HeavyLocker Math

January 9, 2025

A MATHEMATICAL ANALYSIS

A.1 Proof of Theorem 4.1

Theorem 4.1 Assuming we use a single hash h_1, and the sketch has inserted s distinct items, let p_1 be the probability of a collision when we insert a new item, we have:

$$p_1 \approx \frac{1}{\sqrt{2\pi}\beta} e^{-\frac{\beta^2}{2}}$$

where $\beta = \frac{d-\frac{s}{w}}{\sqrt{\frac{s}{w}(1-\frac{1}{w})}}$. Assume the new item is to be inserted into the i-th bucket, then a collision occurring means that among the previous s items, at least d items have already been mapped to the i-th bucket. Consider the 0-1 variable X_j, for $j = 1, 2, \ldots, s$, where $X_j = 1$ if and only if the j-th item maps to the i-th bucket. Let $X = \sum_{j=1}^{s} X_j$, then:

$$E(X_j) = \frac{1}{w} \qquad Var(X_j) = \frac{1}{w}(1 - \frac{1}{w})$$

Since $X_1, X_2, \ldots, X_n$ are i.i.d., let:

$$\mu = E(X) = \frac{s}{w} \qquad \sigma^2 = Var(X) = \frac{s}{w}(1 - \frac{1}{w})$$

The Central Limit Theorem states that as $s \to \infty$, $\frac{X-\mu}{\sigma}$approaches a standard normal distribution $N(0, 1)$, we have:

$$p_1 \leq \Pr[|X - \mu| \geq d - \mu] \approx \frac{1}{\sqrt{2\pi}} \int_{\beta}^{\infty} e^{-\frac{1}{2}t^2} dt \approx \frac{1}{\sqrt{2\pi}\beta} e^{-\frac{\beta^2}{2}}$$

A.2 Proof of Theorem 4.2

Theorem 4.2 Assume we use k independent hash functions $h_1, \cdots, h_k$, and the sketch already contains s distinct items. Let p_k be the probability of a collision when we insert a new item, we have:

$$p_k \approx p_1^k$$

We ignore the impact of discarded items. Since the probability that any item f ultimately selects the i-th bucket remains $\frac{1}{w}$, when using multiple hashes, the probability that any given bucket is filled is the same as with a single hash. Therefore, when an item arrives, a hash collision means that all k buckets are full, and thus we have $p_k \approx p_1^k$. So, we can find that the hash collision probability decreases exponentially with the number of hash functions, proving the effectiveness of multi-hashing in Section 3.4 Optimization-2.

A.3 Proof of Theorem 4.3

Theorem 4.3 Assuming a heavy hitter e_1 is overestimated, its actual size is f_1, and HeavyLocker provides an estimated size $\hat{f_1}$. We have:

$$\Pr[\hat{f_1} \leq f_1 + \epsilon] \geq 1 - \sqrt{\frac{p(1-p)}{8\pi(p-\theta)\epsilon}} e^{-2\frac{(p-\theta)\epsilon}{p(1-p)}}$$

where $p = \frac{f_1}{n}$. Assume that the last time e_1 replaced another item, a total of n_0 items had arrived. Consider the random variable $X = \sum_{i=1}^{n_0} X_i$, where $X_i = 1$ indicates that the i-th item is e_1. We know that for all i, X_i follows a Bernoulli distribution with probability $\frac{f_1}{n}$. Let $p = \frac{f_1}{n}$, therefore, X is distributed as $\mathcal{B}(n_0, p)$. So we have:

$$\mu = E(x) = n_0 p \quad \sigma = Var(X) = n_0 p(1-p)$$

According to the Central Limit Theorem, when n_0 is sufficiently large, $\frac{X-\mu}{\sigma}$ approximates a standard normal distribution, $\mathcal{N}(0, 1)$.
Now consider the item replaced by e_1, suppose its size is f_2. Since the bucket was not locked, we derive that $f_2 < \theta n_0$. So we have: $\Pr[\hat{f_1} \geq f_1 + \epsilon] = \Pr[X \leq f_2 - \epsilon]$
$< \Pr[X \leq \theta n_0 - \epsilon] = \frac{1}{\sqrt{2\pi}\beta} e^{-\frac{\beta^2}{2}}$ where $\beta = \frac{pn_0 - \theta n_0 + \epsilon}{\sqrt{n_0 p(1-p)}}$. Subsequently, we have: $\beta = \frac{1}{\sqrt{p(1-p)}}((p-\theta)\sqrt{n_0} + \frac{\epsilon}{\sqrt{n_0}}) \geq 2\sqrt{\frac{(p-\theta)\epsilon}{p(1-p)}}$ Substituting β into the previous equation yields the result.

