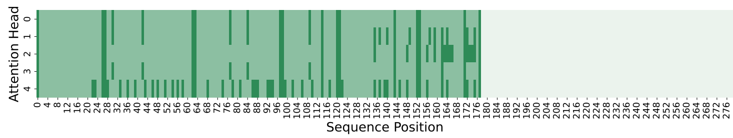
Appendix

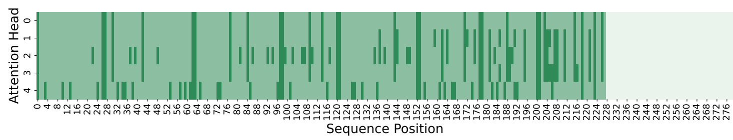
A More Observation Plots

A.1 Repetitive Attention Pattern

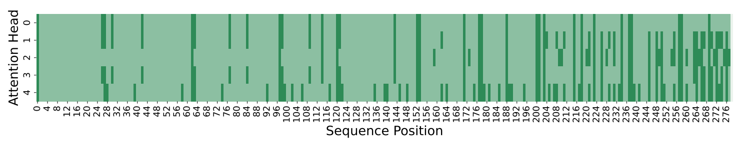
We provide the attention map similar to Figure 1 but from a different transformer layer on the same text in Figure 6, Figure 7, Figure 8 and Figure 9. A repetitive pattern and attention sparsity can be observed across layers.



