

Full Proof

Here are proofs about space complexity to achieve ϵ -approximation with $1 - \delta$ probability.

The target is to guarantee $\Pr[|ERR| > \epsilon N] \leq \delta$

With Chernoff bound: $\Pr[|ERR| > t] \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right)$, for sub-Gaussian ERR with a variance of σ^2 .

Disk-Bounded Chunk Sketches

The total variance is $\sigma^2 = \frac{N}{c} \sum_{h=1}^{H_c-1} 4^{h-1} \cdot 1 \leq \frac{N}{3c} 4^{H_c-1}$

To achieve $2 \exp\left(-\frac{\epsilon^2 N^2}{2\sigma^2}\right) \leq \delta$, there should be $-\frac{\epsilon^2 N^2}{2\sigma^2} \leq \log \frac{\delta}{2}$. Thus $-\epsilon^2 N^2 \leq 2\sigma^2 \log \frac{\delta}{2} \rightarrow \epsilon^2 N^2 \geq 2\sigma^2 \log \frac{2}{\delta} \rightarrow \sigma^2 \leq \epsilon^2 N^2 / \log \frac{2}{\delta} / 2$

When H_c is small enough, there will be $\frac{N}{3c} 4^{H_c-1} \leq \epsilon^2 N^2 / \log \frac{2}{\delta} / 2$ satisfying

above. Thus $4^{H_c} \leq \epsilon^2 6cN \log \frac{\delta}{2}$, i.e., $2^{H_c} \leq \sqrt{6cN} \epsilon \sqrt{\log \frac{\delta}{2}}$

Thus, the total size of chunk sketches $\leq \frac{N}{2^{H_c}} = \sqrt{\frac{N}{6c}} \cdot \frac{1}{\epsilon} \sqrt{\log \frac{2}{\delta}}$

Disk-Bounded SSTable Sketches

When $\sqrt{T} < T_s < T$, the variance of error of SSTable sketches is at most

$\left(1 + \frac{4^{H_c}}{T - \left(\frac{T}{T_s}\right)^2}\right)$ times of concatenated chunk sketches.

Now the target is to satisfy $\left(1 + \frac{4^{H_c}}{T - \left(\frac{T}{T_s}\right)^2}\right) \cdot \frac{N}{3C} 4^{H_c-1} \leq \epsilon^2 N^2 / \log \frac{2}{\delta} / 2$

$$\left(1 + \frac{4^{H_c}}{T - \left(\frac{T}{T_s}\right)^2}\right) \cdot 4^{H_c} \leq \epsilon^2 6CN \log \frac{\delta}{2}$$

At level L, the size of top sketch is $\frac{N_c}{2^{H_c}} \cdot T_s^L$. Then the size of largest top sketch is $O(N^{\log_T T_s} / 2^{H_c})$.

Let $T_s = T^{\frac{1}{2}(1+b)}$, $0 < b < 1$. Then $T/T_s^2 = T^{-b}$, $\log_T T_s = \frac{1}{2}(1+b)$, $\frac{4^{H_c}}{(1-T/T_s^2)T} = \frac{4^{H_c}}{(1-T^{-b})T}$

$$4^{H_c} \cdot 4^{H_c} = 2^{4H_c} = O\left(\epsilon^2 CNT(1 - T^{-b}) \log \frac{\delta}{2}\right)$$

$$2^{4H_c} = O\left(\epsilon^2 CN \log \frac{\delta}{2}\right)$$

The space complexity is $O\left(\frac{N^{\log_T T_s}}{2^{H_c}}\right)$

$$\begin{aligned} &= O\left((CT)^{-\frac{1}{4}} \cdot N^{\left(\frac{1}{4} + \frac{1}{2}b\right)} \cdot \left(\frac{(T^b - 1)}{T^b}\right)^{-\frac{1}{4}} \cdot \epsilon^{-\frac{1}{2}} \cdot \left(\log \frac{1}{\delta}\right)^{\frac{1}{4}}\right) \\ &= O\left((CT)^{-\frac{1}{4}} \cdot N^{\left(\frac{1}{4} + \frac{1}{2}b\right)} \cdot \epsilon^{-\frac{1}{2}} \cdot \left(\log \frac{1}{\delta}\right)^{\frac{1}{4}}\right) \end{aligned}$$

Memory-Constrained Chunk Sketches

To make the merge-and-compressed chunk sketches more accurate than streaming KLL, in other words, $\sum_{i=1}^{H_c-1} 2^{2(i-1)} \frac{n}{n_c} \leq \sum_{i=1}^{H_c-1} m_i \omega_i^2$.

