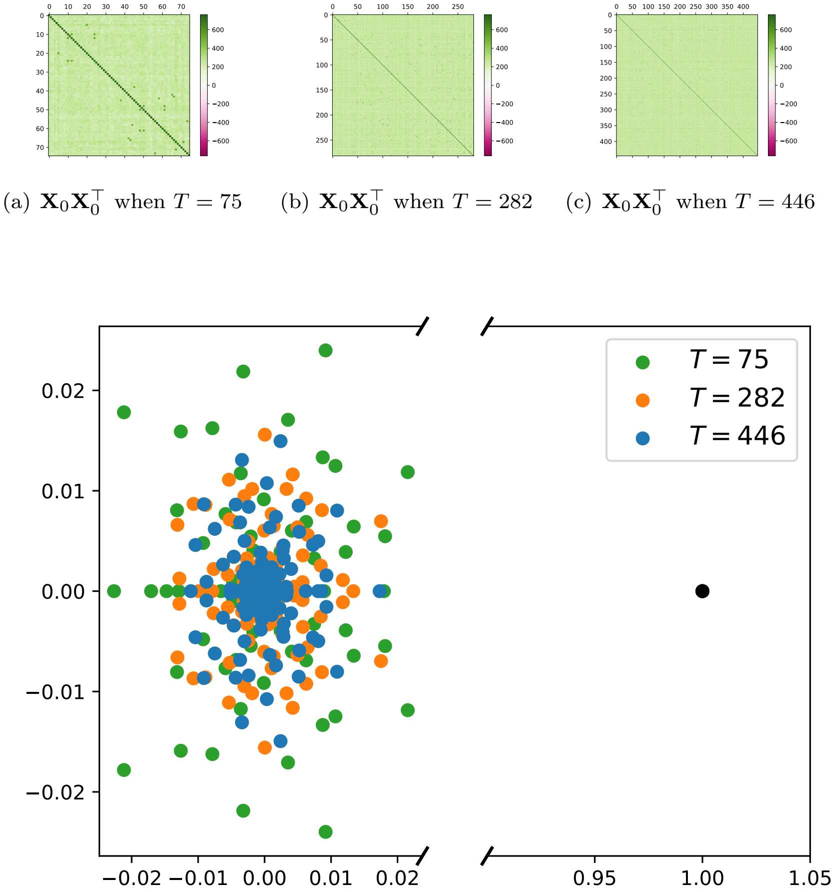
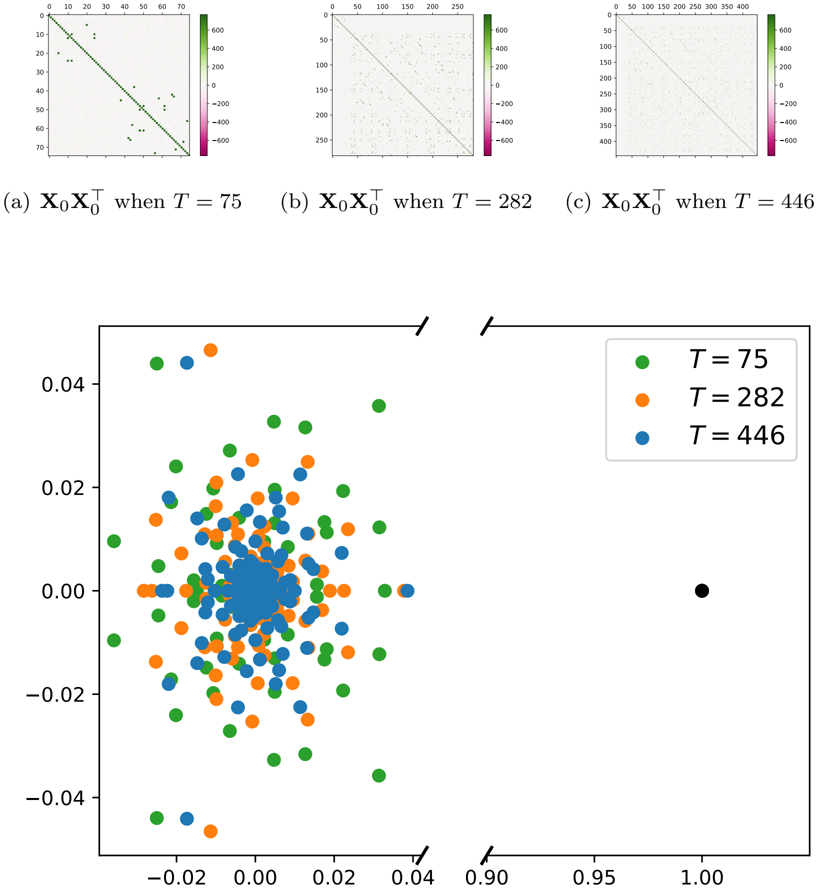