(a) Attention map at position 178

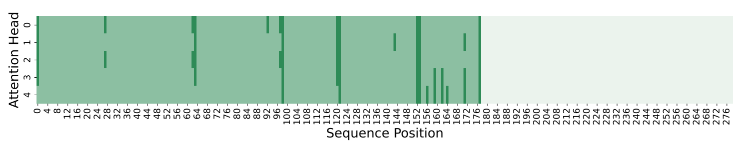


(b) Attention map at position 228

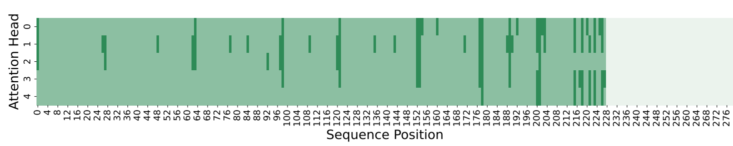


(c) Attention map at position 278

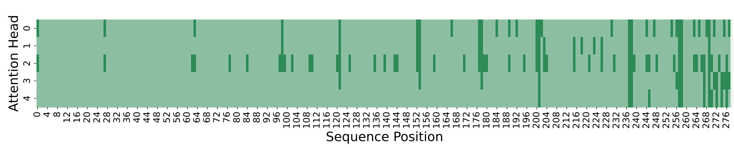
Figure 6: Attention Map at Layer 5



(a) Attention map at position 178

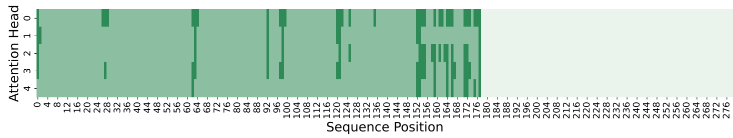


(b) Attention map at position 228

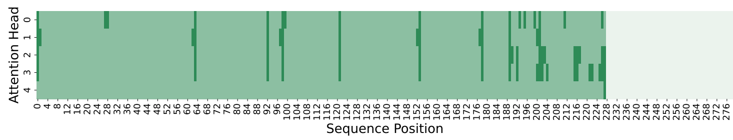


(c) Attention map at position 278

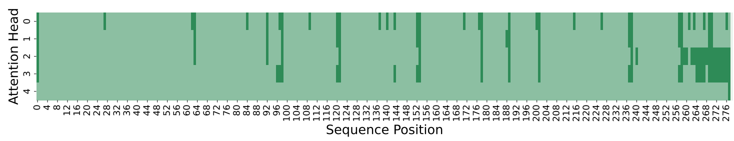
Figure 7: Attention Map at Layer 10



(a) Attention map at position 178

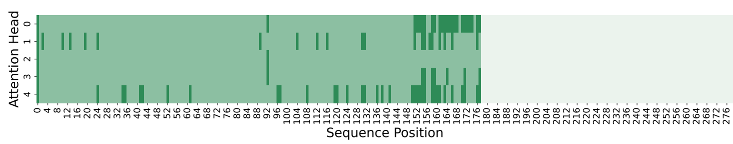


(b) Attention map at position 228

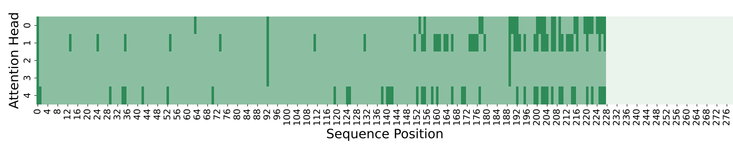


(c) Attention map at position 278

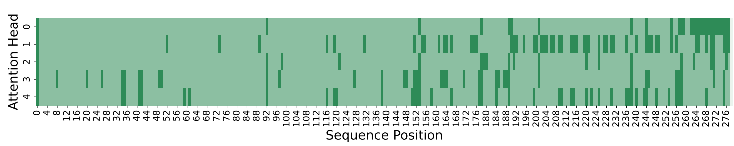
Figure 8: Attention Map at Layer 15



(a) Attention map at position 178



(b) Attention map at position 228

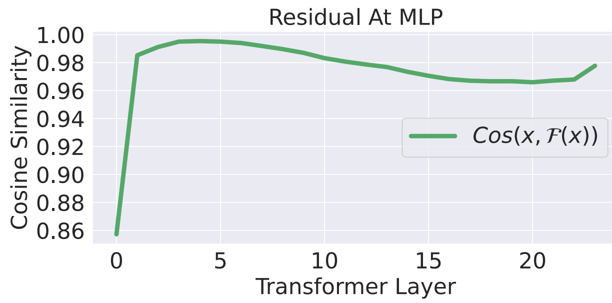


(c) Attention map at position 278

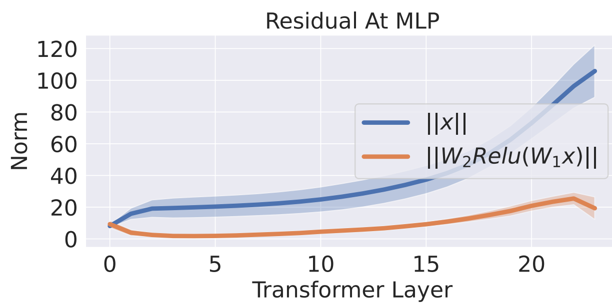
Figure 9: Attention Map at Layer 20

A.2 Cross Layer Cosine Similarity

In Section 3.3, our analysis assumes a large cosine similarity between the input and output of $\mathcal{F}$. Here, we provide empirical evidence to support such an assumption in Figure 10. Because of the residual connection in $\mathcal{F}$ and the domination of x, the cosine similarity between x and $\mathcal{F}(x)$ is extremely high.



(a) Cosine Similarity



(b) Norm Comparision

Figure 10: x and $\mathcal{F}(x)$ is high in cosine similarity

B Proofs

B.1 Proof of Theorem 3.1

We consider the token generation process of a simplified model: a single-layer transformer model with single-head attention.

$$x_{t+1} = \mathcal{F}\left(a_t\right), \text{where } a_t = \texttt{softmax}\left(1/t \cdot x_t W_Q W_K^\top X_{t-1}^\top\right) X_{t-1} W_V W_O \tag{5}$$

$x_t \in \mathbb{R}^{1\times d}$ is a row vector. $X_{t-1} \in \mathbb{R}^{(t-1)\times d}$ denotes the aggregation of $x_1, \ldots, x_{t-1}$, where the jth row is x_j. $W_Q, W_K, W_V \in \mathbb{R}^{d\times p}$ and $W_O \in \mathbb{R}^{p\times d}$ are the attention weights. Lastly, $\mathcal{F} : \mathbb{R}^{1\times d} \to \mathbb{R}^{1\times d}$ denotes the MLP block following attention block, a two-layer MLP with skip connections, given by

$$\mathcal{F}(x) = x + W_2 \texttt{relu}(W_1 x) \tag{6}$$

We are interested in the attention scores $\alpha_t = \texttt{softmax}(1/t \cdot x_t W_Q W_K^\top X_{t-1}^\top)$. Notice that $\alpha_{t,j}$ scales with $x_t W_Q W_K^\top x_j^\top$. We first re-state the Theorem 3.1 below.

Theorem B.1. *Let $A = W_V W_O W_Q W_K^\top$ and let $\lambda_K, \lambda_Q, \lambda_V, \lambda_O$ denote the largest singular values of W_K, W_Q, W_V, W_O, respectively. Consider the transformer in (5) with normalized inputs $\|x_t\|_2 = 1$ for all t. Let $c, \epsilon > 0$ be constants. Assume that $a_t x_{t+1}^\top \geq (1-\delta)\|a_t\|_2$ with $\delta \leq \left(\frac{c\epsilon}{\lambda_Q \lambda_K \lambda_V \lambda_O}\right)^2$. Then for all x_ℓ satisfying $x_\ell A x_\ell^\top \geq c$ and $x_\ell A x_\ell \geq \epsilon^{-1} \max_{j\in[t], j\neq \ell} x_j A x_\ell^\top$, it holds that*

$$\frac{x_\ell A x_\ell^\top}{\|a_t\|_2}(\alpha_{t,\ell} - 3\epsilon) \leq x_{t+1} W_Q W_K^\top x_j^\top \leq \frac{x_\ell A x_\ell^\top}{\|a_t\|_2}(\alpha_{t,\ell} + 3\epsilon) \tag{7}$$

As a preparation of the proof, we first show two lemmas.

Lemma B.1. *Let $x_1, x_2 \in \mathbb{R}^{1\times m}$ satisfies $\|x_1\|_2 = \|x_2\|_2 = 1$ and $x_1 x_2^\top \geq 1 - \delta$ for some $\delta \in (0, 1)$. Then for all $y \in \mathbb{R}^{1\times m}$ we have*

$$\left|x_1 y^\top - x_2 y^\top\right| \leq \sqrt{2\delta}\,\|y\|_2$$

Proof. Let $x_2 = x_2^{\parallel} + x_2^{\perp}$ where

$$x_2^{\parallel} = x_1 x_2^\top \cdot x_1; \quad x_2^{\perp} = x_2 - x_2^{\parallel}$$

Then it is easy to see that $x_2^{\perp} x_1^\top = 0$. By the Pythagorean Theorem, we have

$$\left\| x_2^{\perp} \right\|_2^2 = \|x_2\|_2^2 - \left\| x_2^{\parallel} \right\|_2^2 = \delta(2 - \delta)$$

Therefore, we have

$$\begin{aligned} \|x_1 - x_2\|_2^2 &= \left\| (x_1 - x_2^{\parallel}) - x_2^{\perp} \right\|_2^2 \\ &= \left\| \left(1 - x_1 x_2^\top\right) x_1 - x_2^{\perp} \right\|_2^2 \\ &= \left(1 - x_1 x_2^\top\right)^2 + \left\| x_2^{\perp} \right\|_2^2 \\ &= 2\delta \end{aligned}$$

Thus, the Cauchy-Schwarz inequality implies

$$\left| x_1 y^\top - x_2 y^\top \right| \le \|x_1 - x_2\|_2 \cdot \|y\|_2 = \sqrt{2\delta} \, \|y\|_2$$