That requires $K\gamma^H \leq \frac{3}{2} \frac{(2\gamma)^{H_c-2\gamma}}{(2\gamma-1)(4^{H_c-1}-1)} n_c$

$$K\gamma^H \leq \Gamma \frac{\gamma^{H_c}}{2^{H_c}} n_c$$

For any top capacity K in the streaming KLL,

$$\gamma^H \left(\frac{2}{\gamma}\right)^{H_c} \leq \frac{\Gamma n_c}{K}$$

$$\left(\frac{2}{\gamma}\right)^{H_c} \leq \frac{\Gamma n_c}{\gamma^H K}$$

Note that $(AB)^C = (A^C)^{\log_A AB}$

$$\left(\frac{2}{\gamma}\right)^{H_c} = (2^{H_c})^{\log_2 2/\gamma} \leq \frac{\Gamma n_c}{\gamma^H K}, \quad \log_2 2/\gamma \in (1,2)$$

I/O cost is

$$\begin{aligned} \frac{n}{2^{H_c}} &= O\left(\frac{n}{\left(\frac{\Gamma n_c}{\gamma^H K}\right)^{1/\log_2 \frac{2}{\gamma}}}\right) = O\left(\frac{n}{\left(\frac{\Gamma n_c}{\gamma^H K}\right)^{1/\log_2 \frac{2}{\gamma}}}\right) = O\left(\frac{N}{\left(N_c \left(\frac{1}{\gamma}\right)^{\log N}\right)^{1/\log_2 \frac{2}{\gamma}}}\right) \\ &= O\left(\frac{N}{\left(N_c (N)^{\log \frac{1}{\gamma}}\right)^{1/\log_2 \frac{2}{\gamma}}}\right) = O\left(\frac{N^{1 - \frac{\log \frac{1}{\gamma}}{\log_2 \frac{2}{\gamma}}}}{N_c^{1/\log_2 \frac{2}{\gamma}}}\right) = O\left(\frac{N^{1 - \frac{\log \frac{1}{\gamma}}{1 + \log \frac{1}{\gamma}}}}{N_c^{1/\log_2 \frac{2}{\gamma}}}\right) \\ &= O\left(\frac{N^{\frac{1}{1 + \log \frac{1}{\gamma}}}}{N_c^{1/(1 + \log \frac{1}{\gamma})}}\right) = O\left(\left(\frac{N}{N_c}\right)^{1/(1 + \log \frac{1}{\gamma})}\right) \end{aligned}$$

When $\gamma = 2/3$, that is $O\left(\left(\frac{N}{N_c}\right)^{0.631}\right)$

Memory-Constrained SSTable Sketches

The height of the top sketch in level L SSTable is $H_c + L \cdot \log \frac{T}{T_s}$

The target is to make the merge-and-compacted SSTable sketches more accurate than streaming KLL, invoke the lemmas bounding SSTable sketch error with Chunk sketch error and we have:

$$\left(1 + \frac{4^{H_c}}{T - \left(\frac{T}{T_s}\right)^2}\right) \cdot \sum_{i=1}^{H_c-1} 2^{2(i-1)} \frac{N}{N_c} \leq \sum_{i=1}^{H_c-1+L \cdot \log \frac{T}{T_s}} m'_i \omega_i^2$$

Let $T_s = T^{\frac{1}{2}(1+b)}$, $0 < b < 1$. Then $T/T_s^2 = T^{-b}$, $\log_T T_s = \frac{1}{2}(1+b)$

The target is to satisfy

$$\begin{aligned} \left(1 + \frac{4^{H_c}}{T - \left(\frac{T}{T_s}\right)^2}\right) \cdot \frac{K\gamma^H}{2^{2 \cdot \frac{L}{2}(1+b)}} &\leq \Gamma \frac{\gamma^{H_c + \frac{L}{2}(1+b)}}{2^{H_c + \frac{L}{2}(1+b)}} n_c \\ \left(1 + \frac{4^{H_c}}{T - \left(\frac{T}{T_s}\right)^2}\right) \cdot \frac{K\gamma^H}{2^{\frac{L}{2}(1+b)}} &\leq \Gamma \frac{\gamma^{H_c + \frac{L}{2}(1+b)}}{2^{H_c}} n_c \\ \left(1 + \frac{4^{H_c}}{T - \left(\frac{T}{T_s}\right)^2}\right) \cdot K\gamma^H &\leq \Gamma \left(\frac{\gamma}{2}\right)^{H_c} (2\gamma)^{\frac{L}{2}(1+b)} n_c \end{aligned}$$