(a) $\mathbf{X}_0\mathbf{X}_0^\top$ when $T = 75$ (b) $\mathbf{X}_0\mathbf{X}_0^\top$ when $T = 282$ (c) $\mathbf{X}_0\mathbf{X}_0^\top$ when $T = 446$

(d) Eigenvalues in the complex plane of the attention matrix in the first head of layer 1.

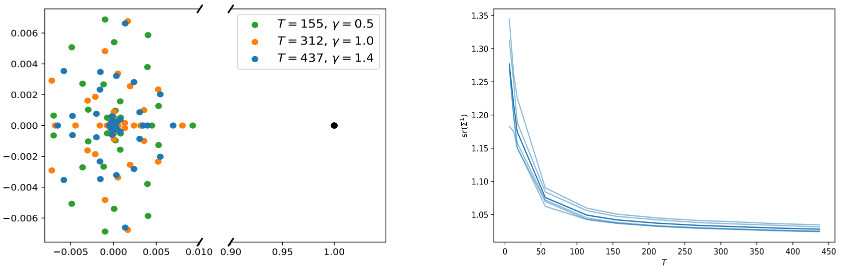
Figure I: Spectral gap (second row) persists even when the inputs are not orthogonal (see first row). Experiment conducted with a randomly initialised RoBERTa encoder, hence $d = 768$ and $T < 512$ by construction, on real text data from our abstract, processed by a pretrained tokenizer available on HuggingFace. We present a single realisation, though this behavior is consistently observed across multiple runs.



(a) $\mathbf{X}_0\mathbf{X}_0^\top$ when $T = 75$ (b) $\mathbf{X}_0\mathbf{X}_0^\top$ when $T = 282$ (c) $\mathbf{X}_0\mathbf{X}_0^\top$ when $T = 446$

(d) The eigenvalues in the complex plane formed by the attention matrix in the first head of layer 1 align with our proof of concept.

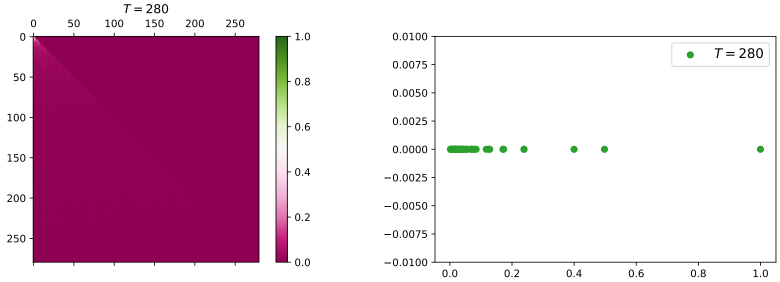
Figure II: Spectral gap (second row) persists on non-synthetic yet random data. Using a randomly initialized RoBERTa encoder ($d = 768$, $T < 512$ by construction), we substitute the word embeddings given by the pretrained tokenizer with random ones, making the input covariance matrix closer to identity (first row). We present a single realisation, though this behavior is consistently observed across multiple runs.



(a) Spectral gap in the attention matrix of the first head of layer 1.

(b) Stable rank quickly collapses in width, even when $\gamma > 1$.

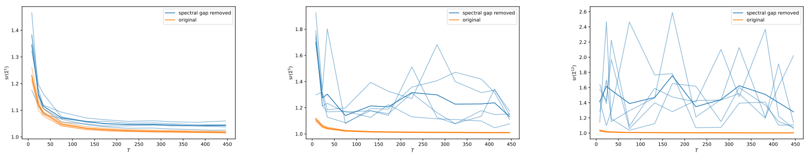
Figure III: Spectral gap (left) persists in TinyBert encoders for which $\gamma > 1$, highlighting the generality of our result, which constrained the value of γ only for the sake of having a mathematical proof. As an almost direct consequence of the spectral gap in the attention matrix, stable rank rapidly collapses (right) in width. Here $d = 312$ by design and we increase γ by feeding the network with longer sequences. Whilst the eigenvalues (left) are shown for one specific realisation (but consistently observed across runs), the stable rank (right) is averaged over 5 runs.