□

Lemma B.2. *Let $\ell \in [t]$ be given. Suppose that $x_\ell A x_\ell^\top > \epsilon^{-1} \left| x_j A x_\ell^\top \right|$ for all $j \neq \ell$. Then we have*

$$\left(\mathcal{S}(t)_\ell - \epsilon\right) x_\ell^\top a x_\ell \le x_\ell^\top W_K^\top W_Q a_t \le \left(\mathcal{S}(t)_\ell + \epsilon\right) x_\ell^\top a x_\ell$$

Proof. Notice that

$$a_t = \alpha_t X_{t-1} W_V W_O = \left(\sum_{j=1}^{t-1} \alpha_{t,j} x_j \right) W_V W_O$$

Thus, we have

$$a_t W_Q W_K^\top x_\ell^\top = \left(\sum_{j=1}^{t-1} \alpha_{t,j} x_j \right) W_V W_O W_Q W_K^\top x_\ell^\top = \sum_{j=1}^{t-1} \alpha_{t,j} x_j A x_\ell^\top$$

Therefore

$$\begin{aligned} \left| a_t W_Q W_K^\top x_\ell^\top - \alpha_{t,\ell} x_\ell A x_\ell^\top \right| &= \left| \sum_{j=1, j\neq\ell}^{t-1} \alpha_{t,j} x_j A x_\ell^\top \right| \\ &\le \sum_{j=1, j\neq\ell}^{t-1} \alpha_{t,j} \left| x_j A x_\ell^\top \right| \\ &\le \epsilon x_\ell A x_\ell^\top \sum_{j=1, j\neq\ell}^{t-1} \alpha_{t,j} \\ &\le \epsilon x_\ell A x_\ell^\top \end{aligned}$$

where in the second inequality we use $\epsilon^{-1} \left| x_j A x_\ell^\top \right| \le x_\ell A x_\ell^\top$ and in the third inequality we use $\sum_{j=1, j\neq\ell}^{t-1} \alpha_{t,j} \le \sum_{j=1}^{t-1} \alpha_{t,j} = 1$. This implies that

$$\left(\alpha_{t,\ell} - \epsilon\right) x_\ell A x_\ell^\top \le a_t W_Q W_K^\top x_\ell^\top \le \left(\alpha_{t,\ell} + \epsilon\right) x_\ell A x_\ell^\top$$

□

Now we proceed to the main body of the proof. Assume that $\|x_\ell\|_2 = 1$ for all ℓ. Using Lemma (B.1), if $a_t x_{t+1}^\top \geq (1-\delta)\|a_t\|_2$, then we have

$$\left| \|a_t\|_2^{-1} a_t W_Q W_K^\top x_\ell^\top - x_{t+1} W_Q W_K^\top x_\ell^\top \right| \leq \sqrt{2\delta} \left\| W_Q W_K^\top x_\ell^\top \right\|_2$$

Recall that λ_Q, λ_K are the maximum singular values of W_Q and W_K, respectively. Then it holds that $\left\| W_Q W_K^\top x_\ell^\top \right\|_2 \leq \lambda_Q \lambda_K \|x_\ell\|_2$. Using $\|x_\ell\|_2 = 1$, we have

$$\left| \|a_t\|_2^{-1} a_t W_Q W_K^\top x_\ell^\top - x_{t+1} W_Q W_K^\top x_\ell^\top \right| \leq \sqrt{2\delta} \lambda_Q \lambda_K$$

Notice that

$$\begin{aligned}
\|a_t\|_2 &= \left\| \left(\sum_{j=1}^{t-1} \alpha_{t,j} x_j \right) W_V W_O \right\| \\
&\leq \lambda_O \lambda_V \left\| \sum_{j=1}^{t-1} \alpha_{t,j} x_j \right\|_2 \\
&\leq \lambda_O \lambda_V \sum_{j=1}^{t-1} \alpha_{t,j} \|x_j\|_2 \\
&= \lambda_O \lambda_V
\end{aligned}$$

Then since $\delta \leq \left(\frac{c\epsilon}{\lambda_Q \lambda_K \lambda_V \lambda_O} \right)^2$, we have

$$\left| \|a_t\|_2^{-1} a_t W_Q W_K^\top x_\ell^\top - x_{t+1} W_Q W_K^\top x_\ell^\top \right| \leq \frac{2c\epsilon}{\lambda_V \lambda_O} \leq \frac{2c\epsilon}{\|a_t\|_2}$$

Since by Lemma (B.2), we have

$$\left| a_t W_Q W_K^\top x_\ell^\top - \alpha_{t,\ell} x_\ell A x_\ell^\top \right| \leq \epsilon x_\ell^\top a x_\ell$$

It must hold that

$$\left| x_{t+1} W_Q W_K^\top x_\ell^\top - \|a_{t+1}\|_2^{-1} \alpha_{t,\ell} x_\ell A x_\ell^\top \right| \leq \frac{\epsilon}{\|a_t\|_2} x_\ell^\top a x_\ell + \frac{2c\epsilon}{\|a_t\|_2}$$

Since $x_\ell^\top a x_\ell \geq c$, it holds that

$$\frac{2c\epsilon}{\|a_t\|_2} \leq \frac{2\epsilon}{\|a_t\|_2} x_\ell^\top a x_\ell$$

which implies that

$$\left| x_{t+1} W_Q W_K^\top x_\ell^\top - \|a_t\|_2^{-1} \alpha_{t,\ell} x_\ell A x_\ell^\top \right| \leq \frac{3\epsilon}{\|a_t\|_2} x_\ell^\top a x_\ell$$

Therefore

$$\frac{x_\ell A x_\ell^\top}{\|a_t\|_2} (\alpha_{t,\ell} - 3\epsilon) \leq x_{t+1} W_Q W_K^\top x_\ell^\top \leq \frac{x_\ell A x_\ell^\top}{\|a_t\|_2} (\alpha_{t,\ell} + 3\epsilon)$$

B.2 Proof of Theorem 4.1

Let $\{\tilde{x}_t\}_{t=0}^T$ denote the tokens generated by the transformer with budget KV cache as in Algorithm 2 with $m = 1$:

$$\tilde{x}_{t+1} = \mathcal{F}(\tilde{a}_t), \text{ where } \tilde{a}_t = \texttt{softmax}\left({}^{1}\!/_{t} \cdot \tilde{x}_t W_Q \tilde{\mathcal{K}}_t^\top \right) \tilde{\mathcal{V}}_t^\top W_O$$

Notice that when $m = 1$, i.e., in each iteration, we drop one token with the lowest score, the cache will always maintain B tokens. If the ranking of the attention scores does not change in each iteration, Algorithm 2 will always drop tokens with the smallest attention scores.

