Using Chernoff Bound in the Proof of Lemma 6 of Approximate Similarity Search Under Edit Distance using Locality-Sensitive Hashing

Lemma (Lemma 6 in [McC21])**.** *For any string x of length d,* $\Pr_\rho(\tau(x,\rho)$ *is complete*$) \geq 1 - 1/n^2$.

Proof of Lemma 6. Recall that a transcript $\tau(x,\rho)$ is complete if $|\tau(x,\rho)| < 8d/(1-p_a)+6\log n$. If the transcript contains l insert operations, $|\tau(x,\rho)| \leq d+l$ since the maximum length of a string is d.
We check the bounds for the probability of $l > 7d/(1-p_a) + 6\log n$.

Consider success to be when we do not have hash-insert operations. This behaves like a geometric progression with probability $= (1-p_a)$. The expected number of hash-inserts we need to get an operation which is not a hash-insert would be $\frac{1}{1-p_a}$. Hence,

$$E[l] = \frac{d}{1-p_a}$$

The relevant Chernoff bound is:

$$\begin{aligned} \Pr(X \geq (1+\delta)E[X]) &\leq \left(\frac{e^\delta}{(1+\delta)^(1+\delta)}\right)^{E[X]} \\ &= e^{[\delta-(1+\delta)\ln(1+\delta)]*E[X]} \qquad (1) \end{aligned}$$

We will be manipulating Equation (1) in the sequel. To calculate $\Pr(l > 7d/(1-p_a) + 6\log n)$ with $E[X] = d/(1-p_a)$, we set $\delta = 6 + \frac{6(1-p_a)\log n}{d}$. We know that $\delta - (1+\delta)\ln(1+\delta) \leq -\delta^2/3$, when $0 \leq \delta \leq 1$ [HR90]. However, our value of $\delta = 6 + \frac{6(1-p_a)\log n}{d} > 6$, hence we cannot use the above bound. Instead, we use the fact that:

$$\delta - (1+\delta)\ln(1+\delta) \leq -\delta/3$$

when $\delta > 1$.

Substituting this in Equation (1) and using the fact that $\frac{6d}{(1-p_a)} > 0$, we get:

$$\begin{aligned} \Pr(l \geq (1+\delta)E[X]) &\leq e^{-\delta E[X]/3} \\ &< e^{-(6d/(1-p_a)+6\log n)/3} \\ &< e^{-(6\log n)/3} \\ &= 1/n^2 \end{aligned} \tag{2}$$

Hence, $\Pr(l < 7d/(1-p_a) + 6\log n) > 1 - 1/n^2$

□

References

[HR90] T. Hagerup and C. Rüss. A Guided Tour of Chernoff Bounds. *Information Processing Letters*, 33:305–308, 1989/90.

[McC21] S. McCauley. Approximate similarity search under edit distance using locality-sensitive hashing. In *ICDT 2021*, pages 21:1 – 21:22, 2021.