(a) The attention matrix is triangular due to the causal mask.

(b) The eigenvalues of this attention matrix are real and correspond to the values along its diagonal.

Figure IV: Further investigation of a GPT2 transformer, a decoder-like architecture. The causal mask imposes a spectific triangular structure on the attention matrix (left), which adds an extra layer of complexity that our analysis does not cover. Yet, it is interesting to see emerging a spectral gap (right) on the real line even in this extended setting. By design, $d = 768$ and the input consists of text from our abstract, processed using a pretrained tokenizer, resulting in non-orthogonal inputs.



(a) layer 1 (b) layer 5 (c) layer 12

Figure V: Rank collapse in width is slowed down in a real Transformer encoder (RoBERTa) with our proposed fix, for which $d = 768$. The randomly initialised model processes sentences from this abstract using a pre-trained tokenizer, making the input data non-isometric and reflective of real-world conditions. While our theoretical analysis fails to explain intricate dynamics in such model, our findings appear robust enough to empirically better preserve rank beyond more complex architectures and across depth. The experiment is averaged over 5 runs.

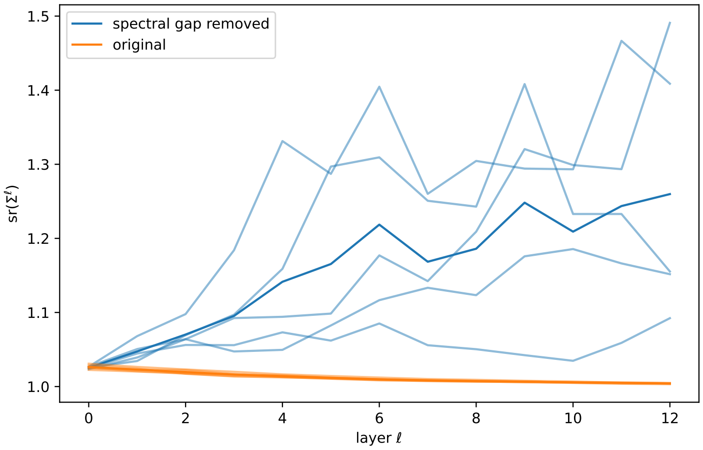


Figure VI: Rank collapse in depth is mitigated in a real Transformer encoder (RoBERTa) with our proposed fix implemented, for $d = 768$. The randomly initialised model processes sentences from this abstract using a pretrained tokenizer, making the input data non-isometric and reflective of real-world conditions, with $T = 282$. While our theoretical analysis is based solely on the attention mechanism, our findings appear robust enough to extend beyond attention, holding empirically in more complex architectures and across depth. The experiment is averaged over 5 runs.

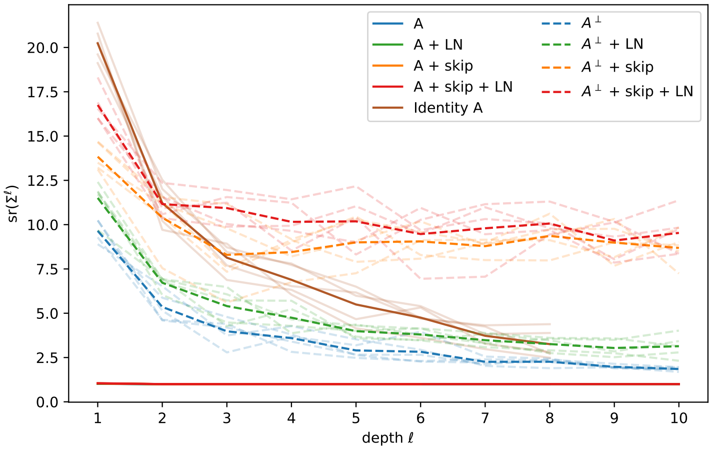


Figure VII: Analogous of Fig. 5 from our draft when consecutive iid Markov matrices are used for information propagation (independent of the input). Notably, all solid lines—except for the case labelled as 'Identity A' where attention is set to the identity (resulting in no attention)—rapidly collapse to 1 due to rank collapse in depth, regardless of additional modules like LayerNorm or skip connections. The only significant change occurs when the spectral gap is removed (denoted as '$A^\perp$'), confirming that mitigating rank collapse in width helps prevent its propagation in depth. Here, $d = T = 150$, the inputs are synthetic and orthogonal, and experiments are averaged over 5 runs.