For reference purposes, let $\{x_t\}_{t=0}^T$ denote the tokens generated by a vanilla transformer defined in (5). We re-state Theorem 4.1 below, which bounds the difference $\|x_t - \tilde{x}_t\|_2$.

Theorem B.2. *Let λ_1, λ_2 denote the largest singular values of W_1 and W_2 in (6). Let*

$$\beta_{t,j} = \frac{\exp\left(1/t \cdot \tilde{x}_t W_Q W_K^\top \tilde{x}_j^\top\right)}{\sum_{i=1}^{t-1} \exp\left(1/t \cdot \tilde{x}_t W_Q W_K^\top \tilde{x}_i^\top\right)}$$

and assume that each $\beta_{t,j} = cv_{t,j}$, where $v_{t,j}$ are sampled from a power-law distribution with pdf $f(x) = c(x+b)^{-k}$. Suppose that $\lambda_V \lambda_O (1 + \lambda_1 \lambda_2)(1 + \lambda_Q \lambda_K) \leq \frac{1}{2}$. Let $T_{\min}$ and $T_{\max}$ denote the starting and maximum sequence lengths, respectively, and let $B \leq T_{\max}$ denote the budget as in Algorithm 2. If for all $t \in [T_{\min}, T_{\max}]$, S_t contains only tokes with at most the largest B values of $\beta_{t,j}$, that is, $|S_t| = B$ and $\min_{j \in S_t} \beta_{t,j} \geq \max_{j \notin \hat{S}_t} \beta_{t,j}$, then for all $\epsilon \in (0,1)$, with probability at least $1 - T_{\max} \exp\left(-\frac{\epsilon^2 b^2 (T_{\min}-1)}{(k-2)^2(u-b)^2}\right) - T_{\max} \exp\left(-\frac{2(T_{\min}-1)(1-B/T_{\max})^2}{(1-\epsilon)^2}\right)$, the following error bound must hold for all $t \in [T_{\min}, T_{\max}]$

$$\mathbb{E}\left[\|x_t - \tilde{x}_t\|_2\right] \leq \frac{2.1(1 - B/T_{\max})}{(1-\epsilon)^2} \left(k - (k-1)\left(\frac{1-\epsilon}{B/T_{\max} - \epsilon}\right)^{1/(k-1)}\right)$$

Define $m_{k,j} = \mathbb{I}\{j \in S_t\}$. With the definition of $m_{k,j}$, $\tilde{a}_t$ can be written as

$$\tilde{a}_t = \left(\sum_{j=1}^{t-1} \tilde{\alpha}_{t,j} \tilde{x}_j\right) W_V W_O; \quad \tilde{\alpha}_{t,j} = \frac{m_{k,j} \exp\left(1/t \cdot \tilde{x}_t W_Q W_K^\top \tilde{x}_j^\top\right)}{\sum_{i=1}^{t-1} m_{k,j} \exp\left(1/t \cdot \tilde{x}_t W_Q W_K^\top \tilde{x}_i^\top\right)} \tag{8}$$

Our first lemma shows the Lipschitzness of the attention module.

Lemma B.3. *Consider two sequences of tokens $\{x_i\}_{i=1}^t$ and $\{y_i\}_{i=1}^t$ where $\|x_i\|_2 = \|y_i\|_2 = 1$ for all $i \in [t]$. Define $X_{t-1}, Y_{t-1} \in \mathbb{R}^{(t-1)\times d}$ as the matrices whose ith row are x_i and y_i, respectively. Let $\Delta_t = \|x_t - y_t\|_2$. Then we have*

$$\left\| \mathtt{softmax}\left(\frac{1}{t} x_t W_Q W_K^\top X_{t-1}^\top\right)_2 - \mathtt{softmax}\left(\frac{1}{t} y_t W_Q W_K^\top Y_{t-1}^\top\right) \right\|_2 \leq 2\frac{\sqrt{t-1}}{t} \lambda_Q \lambda_K \Delta_t$$

Proof. We can decompose the difference as

$$\begin{aligned}
&\left\| \mathtt{softmax}\left(\frac{1}{t} x_t W_Q W_K^\top X_{t-1}^\top\right) - \mathtt{softmax}\left(\frac{1}{t} y_t W_Q W_K^\top Y_{t-1}^\top\right) \right\|_2 \\
&\leq \left\| \mathtt{softmax}\left(\frac{1}{t} x_t W_Q W_K^\top X_{t-1}^\top\right) - \mathtt{softmax}\left(\frac{1}{t} x_t W_Q W_K^\top Y_{t-1}^\top\right) \right\|_2 \\
&\quad + \left\| \mathtt{softmax}\left(\frac{1}{t} x_t W_Q W_K^\top Y_{t-1}^\top\right) - \mathtt{softmax}\left(\frac{1}{t} y_t W_Q W_K^\top Y_{t-1}^\top\right) \right\|_2
\end{aligned}$$

By the Lipschitzness of softmax, we have

$$\begin{aligned}
&\left\| \mathtt{softmax}\left(\frac{1}{t} x_t W_Q W_K^\top X_{t-1}^\top\right) - \mathtt{softmax}\left(\frac{1}{t} x_t W_Q W_K^\top Y_{t-1}^\top\right) \right\|_2 \\
&\leq \frac{1}{t} \left\| x_t W_Q W_K^\top \left(X_{t-1} - Y_{t-1}\right)^\top \right\|_2 \\
&\leq \frac{1}{t} \lambda_Q \lambda_K \|x_t\|_2 \|X_{t-1} - Y_{t-1}\|_2
\end{aligned}$$

