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

Rucha Patil

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

A transcript $\tau(x,\rho)$ is complete if $|\tau(x,\rho)| < 8d/(1-p_a) + 6\log n$.
Let the transcript contain l insert operations, so $|\tau(x,\rho)| = d + l$

We calculate the probability of $l > 7d/(1-p_a) + 6\log n$.
l behaves like a Geometric Progression with common ratio $= (1-p_a)$ (As we insert only if $r_1 < p_a$).
Hence, $E[hash\ inserts\ for\ each\ character] = 1/(1-p_a)$.
$\therefore$ E[l] $= d/(1-p_a)$.

$Pr(X \geq (1+\delta)E[X]) \leq (\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}})^{E[X]}$. (From Ref mentioned in the paper)
As we are calculating $Pr(l > 7d/(1-p_a) + 6\log n)$, $7d/(1-p_a) + 6\log n = (1+\delta)E[l]$
Substituting, $E[l] = d/(1-p_a)$
we get, $7d/(1-p_a) + 6\log n = (1+\delta) * \frac{d}{1-p_a}$
$\implies \frac{7d/(1-p_a)+6\log n}{d/(1-p_a)} = 1+\delta$
$\implies 7 + \frac{6(1-p_a)\log n}{d} = 1+\delta$
$\implies 6 + \frac{6(1-p_a)\log n}{d} = \delta$ —–(1)

$\therefore Pr(l > 7d/(1-pa) + 6\log n) \leq (\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}})^{\mu}$ —–(2)

Now, $(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}})^{\mu} = (\frac{e^{6+\frac{6\log n(1-p_a)}{d}}}{(7+\frac{6\log n(1-p_a)}{d})^{(7+\frac{6\log n(1-p_a)}{d})}})^{d/(1-p_a)}$ —–from (1) —–(3)

$(7+\frac{6\log n(1-p_a)}{d})^{(7+\frac{6\log n(1-p_a)}{d})} = e^{((7+\frac{6\log n(1-p_a)}{d})ln(7+\frac{6\log n(1-p_a)}{d}))}$
Processing: $(7+\frac{6\log n(1-p_a)}{d}ln(7+\frac{6\log n(1-p_a))}{d})$
According to the paper, $d = O(n)$, so we can write $d = cn$, where c can be any positive real number.
And, $0 \leq p_a \leq 1/2$, so $1/2 \leq (1-p_a) \leq 1$
$\therefore \frac{6\log n(1-p_a)}{d} \geq 3\log n/cn$

Combine $3/c = k$

$\therefore \frac{6\log n(1-p_a)}{d} \geq \frac{k\log n}{n}$, where k is a positive real number.

$\therefore \frac{6\log n(1-p_a)}{d} > 0$

$\therefore 7 + \frac{6\log n(1-p_a)}{d} > 7$

$\therefore ln(7 + \frac{6\log n(1-p_a)}{d}) > ln(7) > 2/3$

$\therefore ((7 + \frac{6\log n(1-p_a)}{d})ln(7 + \frac{6\log n(1-p_a)}{d})) > (2/3) * (6 + \frac{6\log n(1-p_a)}{d})$

$\therefore e^{((7+\frac{6\log n(1-p_a)}{d})ln(7+\frac{6\log n(1-p_a)}{d}))} > e^{(2/3)*(6+\frac{6\log n(1-p_a)}{d})}$

$\therefore (7 + \frac{6\log n(1-p_a)}{d})^{7+\frac{6\log n(1-p_a)}{d}} > e^{(2/3)*(6+\frac{6\log n(1-p_a)}{d})}$

Substituting in (3), we get:

$(\frac{e^{6+\frac{6\log n(1-p_a)}{d}}}{7+\frac{6\log n(1-p_a)}{d}^{7+\frac{6\log n(1-p_a)}{d}}})^{d/(1-p_a)} < (\frac{e^{6+\frac{6\log n(1-p_a)}{d}}}{e^{(2/3)*(6+\frac{6\log n(1-p_a)}{d})}})^{d/(1-p_a)}$

$\therefore (\frac{e^{6+\frac{6\log n(1-p_a)}{d}}}{7+\frac{6\log n(1-p_a)}{d}^{7+\frac{6\log n(1-p_a)}{d}}})^{d/(1-p_a)} < (e^{6+\frac{6\log n(1-p_a)}{d}-(2/3)*(6+\frac{6\log n(1-p_a)}{d})})^{d/(1-p_a)}$

$\therefore (\frac{e^{6+\frac{6\log n(1-p_a)}{d}}}{7+\frac{6\log n(1-p_a)}{d}^{7+\frac{6\log n(1-p_a)}{d}}})^{d/(1-p_a)} < (e^{\frac{6+\frac{6\log n(1-p_a)}{d}}{3}})^{d/(1-p_a)}$

$\therefore (\frac{e^{6+\frac{6\log n(1-p_a)}{d}}}{7+\frac{6\log n(1-p_a)}{d}^{7+\frac{6\log n(1-p_a)}{d}}})^{d/(1-p_a)} < e^{\frac{6d/(1-p_a)+6\log n}{3}}$

From (2):

$(\frac{e^{\delta}}{(1+\delta)^(1+\delta)})^{\mu} < e^{\frac{6d/(1-p_a)+6\log n}{3}}$

$\therefore Pr(l > 7d/(1-pa) + 6\log n) < e^{\frac{6d/(1-p_a)+6\log n}{3}}$

$\therefore Pr(l > 7d/(1-pa) + 6\log n) < e^{\frac{2d}{1-p_a}+2\log n}$

The paper mentions $\therefore Pr(l > 7d/(1-pa)+6\log n) < e^{\frac{6d(1-p_a)+6\log n}{3}} < 1/n^2$.