Supplement: Theoretical Limitations of Self-Attention in Neural Sequence Models

Michael Hahn

May 5, 2021

Here I am providing two supplements to the published TACL paper: First, a more formal writeup of the hard attention proof. This has benefited a lot from discussions with Gail Weiss and Will Merrill. Second, I am providing a missing detail in the soft attention proof (thanks for Navin Goyal and Satwik Bhattamishra for spotting this).

S1 Results for Hard Attention

Theorem 1. *Let any hard attention transformer be given, and let $C \in (0,1)$. Then there is a restriction ρ and an integer $c > 0$ such that*

$$|\{i \leq n : \rho_n(i) = *\}| \geq Cn$$

(for all sufficiently large n) and such that the function computed by the transformer on the restricted input depends only on $\leq c$ inputs, independent of input length n.

Definition 2 (c-Transformer)**.** *Let c be a positive integer. A c-transformer is one in which the layer-0 activations $y_j^{(0)}$ depend on the embeddings not just at one position j, but are a function of the embeddings at $\leq c$ input positions:*

$$y_j^{(0)} = f_{n,j}^{inp}((v_{i_1^{j,n}}, p_{i_1^{j,n}}), \ldots, (v_{i_c^{j,n}}, p_{i_c^{j,n}})) \tag{1}$$

for some indices $i_s^{j,n} \in \{1, \ldots, n\}$ ($s = 1, \ldots, c$).

Definition 3. *We say $\rho' \succ \rho$ if, whenever $\rho'_n(i) = *$, then $\rho_n(i) = *$.*

We write ρT for the function resulting from applying ρ to T.

We write $\rho\Sigma^*$ for the set of inputs compatible with ρ.

With this technical notion, we show that we can reduce layers, iteratively removing the lowest layer until no self-attention layer is left:

Lemma 4 (Depth Reduction Lemma)**.** *Given a c-transformer T with L layers, and some restriction ρ such that*

$$|\{i \leq n : \rho_n(i) = *\}| \geq Cn \tag{2}$$

($C \in (0,1]$) for all sufficiently large n. Choose any $C' < C$.

Then there is a restriction $\rho' \succ \rho$ such that

$$|\{i \leq n : \rho'_n(i) = *\}| \geq C'n \tag{3}$$

for all sufficiently large n, and such that there is a $(c \cdot (2^c kH + 1))$-transformer T' with $L-1$ layers, for some integer k (depending on C'), where $H \geq 1$ is the number of attention heads at each layer and position, such that $\rho' T = \rho' T'$.

The lemma implies Theorem 1:

Proof of Theorem 1. The output of the transformer is determined by the last activation $y_n^{(L)}$. Apply the Depth Reduction Lemma iteratively, choosing the constants C' in the lemma appropriately, until only the zero-th layer remains. Then, after applying the resulting restriction, the final activation $y_n^{(L)}$ is now computed by $y_n^{(0)}$, which is determined by a bounded number of input bits. $\square$

S1.1 Proving the Depth Reduction Lemma

In this section, we will prove the Depth Reduction Lemma. We construct the restrictions ρ'_n separately for each n, on the basis of the given restriction ρ_n. In this process, we will only *restrict additional bits*, that is, the only case in which $\rho'_n(i)$ can be different from $\rho_n(i)$ is that $\rho'_n(i)$ may be 0 or 1 where $\rho_n(i)$ was $*$. The construction proceeds in three stages $\rho_n^{(1)}$, $\rho_n^{(2)}$, and $\rho_n^{(3)} = \rho'_n$, which all may restrict additional bits. At the end, we verify that the conclusion of the Depth Reduction Lemma is satisfied for the resulting restriction ρ'_n.

