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

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