\mu}}) \leq \|V\|_2 \|A\|_2 R^2.$$

To bound the second term in the previous inequality, we use Lemma A.6, to get

$$W_p\left((\Gamma^s_{\bar{\mu}})_\sharp\nu^s, (\Gamma^s_{\bar{\nu}})_\sharp\nu^s\right) \leq \left\|\Gamma^s_{\bar{\mu}} - \Gamma^s_{\bar{\nu}}\right\|_{L^\infty(B_R)}.$$

Again, a similar computation as for traditional self-attention shows that

$$\left\|\Gamma^s_{\bar{\mu}} - \Gamma^s_{\bar{\nu}}\right\|_{L^\infty(B_R)} \leq \|V\|_2 (2\|A\|_2 R^2) e^{2\|A\|_2 R^2} W_p(\mu^s,\nu^s),$$

which concludes the proof.

E. Experiments

E.1. Power Iteration

Let $X\in(\mathbb{R}^d)^n$. To compute numerically $\|D_Xf\|_2$, we pick an initialisation

$$u_0 \sim \otimes^{n\times d}\mathcal{N}(0,1)\in\mathbb{R}^{n\times d}$$
$$u_0 \leftarrow u_0/\|u_0\|_F$$

and then repeat the following steps until convergence:

$$v_k = (D_Xf)^\top(D_Xf)u_k$$
$$\mu_k = v_k^\top u_k$$
$$u_{k+1} = \frac{v_k}{\|v_{k+1}\|_F}$$

where v_k is computed by doing successively a Jacobian-vector product and a vector-Jacobian product. It is well known (Sra et al., 2012) that with this method, μ_k converges to $\|D_Xf\|_2$.

E.2. GPT-2

We do the same experiments as in Section 5 on GPT-2 pretrained (Radford et al., 2019). We see in Figure 4 that the behavior is different than what we observe in Figure 1. This is explained by the fact that the magnitude of tokens depends on the sequence length for pretrained GPT-2, due to the learned positional encoding (see Figures 5 and 6). Therefore, the bound of Theorem 3.3 still holds but the growth rate can be faster than $\sqrt{n}$.

E.3. Self-attention with biases

Some Transformer architectures such as BERT (Devlin et al., 2018) add biases $(b_Q,b_K,b_V)\in\mathbb{R}^k$ in the formula of self-attention:

$$f^b\colon (x_1,\dots,x_n)\mapsto \left(V\textstyle\sum_{j=1}^n P_{ij}x_j + b_V\right)_{1\leq i\leq n}\in(\mathbb{R}^k)^n, \tag{5}$$
$$\text{with}\quad P_i := \mathrm{softmax}\left((Qx_i+b_Q)^\top(Kx_j+b_K))_{1\leq j\leq n}\right),$$

where we absorbed the factor $1/\sqrt{k}$ in Q,K,b_Q and b_K to alleviate notations.

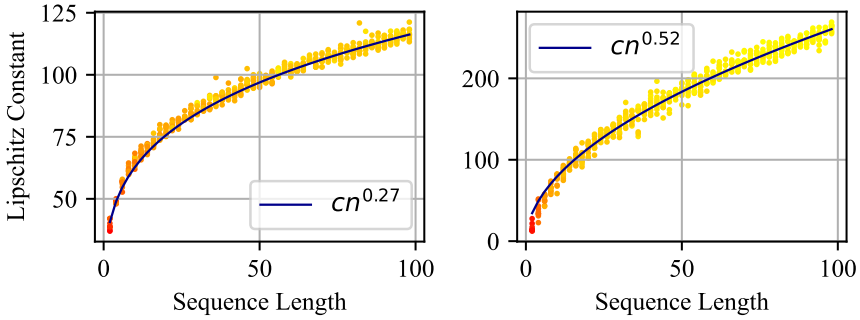

Figure 4. Scatter plots of the local Lipschitz constant of masked self-attention for GPT-2 pretrained as a function of the sequence length, on the dataset Alice in Wonderland. The first column corresponds to masked self-attention layer 0, and the second column to layer 6.