Throughout the proof, we will need a few parameters independent of n: First, we need an integer k that has to be sufficiently large for the proof to succeed, and will be fixed later in the proof. Second, we need parameters $\eta \in (0, \frac{1}{2})$, $q \in (0,1)$ and $\delta > 0$; they can be chosen as follows:

Definition 5. *Choose $\eta \in (0, \frac{1}{2})$ small, $q \in (0,1)$, and $\delta > 0$ (such that $(1+\delta)q \in (0,1)$) in such a way as to achieve*

$$(1-2\eta) \cdot (1-(1+\delta)q) = C'/C \tag{4}$$

A possible choice to satisfy this is $(1+\delta)q = \frac{1}{2}$, $2\eta = 1 - 2C'/C$.

Lemma 6 (Stage 1)**.** *There is N and a restriction $\rho^{(1)} \succ \rho$ such that*

1. *each $\rho^{(1)}$-free input bit serves as an input to at most $\leq \frac{1}{\eta} c/C$ many different layer-0 heads, when applying $\rho_n^{(1)}$.*
2. *For $n > N$,*

$$\#\{i \leq n : \rho_n^{(1)}(i) = *\} \geq (1-\eta)Cn \tag{5}$$

Proof. Assume the number of input bits feeding into more than $\frac{1}{\eta} c/C$ different layer-0 activations is $\geq \eta Cn$. Then the number of pairs of input bits and depending layer-0 activations is $> \eta Cn \cdot \frac{1}{\eta} c/C = nc$. But there are at most nc such pairs, because there are n layer-0 activations, each of which depends on $\leq c$ inputs. So the number of input bits with $> \frac{1}{\eta} c/C$ depending layer-0 heads is $\leq \eta Cn$. We can obtain $\rho_n^{(1)}$ from ρ_n by restricting these input bits to some fixed value in $\{0,1\}$ (it doesn't matter which one), and the set $\{i \leq n : \rho_n^{(1)}(i) = *\}$ still has at least $(1-\eta)Cn$ elements, for all sufficiently large n. $\square$

We write (h,i) for a layer-1 attention head h ($h = 1, \dots, H$) at position i ($i = 1, \dots, n$). Let $V_\rho(i)$ denote the possible values of $y_i^{(0)}$. As $y_i^{(0)}$ depends on $\leq c$ input bits, we have:

$$|V_\rho(i)| \leq 2^c \tag{6}$$

Definition 7. *For a restriction* ρ*, a head* (h,i)*, a value* $z \in V_\rho(i)$*, and each position* $j \in \{1,\dots,n\}$*, set*

$$A_{((h,i),z),j,\rho} := \max_{x_1 \dots x_n \in \rho\Sigma^n \,:\, y_i^{(0)} = z} f_{1,h}^{att}(z, y_j^{(0)}) \tag{7}$$

For each value $z \in V_\rho(i)$*, we rank the positions* $\{1,\dots,n\}$ *downwards by this value, obtaining a sequence (in the case of ties, we resolve as we do when computing hard attention)*

$$J_{((h,i,z),\rho} := \left(j_1^{(z)}, \dots, j_n^{(z)}\right) \tag{8}$$

For each $((h,i),z)$*, obtain the sequence*

$$1 \le i_1^{(h,i,z,\rho)} < i_2^{(h,i,z,\rho)} < \dots < i_L^{(h,i,z,\rho)} \le n \tag{9}$$

of those indices j *such that there is some* ρ*-free input* x_q *that feeds into the activation at* j *and no activation and* $j' < j$*.*

Definition 8 (Satisfaction)**.** *Let* σ *be a restriction, and* $k \in \mathbb{N}$*, and assume* $z \in V_\sigma(i)$*. We say that a pair* $((i,h),z)$ *is* (k,σ)*-**satisfied** if its function value depends on at most* $\le ck$ *many input bits when applying* ρ*.*

Lemma 9 (Satisfaction and Dependency)**.** *If* $((h,i),z)$ *is* (k,σ)*-unsatisfied, then the sequence*

$$\left(i_s^{(h,i,z,\rho)} : s = 1,\dots,L\right) \tag{10}$$

has length L *at least* $\ge k$*.*

Proof. Assume some of the layer-0 heads it (k,ρ)-depends on. The higher-ranked layer-0 heads can only have a total of $\le ck$ inputs, contradiction. □