A.4 Proof of Theorem 4.4

Theorem 4.4 Assuming a heavy hitter e_1 is underestimated, its actual size is f_1, and HeavyLocker provides an estimated size of $\hat{f_1}$. Then, we have:

$$\Pr[\hat{f_1} \geq f_1 - \epsilon] \geq 1 - (\frac{p-\theta}{\theta\epsilon})^{\frac{p}{\theta}}$$

where $p = \frac{f_1}{n}$ Assume that the smallest item in the bucket where e_1 is located is e_2, and its size when replaced by e_1 was f_2. At that time, a total of n_0 items had arrived. Since the bucket was not locked, $f_2 < \theta n_0$. Given that items arrive at a uniform rate, we can simply assume that for each arrival of e_2, there are $\frac{f_1}{f_2}$ arrivals of e_1. Therefore, the probability that e_2 is never replaced is approximately:

$$\Pr[\hat{f}_1 < f_1 - \epsilon] \approx (\prod_{i=1}^{f_2} \frac{i}{i+1})^{\frac{f_1}{f_2}}$$

Note that we have $f_2 < f_1 - \epsilon$, $f_1 = pn_0$, and $f_2 < \theta n_0$. Considering the monotonicity of the right side of the inequality, we find that the maximum value occurs when $f_2 = \theta n_0$ and $(p - \theta)n_0 = \epsilon$. Therefore, we have:

$$\Pr[\hat{f}_1 < f_1 - \epsilon] < (\frac{p - \theta}{\theta \epsilon})^{\frac{p}{\theta}}$$

A.5 Proof of Theorem 4.5

Theorem 4.5 The lock mechanism reduces errors, thereby increasing accuracy. Taking the upper bound of overestimation as an example, the lock operation limits the upper bound of items competing with e_1, thereby keeping the error within the order of magnitude of $-\epsilon$. Without the lock operation, let us make a simple estimate of the expected error:

Lamma 1 Assuming all items arrive at a uniform rate, items e_1 and e_2 compete for a cell using the RAP replacement strategy. Suppose the sizes of e_1 and e_2 are f_1 and f_2 respectively, with $f_1 < f_2$. The estimation of f_1 is $\hat{f}_1$. The expected overestimation of f_1 is:

$$E(max(\hat{f}_1 - f_1), 0) \approx \frac{f_1 f_2^2 - f_1^2 f_2}{(f_1 + f_2)^2}$$

We simply know that e_1 occupies the cell for a fraction of $\frac{f_1}{f_1+f_2}$ of the time, and e_2 occupies it for $\frac{f_2}{f_1+f_2}$ of the time. The ratio of the growth rates of e_1 and e_2 is $\frac{f_1}{f_2}$. Therefore, the expected value of the stored value is $\frac{f_1^2+f_2^2}{f_1+f_2}$. Considering that the probability of f_1 ultimately occupying the cell is $\frac{f_1}{f_1+f_2}$, we obtain:

$$E(max(\hat{f}_1 - f_1), 0)) \approx \frac{f_1}{f_1 + f_2}(\frac{f_1^2 + f_2^2}{f_1 + f_2} - f_1) = \frac{f_1 f_2^2 - f_1^2 f_2}{(f_1 + f_2)^2}$$

Clearly, we note that the above error bound estimation is of polynomial order and related to the distribution of all items. However, with the addition of the lock operation, we can provide an error estimate at an exponential negative order based solely on the lock threshold, which represents a significant improvement. Similarly, the lock operation greatly reduces the lower error bounds of underestimation, which will not be elaborated on here.

A.6 Proof of Theorem 4.6

Theorem 4.6 The error in Global HeavyLocker is less than that in Local HeavyLocker. Note that in our previous error estimations, we omitted a precondition probability—that f_1 is the smallest or smaller in its bucket. Obviously, in Global HeavyLocker, this precondition probability is much smaller: because Global HeavyLocker effectively increases the number of buckets and cell numbers. Therefore, the aforementioned facts hold.