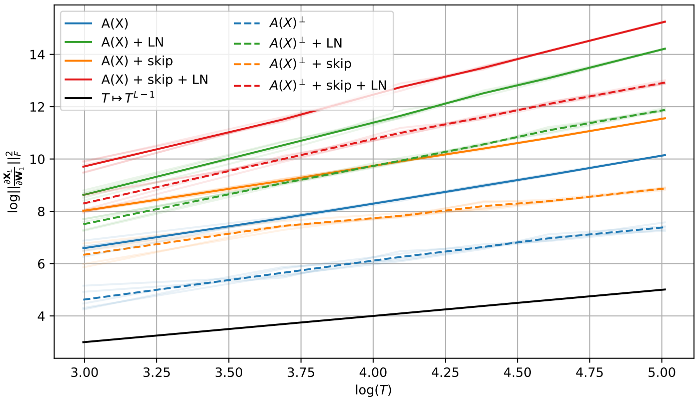
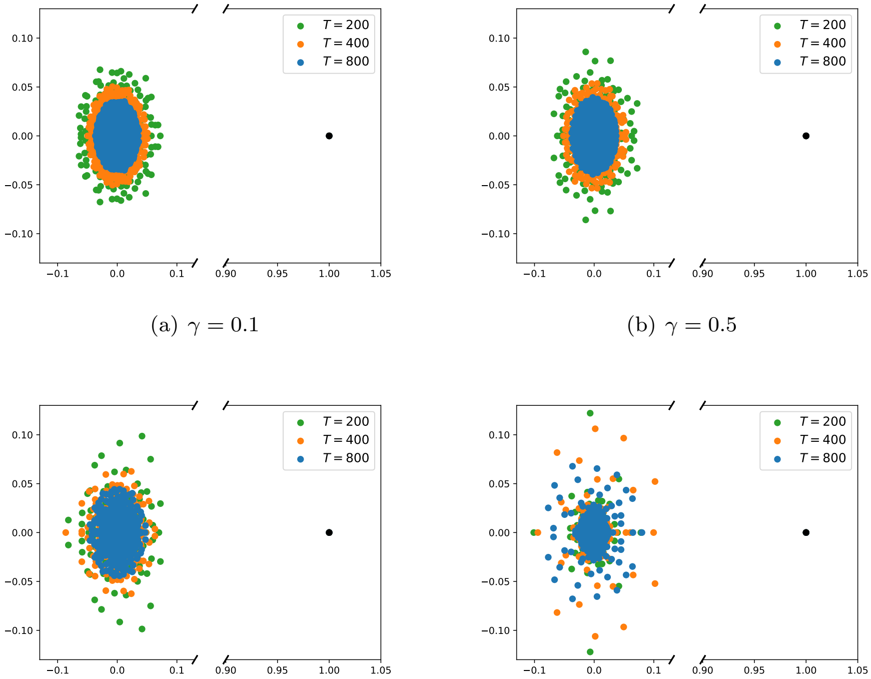


Figure VIII: Gradients already explode in magnitude at the second layer ($L = 2$). In the case of attention matrices as described in equation (1), i.e. input-dependent key-query attention, our proposed fix slows down the gradients' growth by reducing rank collapse in width and thus in depth. The complex interplay between gradients and rank collapse in deep layers presents new research directions, likely requiring a more empirical approach due to the mathematical complexity involved. Here, $d = T$, the inputs are synthetic and orthogonal, and experiments are averaged over 5 runs.



(a) $\gamma = 0.1$

(b) $\gamma = 0.5$

(c) $\gamma = 2$

(d) $\gamma = 10$

Figure IX: The eigenvalue distribution of a single attention layer for non-isometric synthetic input and different values of $\gamma = \frac{T}{d}$. The input is generated by drawing i.i.d. Gaussian entries followed by normalising the rows, so each token has unit length but they are not necessarily orthogonal. The gap between the bulk and the edge persists despite the bulk looking different for $\gamma > 1$.