Lemma 10 (Preservation of Satisfaction)**.** *Let* σ *be a restriction, and* $k \in \mathbb{N}$*. If* $((i,h),z)$ *is* σ*-satisfied, and* $\sigma' \succ \sigma$*, then* $((i,h),z)$ *is also* σ'*-satisfied.*

Proof. Immediate. □

Definition 11. *An unsatisfied tuple* $((h,i),z)$ (k,ρ)*-**depends** on some input* x_i *if* $\rho(i) = *$ *and* x_i *appears as an input to some* $j_r^{(h,i,z,\rho)}$ *for* $r \le i_k^{(h,i,z,\rho)}$*.*

Definition 12. *An unsatisfied tuple* $((h,i),z)$ (k,ρ)*-**depends** on some layer-0 head* j *if* $j = j_s^{(h,i,z,\rho)}$ *for some* $s \le i_k$*.*

Lemma 13. $((h,i),z)$ (k,ρ)*-depends on* x_i *iff* x_i *appears as an input to some* $j_{i_s}^{(h,i,z,\rho)}$ $(s \le i_k)$*.*
Hence, $((h,i),z)$ (k,ρ)*-depends on at most* $\le ck$ *input bits.*

Proof. From the definitions. □

Definition 14. *Two unsatisfied tuples* $((h,i),z)$*,* $((h',i'),z')$ *are* (k,ρ)*-**neighbors** if some* $j_{i_s^{(h,i,z,\rho)}}$ *for one and* $j_{i_{s'}^{(h',i',z',\rho)}}$ *for the other both* (k,ρ)*-depend on some input bit* x_l*.*

Lemma 15. *Let* ρ *be a restriction, and* $k \in \mathbb{N}$*. Assume the layer-0 head at position* j *has more than* $2^c kH$ *many* (k,ρ)*-depending* (k,ρ)*-unsatisfied tuples* $((h,i),z)$*. Then there is a restriction* $\rho' \succ \rho$*, restricting only* $\le c$ *additional inputs, such that at least* kH *many* (k,ρ)*-unsatisfied tuples* $((h,i),z)$ *become* (k,ρ')*-satisfied.*

Proof. Let ρ be a restriction, and $k \in \mathbb{N}$. Assume the layer-0 head at position j has more than $2^c kH$ many (k,ρ)-depending (k,ρ)-unsatisfied tuples $((h,i),z)$. For each (k,ρ)-depending (k,ρ)-unsatisfied tuple $((h,i),z)$, collect the value q' of $y_j^{(0)}$ ($q' \in V_\rho(j)$) resulting in $A_{((h,i),z),j,\rho}$. There are $> 2^c kH$ such tuples, but only 2^c possible values q'. So one value q of them must occur $> kH$ times, by the Pigeonhole Principle. Thus, this $q \in V_\rho(j)$ is such that

$$f_{1,h}^{att}(z,q) = A_{((h,i),z),j,\rho} \tag{11}$$

for at least $> kH$ many of these (k,ρ)-depending tuples $((h,i),z)$.

For such a tuple $((h,i),z)$, j now blocks attention on any lower-ranked elements of the ranking. The higher-ranked elements of the ranking can only depend on a total of $\leq ck$ input bits by Lemma 13. □

Definition 16 (Sequence of Restrictions)**.** *Define a (finite or infinite) sequence of restrictions $\rho^{(1)} = \sigma_1 \prec \sigma_2 \prec \dots$ as follows:*

1. $\sigma_1 := \rho^{(1)}$
2. *Let σ_i be given ($i \geq 1$). If a layer-0 head has more than $2^c kH$ many (k,σ_i)-depending (k,σ_i)–unsatisfied tuples $((h,i),z)$, fix $\leq c$ input bits to make $\geq kH$ tuples satisfied, using the preceding lemma, obtaining σ_{i+1}. Otherwise, terminate the procedure.*

Lemma 17. *There are K,N such that for all $k > K$, $n > N$, this procedure terminates with $\rho'_n \succ \rho_n^{(1)}$ such that*