Since $\|x_t\|_2 = 1$ and $\|X_{t-1} - Y_{t-1}\|_2 = \left(\sum_{j=1}^{t-1} \|x_j - y_j\|_2\right)^{\frac{1}{2}} \leq \sqrt{t-1}\Delta_t$, we have

$$\left\| \mathtt{softmax}\left(x_t W_Q W_K^\top X_{t-1}^\top\right) - \mathtt{softmax}\left(x_t W_Q W_K^\top Y_{t-1}^\top\right) \right\|_2 \leq \frac{\sqrt{t-1}}{t} \lambda_Q \lambda_K \Delta_t$$

Similarly,

$$\begin{aligned}
&\left\| \mathtt{softmax}\left(\frac{1}{t} x_t W_Q W_K^\top Y_{t-1}^\top\right) - \mathtt{softmax}\left(\frac{1}{t} y_t W_Q W_K^\top Y_{t-1}^\top\right) \right\|_2 \\
&\leq \frac{1}{t} \left\| (x_t - y_t) W_Q W_K^\top Y_{t-1}^\top \right\|_2 \\
&\leq \frac{1}{t} \lambda_Q \lambda_K \|Y_{t-1}\|_F \|x_t - y_t\|_2
\end{aligned}$$

Since $\|x_t - y_t\|_2 = \Delta_t$ and $\|Y_{t-1}\|_2 = \sqrt{t-1}$, we have

$$\left\|\texttt{softmax}\left(\frac{1}{t}x_t W_Q W_K^\top Y_{t-1}^\top\right) - \texttt{softmax}\left(\frac{1}{t}y_t W_Q W_K^\top Y_{t-1}^\top\right)\right\|_2 \leq \frac{\sqrt{t-1}}{t}\lambda_Q\lambda_K\Delta_t$$

Combining the two bounds gives

$$\left\|\texttt{softmax}\left(\frac{1}{\sqrt{t}}x_t W_Q W_K^\top X_{t-1}^\top\right) - \texttt{softmax}\left(\frac{1}{\sqrt{t}}y_t W_Q W_K^\top Y_{t-1}^\top\right)\right\|_2 \leq 2\frac{\sqrt{t-1}}{t}\lambda_Q\lambda_K\Delta_t$$

□

Our second lemma shows the difference between the output of the sampled and vanilla transformer when the input is the same.

Lemma B.4. *Let $\tilde{a}_t$ be defined as in (8). Define b_t as*

$$b_t = \left(\sum_{j=1}^{t-1}\beta_{t,j}\tilde{x}_j\right)W_V W_O; \quad \beta_{t,j} = \frac{\exp\left({}^1/_t \cdot \tilde{x}_t W_Q W_K^\top \tilde{x}_j^\top\right)}{\sum_{i=1}^{t-1}\exp\left({}^1/_t \cdot \tilde{x}_t W_Q W_K^\top \tilde{x}_i^\top\right)} \tag{9}$$

Assume that $\|x_j\|_2 = 1$ for all $j \in [t]$. Then we have

$$\|\tilde{a}_t - b_t\|_2 \leq \lambda_V\lambda_O\sum_{j\notin\hat{S}_t}\beta_{t,j}$$

Proof. A direction computation yields

$$\tilde{a}_t - b_t = \left(\sum_{j=1}^{t-1}\left(\tilde{\alpha}_{t,j} - \beta_{t,j}\right)\tilde{x}_j\right)W_V W_O$$

Thus, $\|\tilde{a}_t - b_t\|_2$ can be bounded as

$$\|\tilde{a}_t - b_t\|_2 \leq \lambda_V\lambda_O\sum_{j=1}^{t-1}\left(\tilde{\alpha}_{t,j} - \beta_{t,j}\right)\|\tilde{x}_j\|_2 = \lambda_V\lambda_O\sum_{j=1}^{t-1}\left(\tilde{\alpha}_{t,j} - \beta_{t,j}\right)$$

since $\|\tilde{x}_j\|_2 = 1$ for all $j \in [t]$. Now we analyze $\tilde{\alpha}_{t,j} - \beta_{t,j}$. Let $\hat{S}_t = S_t \setminus \{t\}$. Then $m_{k,j} = 1$ if and only if $j \in \hat{S}_t$. For convenience, let $r_{t,j} = {}^1/_t \cdot \tilde{x}_t W_Q W_K^\top \tilde{x}_j^\top$. Thus, β can be written as

$$\beta_{t,j} = \frac{\exp\left(r_{t,j}\right)}{\sum_{i\in\hat{S}_t}\exp\left(r_{t,i}\right) + \sum_{i\notin\hat{S}_t}\exp\left(r_{t,i}\right)}$$

Furthermore, for all $j \notin \hat{S}_t$, we have $\tilde{\alpha}_{t,j} = 0$. For all $j \in \hat{S}_t$, we have

$$\tilde{\alpha}_{t,j} = \frac{\exp\left(r_{t,j}\right)}{\sum_{i\in\hat{S}_t}\exp\left(r_{t,i}\right)}$$

Therefore, for all $j \in \hat{S}_t$, we have

$$\begin{aligned}
\beta_{t,j} - \tilde{\alpha}_{t,j} &= \exp\left(r_{t,j}\right)\cdot\frac{\sum_{i\notin\hat{S}_t}\exp\left(r_{t,i}\right)}{\left(\sum_{i\in\hat{S}_t}\exp\left(r_{t,i}\right)\right)\left(\sum_{i\in\hat{S}_t}\exp\left(r_{t,i}\right) + \sum_{i\notin\hat{S}_t}\exp\left(r_{t,i}\right)\right)} \\
&= \frac{\exp\left(r_{t,j}\right)}{\sum_{i\in\hat{S}_t}\exp\left(r_{t,i}\right)}\cdot\frac{\sum_{i\notin\hat{S}_t}\exp\left(r_{t,i}\right)}{\sum_{i\in\hat{S}_t}\exp\left(r_{t,i}\right) + \sum_{i\notin\hat{S}_t}\exp\left(r_{t,i}\right)} \\
&= \tilde{\alpha}_{t,j}\sum_{i\notin\hat{S}_t}\beta_{t,j}
\end{aligned}$$