Recall that $\frac{4^{H_c}}{(1-T/T_s^2)T} = \frac{4^{H_c}}{(1-T^{-b})T}$

$$\frac{4^{H_c}}{(1-T^{-b})T} \cdot K\gamma^H \leq \Gamma \left(\frac{\gamma}{2}\right)^{H_c} (2\gamma)^{\frac{L}{2}(1+b)} n_c$$

Again, the I/O cost is $O(N^{\log_T T_s} / 2^{H_c})$. However, L is determined by N and we need further analyze.

$$n_c \cdot T^L = n; \quad 2^L = (T * 2/T)^L = (T^L)^{\log_T 2}$$

$$(2\gamma)^{\frac{L}{2}(1+b)} = ((T^L)^{\log_T(2\gamma)})^{\frac{1}{2}(1+b)} = \left(\frac{n}{n_c}\right)^{\frac{1}{2}(1+b) \log_T(2\gamma)}$$

Now the target is to satisfy

$$\begin{aligned}
\left(\frac{8}{\gamma}\right)^{H_c} &\leq \frac{\Gamma(1 - T^{-b})T(2\gamma)^{\frac{L}{2}(1+b)}n_c}{K\gamma^H} \\
\left(\frac{8}{\gamma}\right)^{H_c} &= (2^{H_c})^{\log_2 8/\gamma} \leq \frac{(1 - T^{-b})T\left(\frac{n}{n_c}\right)^{\frac{1}{2}(1+b)\log_T(2\gamma)}n_c}{K\gamma^H} \\
2^{H_c} &\leq \left(\frac{(1 - T^{-b})Tn_c}{K\gamma^H}\right)^{\frac{1}{\log_2 \frac{8}{\gamma}}} \left(\frac{n}{n_c}\right)^{\frac{1(1+b)}{2\log_2 \gamma}\log_T(2\gamma)} \\
&= \left(\frac{(1 - T^{-b})Tn_c}{K\gamma^H}\right)^{1/\log_2 \frac{8}{\gamma}} \left(\frac{n}{n_c}\right)^{\frac{(1+b)\log_T(2\gamma)}{2 \cdot \frac{1}{3+\log_2 \frac{1}{\gamma}}}}
\end{aligned}$$

Now H_c is bounded and the I/O cost is

$$\begin{aligned}
O(N^{\log_T T_s}/2^{H_c}) &= O\left(\frac{N^{(1+b)/2}}{2^{H_c}}\right) \\
&= O\left(\frac{N^{\frac{(1+b)}{2}}}{\left((1 - T^{-b})Tn_c(N)^{\log \frac{1}{\gamma}}\right)^{\frac{1}{\log_2 \frac{8}{\gamma}}} \left(\frac{n}{n_c}\right)^{\frac{1+b}{2\log T} \cdot \frac{\log 2\gamma}{\log \frac{8}{\gamma}}}}}\right) \\
&= O\left(\frac{N^{\frac{(1+b)}{2} - \frac{(1+b)}{2} \cdot \frac{1}{\log T} \cdot \frac{\log 2\gamma}{\log \frac{8}{\gamma}} - \log \frac{1}{\gamma} \cdot \frac{1}{\log \frac{8}{\gamma}}}}{(N_c)^{\frac{1}{\log \frac{8}{\gamma}} \left(1 - \frac{1+b}{2\log T} \cdot \log 2\gamma\right)}}}\right) \\
&= O\left(\frac{N^{\frac{(1+b)}{2} - \frac{(1+b)}{2} \cdot \frac{1}{\log T} \cdot \frac{\log 2\gamma}{\log \frac{8}{\gamma}} - \log \frac{1}{\gamma} \cdot \frac{1}{\log \frac{8}{\gamma}}}}{(N_c)^{\frac{1}{\log \frac{8}{\gamma}} \left(1 - \frac{1+b}{2\log T} \cdot \log 2\gamma\right)}}}\right)
\end{aligned}$$

When $\gamma = \frac{2}{3}$, $T=10$, $T_s=5$, $b=0.3979$, it is $O\left(\frac{N^{0.5114}}{N_c^{0.32}}\right)$