1. *We have*
$$\#\{i \leq n : \rho'_n(i) = *\} \geq (1-2\eta)Cn \tag{12}$$
2. *No layer-0 head has more than $2^c kH$ many (k,ρ')-depending (k,ρ')-unsatisfied tuples $((h,i),z)$.*

Proof. Due to Lemma 10, this procedure can be iterated at most until each tuple $((h,i),z)$ is (k,σ_i)-satisfied, that is, at most

$$\frac{2^c Hn}{kH} = \frac{2^c n}{k} \tag{13}$$

times. Let U_n be the number of times this procedure is iterated ($U_n \leq \frac{2^c n}{k}$). At the end, for $n > N$,

$$\#\{i \leq n : (\sigma_U)(i) = *\} \geq (1-\eta)Cn - cU_n \geq \left((1-\eta)C - \frac{2^c c}{k}\right)n \tag{14}$$

By choosing k so large that $\frac{2^c c}{k} \leq \eta C$, we find that

$$\#\{i \leq n : (\sigma_U)_n(i) = *\} \geq (1-2\eta)Cn \tag{15}$$

for every $n > N$. For the second claim, if this were not the case, the procedure would not have terminated at ρ'_n. □

Corollary 18 (Stage 2)**.** *There is K,N such that, for each $k > K$, there is a restriction $\rho^{(2,k)} \succ \rho^{(1)}$ such that*

1. $\#\{i \leq n : \rho_n^{(2,k)}(i) = *\} \geq (1-2\eta)Cn$ *for each $n > N$*
2. *Every $(k,\rho^{(2,k)})$-unsatisfied $((h,i),z)$ has at most $f \leq \frac{2^{2c}}{\eta}c^2k^2H/C$ many $(k,\rho^{(2,k)})$-unsatisfied $(k,\rho^{(2,k)})$-neighbors.*

Proof. Let $\rho^{(2,k)}$ be as given by Lemma 17. The first assertion is immediate from that lemma. For the second assertion, by that lemma, each layer-0 head has at most $\leq 2^c kH$ many $(k,\rho^{(2)})$-depending $(k,\rho^{(2)})$-unsatisfied tuples $((h,i),z)$. Using Lemma 6 and Lemma 13, each input bit has at most $\leq \frac{2^c}{\eta}kcH/C$ many $(k,\rho^{(2)})$-depending $(k,\rho^{(2)})$-unsatisfied tuples. On the other hand, a tuple $((h,i),z)$ can $(k,\rho^{(2)})$-depend on $\leq kc$ inputs by Lemma 13. Multiplying these two bounds gives $\leq \frac{2^{2c}}{\eta}k^2c^2H/C$. □

In order to construct the third and final restriction $\rho_n^{(3)}$, we apply the "probabilistic method": We define a probability distribution over restrictions $\rho_n^{(3)}$, and show that the probability assigned to restrictions of the type we require is strictly greater than zero, showing that such a restriction exists.

Definition 19. *Let $k > K$. For each input length n, define the distribution over restrictions $\rho_n^{(3,k)} \succ \rho_n^{(2,k)}$ that independently assigns to each input position $i \in \{1,\ldots,n\}$ the symbol 1 or 0 with probability $q/2$ each ($q \in (0,1)$ from Definition 5), and $*$ with probability $1-q$. On those input bits where $\rho_n^{(2,k)}(i) \neq *$, we restrict this random restriction to agree with $\rho_n^{(2,k)}(i)$.*

Definition 20. *Let $k > K$, and consider a $(k,\rho^{(2,k)})$-unsatisfied tuple $((h,i),z)$. By Lemma 9, the sequence*

$$\left(y^{(0)}_{j_{i_s^{(z)}}} : s = 1,\ldots,L\right) \tag{16}$$

has length at least $\geq k$.

Define $X_{i,h,k}^{(z)}$ to be the event that, for this tuple, none of the k layer-0 head it depends on ($s = 1,\ldots,k$) is fixed by $\rho^{(3,k)}$ to the value

$$\arg_{q\in V_{\rho^{(2,k)}}(j_{i_s^{(h,i,z,\rho^{(2)})}})} \max f^{att}_{1,h}(z,q) \tag{17}$$

(or any element of the argmax, if multiple values achieve this attention weight).