Therefore, the bound of $\|\tilde{a}_t - b_t\|_2$ can be written as

$$\|\tilde{a}_t - b_t\|_2 \leq \lambda_V\lambda_O\left(\sum_{j\in\hat{S}_t}^{t-1}\tilde{\alpha}_{t,j}\sum_{i\notin\hat{S}_t}\beta_{t,j} - \sum_{j\notin\hat{S}_t}\beta_{t,j}\right) = 2\lambda_V\lambda_O\sum_{j\notin\hat{S}_t}\beta_{t,j}$$

where the last equality follows from $\sum_{j\in\hat{S}_t}\tilde{\alpha}_{t,j} = 1$. □

Our last lemma shows the Lipschitzness of the MLP in (6).

Lemma B.5. *Let λ_1, λ_2 denote the largest singular values of W_1, W_2 in (6). For all $x_1, x_2 \in \mathbb{R}^d$, we have*

$$\|\mathcal{F}(x_1) - \mathcal{F}(x_2)\| \le (1 + \lambda_1\lambda_2)\,\|x_1 - x_2\|_2$$

Proof. Direct computation yields

$$\begin{aligned}
\|\mathcal{F}(x_1) - \mathcal{F}(x_2)\| &= \|(x_1 + W_2\texttt{relu}\,(W_1x_1)) - (x_2 + W_2\texttt{relu}\,(W_1x_2))\| \\
&\le \|x_1 - x_2\|_2 + \|W_2\texttt{relu}\,(W_1x_1) - W_2\texttt{relu}\,(W_1x_2)\| \\
&\le \|x_1 - x_2\|_2 + \lambda_2\,\|\texttt{relu}\,(W_1x_1) - \texttt{relu}\,(W_1x_2)\| \\
&\le \|x_1 - x_2\|_2 + \lambda_2\,\|W_1\,(x_1 - x_2)\|_2 \\
&\le \|x_1 - x_2\|_2 + \lambda_1\lambda_1\,\|x_1 - x_2\|_2 \\
&= (1 + \lambda_1\lambda_2)\,\|x_1 - x_2\|_2
\end{aligned}$$

where in the third inequality we use the fact that $\texttt{relu}(\cdot)$ is 1-Lipschitz. □

Now we turn to the proof of our main theorem. Combining all of the results, we have

$$\begin{aligned}
a_t - \tilde{a}_t &= \left(\sum_{j=1}^{t-1} \alpha_{t,j} x_j\right) W_V W_O - \left(\sum_{j=1}^{t-1} \tilde{\alpha}_{t,j} \tilde{x}_j\right) W_V W_O \\
&= \underbrace{\left(\sum_{j=1}^{t-1} \alpha_{t,j} x_j\right) W_V W_O - \left(\sum_{j=1}^{t-1} \alpha_{t,j} \tilde{x}_j\right) W_V W_O}_{\mathcal{T}_1} \\
&\quad + \underbrace{\left(\sum_{j=1}^{t-1} \alpha_{t,j} \tilde{x}_j\right) W_V W_O - \left(\sum_{j=1}^{t-1} \beta_{t,j} \tilde{x}_j\right) W_V W_O}_{\mathcal{T}_2} \\
&\quad + \underbrace{\left(\sum_{j=1}^{t-1} \beta_{t,j} \tilde{x}_j\right) W_V W_O - \left(\sum_{j=1}^{t-1} \tilde{\alpha}_{t,j} \tilde{x}_j\right) W_V W_O}_{\mathcal{T}_3}
\end{aligned}$$

Therefore, by triangle inequality, we have

$$\|a_t - \tilde{a}_t\|_2 \le \|\mathcal{T}_1\|_2 + \|\mathcal{T}_2\|_2 + \|\mathcal{T}_3\|_2 \tag{10}$$

To start, the magnitude of $\mathcal{T}_1$ can be bounded as

$$\begin{aligned}
\|\mathcal{T}_1\|_2 &= \left\| \left(\sum_{j=1}^{t-1} \alpha_{t,j}(x_{t,j} - \tilde{x}_{t,j})\right) W_V W_O \right\|_2 \\
&\le \lambda_V \lambda_O \left\| \sum_{j=1}^{t-1} \alpha_{t,j}(x_{t,j} - \tilde{x}_{t,j}) \right\| \\
&\le \lambda_V \lambda_O \sum_{j=1}^{t-1} \alpha_{t,j}\,\|x_{t,j} - \tilde{x}_{t,j}\|_2 \\
&\le \lambda_V \lambda_O \Delta_t \sum_{j=1}^{t-1} \alpha_{t,j} \\
&= \lambda_V \lambda_O \Delta_t
\end{aligned}$$

where in the third inequality we use $\|x_{t,j} - \tilde{x}_{t,j}\|_2 = \Delta_t$ and in the last equality we use $\sum_{j=1}^{t-1} \alpha_{t,j} = 1$. To bound the magnitude of $\mathcal{T}_2$, we apply Lemma B.3, which shows that $\|\alpha_t - \beta_t\| \le 2\frac{\sqrt{t-1}}{t}\lambda_Q\lambda_K\Delta_t$ to get that

$$
\begin{aligned}
\|\mathcal{T}_2\|_2 &= \left\| \left(\sum_{j=0}^{t-1} (\alpha_{t,j} - \beta_{t,j}) \tilde{x}_j \right) W_V W_O \right\|_2 \\
&\le \lambda_V \lambda_O \left\| \left(\sum_{j=0}^{t-1} (\alpha_{t,j} - \beta_{t,j}) \tilde{x}_j \right) \right\|_2 \\
&\le \lambda_V \lambda_O \sum_{j=0}^{t-1} |\alpha_{t,j} - \beta_{t,j}| \, \|\tilde{x}_j\|_2 \\
&\le \lambda_V \lambda_O \|\alpha_t - \beta_t\|_1 \\
&\le \sqrt{t-1}\lambda_V \lambda_O \|\alpha_t - \beta_t\|_2 \\
&\le 2\left(1 - \frac{1}{t}\right) \lambda_Q \lambda_K \lambda_V \lambda_O \Delta_t
\end{aligned}
$$