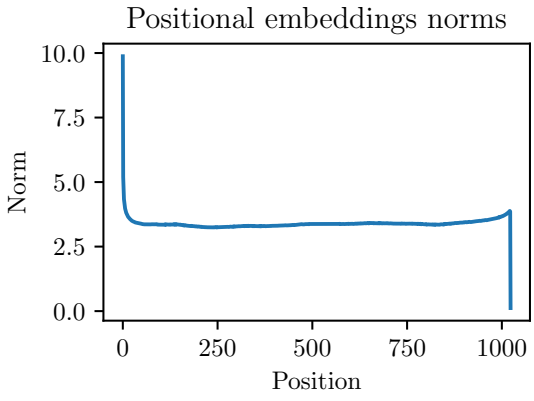

Figure 5. Norm of the positional embeddings of GPT-2 pretrained, ordered by position. The very first tokens are associated to positional embeddings of much larger magnitude, which makes n and R dependent from the very beginning of the architecture.

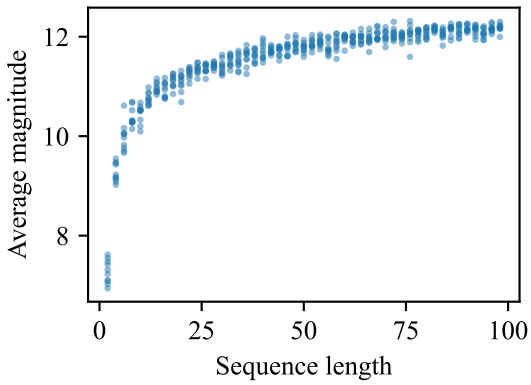

Figure 6. Scatter plot of the average magnitude of tokens $R \coloneqq \sqrt{1/n \sum_{i=1}^{n} |x_i|^2}$ as a function of the sequence length, for the same dataset as in Figure 4, right column.

Remark E.1. Note that b_K has no influence on the value of f^b:

$$\operatorname{softmax}\left((Qx_i + b_Q)^\top (Kx_j + b_K))_{1\leq j\leq n}\right) = \operatorname{softmax}\left((Qx_i + b_Q)^\top Kx_j)_{1\leq j\leq n}\right),$$

as b_K in only involved in terms that are independent of j.

How do biases in self-attention affect the bound in Theorem 3.3? We start by computing the Jacobian of self-attention with biases.

Lemma E.2. *Let $X = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$. Let f^b be a self-attention module with biases, defined as in Equation (5). For all perturbations $\epsilon \coloneqq (\epsilon_1, \dots, \epsilon_n) \in (\mathbb{R}^d)^n$ and all $i \in \{1, \dots, n\}$, we have*

$$(D_X f^b)(\epsilon)_i = V \sum_{j=1}^n P_{ij}(x_j - \sum_{k=1}^n P_{ik}x_k)(Qx_i + b_Q)^\top K\epsilon_j + V \sum_{j=1}^n P_{ij}\epsilon_j + V \sum_{j=1}^n P_{ij}(x_j - \sum_{k=1}^n P_{ik}x_k)x_j^\top A\epsilon_i,$$

with $P_{ij} \coloneqq e^{(Qx_i+b_Q)^\top (Kx_j+b_K)} / \sum_{k=1}^n e^{(Qx_i+b_Q)^\top (Kx_k+b_K)}$.

The same steps as in the proof of Theorem 3.3 lead to the following bound.

Theorem E.3. *Let $Q, K, V \in \mathbb{R}^{k\times d}$ and $A \coloneqq K^\top Q/\sqrt{k}$. Let $R > 0$ and $n \in \mathbb{N}$. Unmasked self-attention with biases f^b with parameters (A, V), defined in Equation (5), is Lipschitz continuous on the set B_R^n, with*

$$\operatorname{Lip}\left(f_{|B_R^n}\right) \leq \sqrt{3}\,\|V\|_2 \left(\|A\|_2^2 R^4 + n\left(\|K\|_2 (\|Q\|_2 R + |b_Q|)^2 R^2 + 1\right)\right)^{1/2}.$$