Define $X_{0,k}$ to be the event that more than $(1+\delta)q$ of the input bits that $\rho_n^{(2,k)}$ maps to $*$ are set to $0/1$ by $\rho_n^{(3,k)}$ (where $\delta \in (0,1)$ was fixed in Definition 5).

Our goal will be to show that a nonzero amount of probability mass is assigned to restrictions ρ'_n avoiding all events. We start by individually bounding the probability of each of these events.

Lemma 21 ($X_{0,k}$ is unlikely)**.** *For any $n > N, k > K$:*

$$\mathbb{P}(X_{0,k}) \leq \exp\left(-\frac{\delta^2 q(1-2\eta)Cn}{3}\right) \tag{18}$$

Proof. Since $\rho_n^{(2,k)}$ had $\geq (1-2\eta)Cn$ unrestricted input bits for $n > N$, this follows by a Chernoff bound (Mitzenmacher and Upfal, 2017, Theorem 4.4). □

Second, we show that the probability of $X_{i,h,k}^{(z)}$ ($i = 1,2,\ldots,n$, $h = 1,\ldots,H$) decays exponentially in k.

Lemma 22 ($X_{i,h,k}^{(z)}$ is unlikely)**.** *If $((h,i),z)$ is (k,ρ)-unsatisfied, then*

$$\mathbb{P}(X_{i,h,k}^{(z)}) \leq (1-(q/2)^c)^{\frac{k}{\frac{1}{\eta}c^2/C}} \tag{19}$$

for each $i = 1,2,\ldots,n$ and $h = 1,\ldots,H$.

Proof. Let $Y^t_{i,h,z,k}$ $(t = 1,\dots,k)$ be the event that the layer-0 activation $y^{(0)}_{j_{i_t^{(h,i,z,\rho^{(2)})}}}$ is not fixed by $\rho^{(3,k)}$ to

$$\arg_{q \in V_{\rho^{(2,k)}}(j_{i_t^{(h,i,z,\rho^{(2)})}})} \max f^{att}_{1,h}(z,q) \tag{20}$$

Note that

$$X^{(z)}_{i,h} = \bigcap_{t=1}^{k} Y^t_{i,h} \tag{21}$$

We have

$$\mathbb{P}(Y^s_{i,h}) \leq 1-(q/2)^c \in (0,1) \tag{22}$$

Any $Y^s_{i,h,z}$ can be statistically dependent on at most

$$c \cdot \frac{1}{\eta} c/C = \frac{1}{\eta} c^2/C \tag{23}$$

other events $Y^{s'}_{i,h,z}$, because each $\rho^{(2,k)}$-free input bit serves as an input to at most

$$\frac{1}{\eta} c/C \tag{24}$$

layer-0 heads (Lemma 6). Therefore, there is a set of

$$\geq \frac{k}{\frac{1}{\eta}c^2/C} \tag{25}$$

independent events among these. Call these $Y^{t_1}_{i,h},\dots,Y^{\frac{k}{\frac{1}{\eta}c^2/C}}_{i,h}$. Then

$$X^{(z)}_{i,h} \subseteq \bigcap_{s=1}^{\frac{k}{\frac{1}{\eta}c^2/C}} Y^{t_s}_{i,h} \tag{26}$$

and thus

$$\mathbb{P}(X^{(z)}_{i,h}) \leq \prod_{s=1}^{\frac{k}{\frac{1}{\eta}c^2/C}} \mathbb{P}(Y^{t_s}_{i,h}) \leq (1-(q/2)^c)^{\frac{k}{\frac{1}{\eta}c^2/C}} \tag{27}$$

for each $i = 1,2,\dots,n$ and $h = 1,\dots,H$. □

Lemma 23. *There are N, K such that, for each $n > N$, $k > K$, the probability of avoiding all events*

$$\{X_{0,k}\} \cup \{X^{(z)}_{i,h,k} : ((h,i),z) \textit{ is } (k,\rho^{(2,k)})\textit{-unsatisfied}\} \tag{28}$$

is strictly greater than zero.