Lastly, to bound the magnitude of $\mathcal{T}_3$, we use Lemma B.4 to get that

$$
\|\mathcal{T}_3\|_2 \le 2\lambda_V \lambda_O \sum_{j \notin \hat{S}_t} \beta_{t,j}
$$

Putting things together for (10), we have

$$
\|a_t - \tilde{a}_t\|_2 \le \lambda_V \lambda_O \left(2 \sum_{j \notin \hat{S}_t} \beta_{t,j} + (2\lambda_Q \lambda_K + 1)\, \Delta_t \right)
$$

By Lemma B.5 we can further show that

$$
\|x_{t+1} - \tilde{x}_{t+1}\|_2 \le (1 + \lambda_1 \lambda_2) \lambda_V \lambda_O \left(2 \sum_{j \notin \hat{S}_t} \beta_{t,j} + (2\lambda_Q \lambda_K + 1)\, \Delta_t \right)
$$

By Theorem B.3, we have that with probability at least $1 - T_{\max} \exp\left(-\frac{\epsilon^2 b^2 (T_{\min} - 1)}{(k-2)^2 (u-b)^2}\right) - T_{\max} \exp\left(-\frac{2(T_{\min} - 1)(1 - B/T_{\max})^2}{(1-\epsilon)^2}\right)$, it holds for all $t \in [T_{\min}, T_{\max}]$ that

$$
\mathbb{E}\left[\sum_{j \notin \hat{S}_t} \beta_{t,j} \right] \le \frac{(1 - B/T_{\max})}{0.98(1-\epsilon)^2} \left(k - (k-1) \left(\frac{1-\epsilon}{B/T_{\max} - \epsilon} \right)^{\frac{1}{k-1}} \right) := \Delta_{\max}
$$

Given that $\mathbb{E}\left[\|x_t - \tilde{x}_t\|\right] \le 2\Delta_{\max}$, we have

$$
\begin{aligned}
\mathbb{E}\left[\|x_{t+1} - \tilde{x}_{t+1}\|_2\right] &\le (1 + \lambda_1 \lambda_2) \lambda_V \lambda_O \left(2\Delta_{\max} + 2\,(2\lambda_Q \lambda_K + 1)\, \Delta_{\max} \right) \\
&\le 4\lambda_V \lambda_O (1 + \lambda_1 \lambda_2)(1 + \lambda_Q \lambda_K) \Delta_{\max}
\end{aligned}
$$

Thus, as long as $\lambda_V \lambda_O (1 + \lambda_1 \lambda_2)(1 + \lambda_Q \lambda_K) \le \frac{1}{2}$, we can guarantee that

$$
\mathbb{E}\left[\|x_{t+1} - \tilde{x}_{t+1}\|_2\right] \le 2\Delta_{\max}
$$

Thus, for all $t \in [T_{\min}, T_{\max}]$, we have that

$$
\mathbb{E}\left[\|x_t - \tilde{x}_t\|_2\right] \le \frac{2.1(1 - B/T_{\max})}{(1-\epsilon)^2} \left(k - (k-1) \left(\frac{1-\epsilon}{B/T_{\max} - \epsilon} \right)^{\frac{1}{k-1}} \right)
$$

B.3 Budgeted Cache

Theorem B.3. *Let $\beta_{t,j}$ be sampled from some power-law distribution $f(x) = c(x+b)^{-\gamma}$ with support on $[0, u-b)$ for some $k > 2$ and $u \geq 5b$. Let S_t be defined in Theorem B.2, and define $\hat{S}_t = S_t \setminus \{t\}$. Then with probability at least $1 - T_{\max} \exp\left(-\frac{\epsilon^2 b^2 (T_{\min}-1)}{(k-2)^2(u-b)^2}\right) - T_{\max} \exp\left(-\frac{2(T_{\min}-1)(1-B)^2}{(1-\epsilon)^2}\right)$ it holds for all $t \in T$ that*

$$\mathbb{E}\left[\sum_{j \notin \hat{S}_t} \beta_{t,j}\right] \leq \frac{(1 - {}^B\!/\!{}_{T_{\max}})}{0.98(1-\epsilon)^2} \left(k - (k-1)\left(\frac{1-\epsilon}{{}^B\!/\!{}_{T_{\max}} - \epsilon}\right)^{\frac{1}{k-1}}\right) \tag{11}$$

We consider the case of maintaining a budget of B by dropping the smallest $\beta_{t,j}$'s. Assume that v_j has pdf $f(x) = c(x+b)^{-k}$ with support on $[0, u-b)$. To make things precise, we first compute c

$$c = \left(\int_0^{u-b} (x+b)^{-k} dx\right)^{-1} = \frac{k-1}{b^{1-k} - u^{1-k}}$$

To start, we notice that

$$\int x(x+b)^{-k} = -\frac{(x+b)^{1-k}((k-1)x+b)}{(k-1)(k-2)} := g(x)$$

Let $C = \sum_{j=1}^{t-1} v_j$, then the expectation of C is

$$\begin{aligned}
\mathbb{E}[C] = (t-1)\mathbb{E}[v_1] &= (t-1)\frac{k-1}{b^{1-k} - u^{1-k}} \int_0^{\infty} x(x+b)^{-k} dx \\
&= (t-1)\frac{k-1}{b^{1-k} - u^{1-k}} (g(u) - g(0)) \\
&= (t-1)\frac{k-1}{b^{1-k} - u^{1-k}} \left(\frac{b^{2-k}}{(k-1)(k-2)} - \frac{u^{1-k}((k-1)u - (k-2)b)}{(k-1)(k-2)}\right) \\
&= \frac{t-1}{k-2} \cdot \frac{b^{2-k} - (k-1)u^{2-k} + (k-2)bu^{1-k}}{b^{1-k} - u^{1-k}}
\end{aligned}$$

