A PROOF

Lemma 3.1

For $1 \leq h < H_c$, The $S_c[h]$ will be compacted in line 4 of Algorithm 1.

When $h = 1$, we have $|S_c[1]| = n_c$ before line 4 and $|S_c[1]| = n_c \mod 2$ after compaction in line 4. After that, we have $|S_c[2]| = \lfloor n_c/2 \rfloor$ before compaction and $|S_c[2]| = \lfloor n_c/2 \rfloor \mod 2$ after compaction. Similarly, we have $|S_c[h]| = \lfloor n_c/s^(h-1) \rfloor \mod 2$ for $h < H_c$.

Since the top compactor $S_c[H_c]$ is not compacted, its size is $|S_c[H_c]| = \lfloor n_c/s^{H_c-1} \rfloor$.

Proposition 3.3

For any KLL sketch S and any value x, $R(x, S) - R(x) = \sum_{h=1}^{h=H} \sum_{i=1}^{m_h} \omega_h \cdot X_{i,h}$, where $X_{i,h} \perp\!\!\!\perp X_{i',h'}$ if $i \neq i'$ or $h \neq h'$. For any $X_{i,h}$, there are $Pr[X_{i,h} = 0] = \frac{1}{2}, Pr[X_{i,h} = -1] = Pr[X_{i,h} = 1] = \frac{1}{4}$. Informally, that means different compaction are independent. Any one compaction at level h may affect the result of $R(x, S)$: 50% probability of unchanged, 25% of an increase of ω_h, and 25% of a decrease of ω_h. This was already shown in KLL paper.

Let s_k denotes $s_k = \sum_{i=1}^{i=k} X_{i,h}$, its easy to examine that $Pr[s_k = v] = \binom{2k+1}{k+v+1}/2^{2k+1}$ for $-k \leq v \leq k$.

Thus we have, for any $k \geq 1$, $Pr[s_k = 0] \geq Pr[s_1 = 0]$, $Pr[|s_k| \geq 1] \geq Pr[|s_1| \geq 1] = Pr[|s_1| = 1]$ and $Pr[|s_k| \geq v] \geq Pr[|s_1| \geq v] = 0 \; for \; v \geq 2$.

Then for the optimal chunk sketch S and any other chunk sketch S'. Since compaction number $m_h = 1$ and $m'_h \geq 1$, we have $Pr[\sum_{i=1}^{m_h} \omega_h \cdot X_{i,h} \geq \epsilon n_c] \leq Pr[\sum_{i=1}^{m'_h} \omega_h \cdot X_{i,h} \geq \epsilon n_c]$ for every level h. Add all levels and we have the proposition.

Lemma 3.4

Imaging a worst case: We have $H = H_c, |S_c[h]| = 1, |S[h]| = K\gamma^{H-h}$ for $1 \leq h \leq H_c - 1$.

We first compact the compactor at level H to increase it 1. Let $|S[H_c]| = 1$ after compaction.

Now we check whether the bottom $H_c - 1$ levels will trigger a compaction at level H_c again or not.

After the following compaction in bottom $H_c - 1$ levels, there would be $|S[H_c]| \leq 1 + \sum_{i=1}^{i=H_c-1} 2^{-i}(1 + |S[H_c - i]|) = 1 + \sum_{i=1}^{i=H_c-1} 2^{-i}(1 + K\gamma^i) \leq 2 + K * \sum_{i=1}^{i=+\infty} (\gamma/2)^i = 2 + K * \frac{\gamma}{2-\gamma]}$

When $K \geq 10$, there will be $|S[H_c]| \leq 2 + K * \frac{\gamma}{2-\gamma]} \leq K\gamma = k_{H_c}$ since $1/2 < \gamma < 1$. So there won't be another compaction at level H_c. So the H will increase at most 1.

Proposition 3.6

According to Hoeffding's inequality, we only need to compare $\sum_{i=1}^{H} 2^{2(i-1)} m_i$ and $\sum_{i=1}^{H} 2^{2(i-1)} m'_i$.

Let S and S' have the same height H. Let $m_h = m'_h$ for $h >= H_c$ since compaction of the two methods mainly differ in the bottom $H_c - 1$ levels.

Now let's compute the compaction number of bottom $H_c - 1$ levels for both methods.

For our proposed merging method, we have $m_h = n/n_s$ for $1 <= h <= H_c - 1$. That's because there are n/n_s chunk sketches to merge, each has compaction 1 in every level. $\sum_{i=1}^{H_c-1} 2^{2(i-1)}(n/n_s) = \frac{n}{n_s} \frac{4^{H_c-1}-1}{3}$

For the original method in data stream, we let $m'_h = n/(k_h * 2^{h-1}) = 2n(\gamma/2)^h/(K\gamma^H)$. Then we have $\sum_{h=1}^{H_c-1} 2^{2(h-1)} m'_h = \frac{n/2}{K\gamma^H} \sum_{h=1}^{H_c-1} (2\gamma)^h = \frac{n/2}{K\gamma^H} \cdot \frac{(2\gamma)^{H_c}-2\gamma}{2\gamma-1}$

Since we want $\sum_{i=1}^{H} 2^{2(i-1)} m_i \leq \sum_{i=1}^{H} 2^{2(i-1)} m_i'$, there should be $K\gamma^H \leq \frac{3}{2} \frac{(2\gamma)^{H_c} - 2\gamma}{(2\gamma-1)(4^{H_c-1}-1)} n_s$

Let $F(\gamma, H_c) = \frac{3}{2} \frac{(2\gamma)^{H_c} - 2\gamma}{(2\gamma-1)(4^{H_c-1}-1)}$, then we want $\gamma^H \leq F(\gamma, H_c) n_s / K$, in other words H should be large enough since $0.5 < \gamma < 1$ and $F(\gamma, H_c) n_s / K$ doesn't change with n.

H will increase as n grows. Now we find a lower bound of n under certain H. A sketch with $H-1$ levels can summarize at most $\sum_{i-1}^{H-1} 2^{i-1} K \gamma^{H-1-i} = K\gamma^{H-2} \sum_{i=0}^{i=H-2} (2/\gamma)^i <$. Then we have $n > K2^H/(4-2\gamma)$ when there are H levels.

When $\gamma^H \leq F(\gamma, H_c) n_s / K \iff 2^H \geq (F(\gamma, H_c) n_s / K)^{\log_\gamma 2}$, there is $n \geq \frac{K}{4-2\gamma} 2^H \geq \frac{K}{4-2\gamma} (F(\gamma, H_c) n_s / K)^{\log_\gamma 2} = K^{1-\log_\gamma 2} \cdot n_c^{\log_\gamma 2} \cdot F(\gamma, H_c)$. That is our proposition.