Proof. We apply the Lovász Local Lemma (Mitzenmacher and Upfal, 2017, Theorem 6.17). Each event $X^{(z)}_{i,h,k}$ is statistically independent of the set

$$\left\{X^{(z')}_{(j,h',k)} : (k,\rho^{(2,k)})\text{-unsatisfied tuples } (j,h',z') \text{ and } (i,h,z) \text{ are not } (k,\rho^{(2,k)})\text{-neighbors}\right\} \tag{29}$$

The complement of this set has cardinality

$$\leq f = \frac{2^{2c}}{\eta} c^2 k^2 H / C \tag{30}$$

as concluded in Corollary 18. Set $A := \frac{1}{k^2}$, $B := \frac{1}{2}$. The number of events $X_{i,h}^{(z)}$ is bounded by $2^c Hn$. By the Lovász Local Lemma, it is sufficient show the following:

$$\mathbb{P}(X_{i,h}^{(z)}) \leq A(1-B)(1-A)^f \tag{31}$$

$$\mathbb{P}(X_0) \leq B(1-A)^{2^c Hn} \tag{32}$$

The Lovász Local Lemma then guarantees that there is some input restriction $\rho_n^{(3)}$ that avoids all events $\{X_0\} \cup \{X_{i,h,k}^{(z)} : i,h,z\}$. For (31), we need

$$D \leq A^{1/k}(1-B)^{1/k}(1-A)^{f/k} \tag{33}$$

where $D = (1-(q/2)^c)^{\frac{1}{\frac{1}{\eta}c^2/C}} \in (0,1)$. For the first term on the right,

$$\lim_{k\to\infty} A^{1/k} = \lim_{k\to\infty} \exp\left(-\log(k^2)/k\right) = 1$$

Also, $\lim_{k\to\infty}(1-A)^{f/k}$ equals

$$\lim_{k\to\infty}\left(1-\frac{1}{k^2}\right)^{\frac{2^{2c}}{\eta}c^2 kH/C} = \lim_{k\to\infty}\left(1-\frac{E^2}{k^2}\right)^k = 1$$

for $E := \frac{2^{2c}}{\eta}c^2 H/C$. So, if we choose k large enough (independently of n), the RHS of (33) can be made arbitrarily close to 1, in particular, greater than D. In order to also satisfy (32), we need

$$\exp\left(-\delta^2 q(1-2\eta)C/3\right) \leq B^{1/n}(1-A)^{2^c H}$$

which holds for n, k large enough (again, choosing k independent of n). $\square$

Corollary 24. *There are K,N such that for $n > N$, $k > K$, for any $\rho_n^{(3,k)}$ provided by Lemma 23, we have*

$$|\{i \leq n : \rho_n^{(3,k)}(i) = *\}| \geq C'n$$

Proof. We have

$$|\{i \leq n : \rho_n^{(3,k)}(i) = *\}| \geq (1-2\eta)\cdot(1-(1+\delta)q)Cn$$

for all sufficiently large n. The claim follows from the choices in Definition 5. $\square$

Proof of the Depth Reduction Lemma. After applying $\rho_n^{(3,k)}$, every layer-1 head $b_{j,1,h}$ depends at most on

1. the c input bits feeding into $y_j^{(0)}$, and
2. for each $h = 1,\ldots,H$, $z \in V_{\rho^{(3,k)}}(j) \subseteq V_{\rho^{(2,k)}}(j)$ such that $((h,j),z)$ is $(k,\rho^{(2,k)})$-**satisfied**, at most $\leq ck$ input bits by the definition of "satisfied".