Let $\Delta = \frac{b^{2-k} - (k-1)u^{2-k} + (k-2)bu^{1-k}}{b^{1-k} - u^{1-k}}$. By Hoeffding's inequality, we have that

$$\mathbb{P}(C \leq (1-\epsilon)\mathbb{E}[C]) \leq \exp\left(-\frac{2\epsilon^2 \mathbb{E}[C]^2}{(t-1)(u-b)^2}\right)$$

This implies that with probability at least $1 - \exp\left(-\frac{2\epsilon^2 \Delta^2 (t-1)}{(k-2)^2(u-b)^2}\right)$ we have

$$C \geq (1-\epsilon)\Delta \frac{t-1}{k-2}$$

Now, we proceed to bound $\sum_{j \notin \hat{S}_t} \beta_{t,j}$ where $\hat{S}_t = \{j \in [t-1] : \beta_{t,j} \geq \frac{\gamma}{C}\}$. Equivalently, we can bound $C^{-1} \sum_{j=1}^{t-1} \mathbb{I}\{v_j \leq \gamma\} v_j$. Its expectation is given by

$$\begin{aligned}
\mathbb{E}\left[C^{-1} \sum_{j=1}^{t-1} \mathbb{I}\{v_j \leq \gamma\} v_j\right] &\leq \frac{k-2}{(t-1)\Delta(1-\epsilon)} \mathbb{E}\left[\sum_{j=1}^{t-1} \mathbb{I}\{v_j \leq \gamma\} v_j\right] \\
&= \frac{k-2}{\Delta(1-\epsilon)} \cdot \frac{k-1}{b^{1-k} - u^{1-k}} \int_0^{\gamma} x(x+b)^{-k} dx \\
&= \frac{(k-1)(k-2)}{\Delta(1-\epsilon)(b^{1-k} - u^{1-k})} (g(\gamma) - g(0))
\end{aligned}$$

We pause here and study how small can we choose γ. Notice that

$$\mathbb{E}\left[\sum_{j=1}^{t-1} \mathbb{I}\{v_j \leq \gamma\}\right] = (t-1)\mathbb{P}(v_j \leq \gamma) = (t-1) \cdot \frac{b^{1-k} - (\gamma+b)^{1-k}}{b^{1-k} - u^{1-k}}$$

By Hoeffding's inequality again, we have that

$$\mathbb{P}\left(\sum_{j=1}^{t-1}\mathbb{I}\left\{v_j\le\gamma\right\}\ge(1-\epsilon)(t-1)\cdot\frac{b^{1-k}-(\gamma+b)^{1-k}}{b^{1-k}-u^{1-k}}\right)$$
$$\le\exp\left(-\frac{2(t-1)\epsilon^2\left(b^{1-k}-(\gamma+b)^{1-k}\right)^2}{\left(b^{1-k}-u^{1-k}\right)^2}\right)$$

Enforcing $\sum_{j=1}^{t-1}\mathbb{I}\left\{v_j\le\gamma\right\}\ge T_{\max}-B$ gives $(\gamma+b)^{1-k}\le b^{1-k}-\frac{1-B/T_{\max}}{1-\epsilon}(b^{1-k}-u^{1-k})$, which can be satisfied as long as $\gamma\ge\left(\left(\frac{B/T_{\max}-\epsilon}{1-\epsilon}\right)^{\frac{1}{1-k}}-1\right)b$. Therefore

$$g(\gamma)=-\left(b^{1-k}-\frac{1-B/T_{\max}}{1-\epsilon}(b^{1-k}-u^{1-k})\right)\frac{b+(k-1)\gamma}{(k-1)(k-2)}$$

We further notice that

$$b^{1-k}-\frac{1-B/T_{\max}}{1-\epsilon}(b^{1-k}-u^{1-k})\ge\frac{B/T_{\max}-\epsilon}{1-\epsilon}(b^{1-k}-u^{1-k})$$

This gives

$$\mathbb{E}\left[C^{-1}\sum_{j=1}^{t-1}\mathbb{I}\left\{v_j\le\gamma\right\}v_j\right]\le\frac{b(1-B/T_{\max})}{\Delta(1-\epsilon)^2}-\frac{(k-1)(B/T_{\max}-\epsilon)\gamma}{\Delta(1-\epsilon)^2}$$
$$\le\frac{b(1-B/T_{\max})}{\Delta(1-\epsilon)^2}\left(k-(k-1)\left(\frac{1-\epsilon}{B/T_{\max}-\epsilon}\right)^{\frac{1}{k-1}}\right)$$

Notice that if $u\ge 5b$, we have

$$\Delta=b-(k-1)\left(\frac{u}{b}\right)^{1-k}\cdot\frac{b-u}{b^{1-k}-u^{1-k}}\le 0.98b$$

Therefore

$$\mathbb{E}\left[C^{-1}\sum_{j=1}^{t-1}\mathbb{I}\left\{v_j\le\gamma\right\}v_j\right]\le\frac{(1-B/T_{\max})}{0.98(1-\epsilon)^2}\left(k-(k-1)\left(\frac{1-\epsilon}{B/T_{\max}-\epsilon}\right)^{\frac{1}{k-1}}\right)$$

holds with probability at least $1-\exp\left(-\frac{\epsilon^2b^2(t-1)}{(k-2)^2(u-b)^2}\right)-\exp\left(-\frac{2(t-1)(1-B/T_{\max})^2}{(1-\epsilon)^2}\right)$. Taking a union bound gives the desired result.