A.7 Proof of Theorem 4.7

Theorem 4.7 Assuming that all streams arrive uniformly, and the bucket threshold at a single point is θ. Then, in the case of multiple data streams: if they are intersecting, the threshold for each Local HeavyLocker is θ; if they are disjoint, the threshold for each Local HeavyLocker is $\theta \cdot ds_{num}$, where ds_{num} represents the number of data stream. Since each item randomly selects a stream, when they are intersecting, the proportion of an item at its chosen data stream remains unchanged. In contrast, when they are disjoint, the expected proportion of an item at any stream is multiplied by the previous number of data streams (ds_{num}). As the threshold should vary with the proportion of each item, this fact holds.

A.8 Proof of Theorem 4.8

Theorem 4.8 Let the precision be denoted by p_0, the target threshold by ϕ, and the lock threshold by θ, the number of heavy hitter is s. We propose the following estimation for p_0:

$$E(p_0) \geq 1 - 4\frac{w}{s}\sqrt{\frac{\phi(1-\phi)}{\pi(\phi-\theta)\epsilon}}e^{\frac{(\phi-\theta)\epsilon}{64\phi(1-\phi)}} - \epsilon_0$$

where $\epsilon_0 < 0.01$. We estimate the error caused by items too close to ϕn as ϵ_0, where 'too close' refers to being greater than $\frac{127}{128}\phi n$. For the remaining items, according to Theorem 3, the probability p_1 that they are overestimated as heavy hitters is:

$$E(p_1) \leq 4\sqrt{\frac{\phi(1-\phi)}{\pi(\phi-\theta)\epsilon}}e^{\frac{(\phi-\theta)\epsilon}{64\phi(1-\phi)}}$$

Given that only the items in the lowest cell of each bucket are prone to overestimation, at most w items could be incorrectly reported as heavy hitters. Therefore, the proportion of items that are overestimated as heavy hitters definitely does not exceed $\frac{w}{s}p_1$. Thus, the above formula for $E(p_0)$ is derived, providing a conservative estimate of the precision rate.

A.9 Proof of Theorem 4.9

Theorem 4.9 Let the recall rate be denoted by r_0, the target threshold by ϕ. Then, we have:

$$E(r_0) \geq 1 - \frac{1}{2\sqrt{2\pi}\beta}e^{-\frac{\beta^2}{2}} - \epsilon_0$$

where $\beta = \frac{d-\frac{1}{w\phi}}{\sqrt{\frac{1}{w\phi}(1-\frac{1}{w})}}$,$\epsilon_0 < 0.01$ In our comprehensive error analysis in Section 4.2 Errorbound, we have determined that the likelihood of underestimation leading to the omission of a true heavy hitter is minimal. Therefore, the main cause of recall error is hash collisions. We simply denote the missing due to underestimation as ϵ_0, which depends on the distribution of items and the size of n, and typically $\epsilon_0 < 0.01$. Focusing on heavy hitters, our primary concern shifts towards hash collisions, which present a more substantial risk for recall inaccuracies. Suppose there are k heavy hitters arriving, and their average number of hash collisions is r_1. So that $k < \frac{1}{\phi}$. Then we can apply Theorem 1 to understand the implications of these collisions:

$$E(r_1) \leq \frac{1}{k}\sum_{s=0}^{k-1} \frac{1}{\sqrt{2\pi}\beta_s} e^{-\frac{\beta_s^2}{2}}$$

where $\beta_s = \frac{d-\frac{s}{w}}{\sqrt{\frac{s}{w}(1-\frac{1}{w})}}$.
Note that the right side of the inequality is a monotonically increasing convex function with respect to s. Therefore, we have:

$$r_1 \leq \frac{1}{2}\frac{1}{\sqrt{2\pi}\beta_k} e^{-\frac{\beta_k^2}{2}} \leq \frac{1}{2}\frac{1}{\sqrt{2\pi}\beta} e^{-\frac{\beta^2}{2}}$$

where $\beta = \frac{d-\frac{1}{w\phi}}{\sqrt{\frac{1}{w\phi}(1-\frac{1}{w})}}$.
Thus, we obtain the required formula.