3. for each $h = 1, \ldots, H$, $z \in V_{\rho^{(3,k)}}(j) \subseteq V_{\rho^{(2,k)}}(j)$ such that $((h, j), z)$ is $(k, \rho^{(2,k)})$-**unsatisfied**, the input bits that the tuple k-depends on, of which there are at most $\leq ck$ by Lemma 13. (Stated differently, every tuple is $(k, \rho^{(3,k)})$-satisfied.)

Thus, each layer-1 activation $y_j^{(1)}$ only depends on $\leq c \cdot (2^c kH + 1)$ input bits.

We can thus remove layer 0, convert layer-1 activations $y_j^{(1)}$ into layer-0 activations $y_j^{(0)}$, and obtain a $(c \cdot (2^c kH + 1))$-transformer performing the same computation as before when $\rho^{(3)}$ is applied. $\square$

S2 Missing Detail in Soft Attention Proof

In the proof of Lemma 5 on Page 11, the inequality at the end of the first column has the form

$$\|b - b'\| < \sum a_w \|y_w - y'_w\| \tag{34}$$

A term is missing: the RHS should be of the form

$$\|b - b'\| < \sum a_w \|y_w - y'_w\| + \sum |a_w – a'_w| y'_w \tag{35}$$

The missing term is also small under the assumptions used in the paper.

First, y'_w is bounded because f^{att} and f^{act} are Lipschitz functions, and the positional embeddings are assumed to be bounded. These assumptions are used in the k=0 step of the proof of Lemma 5, and they are necessary for the proof to work.

Second, $\sum |a_w – a'_w|$ is also in O(1/n). The next page contains a calculation for this claim.

We want to show that

$$\sum_{u \neq i} |\hat{a}^{k,h}_{j,u} - \hat{a}^{k,h\prime}_{j,u}| = O(1/n) \tag{1}$$

To show this, we show that each term is $O(1/n^2)$.

First, note $\hat{a}^{k,h}_{j,u} \in [\frac{\exp(-2A)}{n-1}, \frac{\exp(2A)}{n-1}]$ (the upper bound is given in the paper, the lower bound is analogous).

Also, for the unnormalized attention weights, $|a^{k,h}_{j,u} - a^{k,h\prime}_{j,u}| \leq \frac{Q}{n}$ for some constant Q depending on the parameter matrices and Lipschitz constant of f^{att}.

Let's fix all indices but u, and write

$$c_u := \exp(a_u) \in [\exp(-A), \exp(A)] \tag{2}$$

$$d_u := \exp(a_u) - \exp(a'_u) \tag{3}$$

Because $|a^{k,h}_{j,u} - a^{k,h\prime}_{j,u}| \leq \frac{Q}{n}$, a_u is bounded, and $\exp(\cdot)$ is continuous, therefore $|d_u| \in O(\frac{1}{n})$.

Then

$$\hat{a}_u - \hat{a}_u = \frac{c_u}{\sum_y c_y} - \frac{c_u + d_u}{\sum_y c_y + d_y} = \frac{c_u(\sum_y c_y + d_y) - (c_u + d_u)\sum_y c_y}{\sum_y c_y(\sum_y c_y + d_y)} = \frac{c_u \sum_y d_y - d_u \sum_y c_y}{\sum_y c_y(\sum_y c_y + d_y)} \tag{4}$$

$$\leq \frac{c_u \sum_y |d_y| + \frac{C}{n}\sum_y c_y}{(\sum_y c_y)^2} \leq \frac{\exp(A)C + \frac{C}{n}\sum_y c_y}{(\sum_y c_y)^2} \tag{5}$$

(for some constant C). Considering that $c_u \geq \exp(-A)$, therefore $\sum_y c_y \geq n\exp(-A)$, and this is bounded as

$$\leq \frac{\exp(A)C + \frac{C}{n} n \exp(A)}{n^2 \exp(-2A)} = O(\frac{1}{n^2}) \tag{6}$$

Acknowledgments

Thanks for Gail Weiss, Will Merrill, Navin Goyal, and Satwik Bhattamishra for helpful discussion about the original paper.

References

Mitzenmacher, M. and Upfal, E. (2017). *Probability and Computing*. Cambridge University Press, Cambridge, 2nd edition.