3. 给定输入流 $< b, a, c, a, d, e, a, f, a, d >$, 计数器个数 $k = 3$ 。请逐步写出 Misra-Gries 算法执行的结果。

b 插入 $(b, 1)$

a 插入 $(b, 1)(a, 1)$

c 插入 $(b, 1)(a, 1)(c, 1)$

删除

a 插入 $(a, 1)$

d 插入 $(a, 1)(d, 1)$

e 插入 $(a, 1)(d, 1)(e, 1)$

删除

a 插入 $(a, 1)$

f 插入 $(a, 1)(f, 1)$

a 插入 $(a, 2)(f, 1)$

d 插入 $(a, 2)(f, 1)(d, 1)$

删除 $(a, 1)$

4. 在利用 count sketch 算法计算数据流中的频繁项时, 输入元素 a, 算法返回结果为:

$$\widehat{f_a} = \text{median}_{1 \leq i \leq t}\, g_i(a) C[i]\left[h_i(a)\right]$$

证明: $t = O(\log(1/\delta))$ 。

$$E(\hat{f}_a) = f_a, Var(\hat{f}_a) = \frac{\|f_{-a}\|_2^2}{k}$$

根据 Chebyshev 不等式,

$$\begin{aligned} P\left[\left|\hat{f}_a - f_a\right| \geqslant \varepsilon \|f\|_2\right] &\leqslant P\left[\left|f_a - f_a\right| \geqslant \epsilon \left\|f_{-a}\right\|_2\right] \\ &\leqslant \frac{\operatorname{Var}\left(f_a\right)}{\epsilon^2 \|f_a\|_2^2} = \frac{1}{k\epsilon^2} \end{aligned}$$

取

$$k = O(1/\epsilon^2)$$

, 就可以使得 $\hat{f}_a$ 偏离真实值超过 $\epsilon\|f\|_2$ 的概率小于 $\frac{1}{3}$

针对第 i 个哈希函数的结果，定义一个指示变量

$$Y_i = \begin{cases} 1 \ if \left|\hat{f}_a - f_a\right| \geq \epsilon\|f\|_2 \\ 0 \ otherwise \end{cases}$$

$$E(Y_i) = P(Y_i = 1) < \frac{1}{3}$$

根据 Chernoff 不等式得:

$$P\left(\sum_i Y_i > \frac{t}{2}\right) \leq P\left(\sum_i Y_i > \left(1 + \frac{1}{2}\right)\mu\right) \leq e^{-\frac{\mu}{16}} < \delta$$

$$\frac{t}{3} \leq \mu \leq 16 \ln\frac{1}{\delta}$$

$$所以t = O\left(\log\left(\frac{1}{\delta}\right)\right)$$

5. 将 count sketch 算法的最后一行 "$\hat{f}_a = \text{median}_{1\leq i\leq t}\, g_i(a)C[i]\left[h_i(a)\right]$ " 改为" $\hat{f}_a = \frac{\sum_{i=1}^t g_i(a)C[i][h_i(a)]}{t}$ " ，其他不变。试分析修改后的算法性能。

若 $\hat{f}_a = \frac{\sum_{i=1}^t g_i(a)C[i][h_i(a)]}{t}$ 则方差为 $\text{var}\left(\hat{f}_a\right) = \frac{\left(\|f_{-a}\|_2\right)^2}{tk}$ 由切比雪夫不等式有: $P\left(\left|\hat{f}_a - f_a\right| \geq \epsilon\|f\|_2\right) \leq P\left(\left|\hat{f}_a - f_a\right| \geq \epsilon\left\|f_{-a}\right\|_2\right) \leq \frac{\text{Var}\left(\hat{f}_a\right)}{\epsilon^2\|f_{-a}\|_2^2} = \frac{1}{tk\epsilon^2} < \delta$

所以 $tk = O\left(\frac{1}{\epsilon^2\delta}\right)$

6. 给定数据流 $< 4, 1, 3, 5, 1, 3, 2, 6, 7, 0, 9 >$, 若哈希函数形如 $h(x) = (ax + b)mod8$, 其中 a 和 b 是任意给定的常数。假设给定如下哈希函数:

(1) $h(x) = (3x + 2)mod8$

(2) $h(x) = (7x + 5)mod8$

(3) $h(x) = (5x + 3)mod8$

试着解答以下问题:

a. 利用 Count-Min sketch 算法估计频繁项。

b. 分析算法在对元素 a 计数时的精确度。

c. 如果想要找到 $(\epsilon, \delta)-$ 估计, 需要怎么样修改算法?

a

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	1	1	2	1	3	1	1
1	1	2	1	3	1	1	1
3	1	2	1	1	1	1	1

$\hat{f}_0=1, \hat{f}_1=3, \hat{f}_2=1, \hat{f}_3=2, \hat{f}_4=1, \hat{f}_5=1, \hat{f}_6=1, \hat{f}_7=1, \hat{f}_8=1, \hat{f}_9=3$

频繁项根据不同的设置，可以是 1、9 或者 1、3、9.

b

对于 $j \in [n] \backslash \{a\}$, 定义如下随机变量表示元素 j 在第 i 个哈希函数上与元素 a 冲突与否:

$$Y_{i,j}=\begin{cases}1, & \text{如果} h_i(j)=h_i(a) \\ 0, & \text{否则}\end{cases}$$

当 $Y_{i,j}=1$ 时, 元素 j 和元素 a 在第 i 个哈希函数上发生了冲突, 此时元素 j 将影响到对元素 a 频数的估计。所以, 定义在第 i 个哈希函数上的其他元素对元素 a 频数估计的贡献为

$$X_i=\sum_{j \in [n] \backslash \{a\}} f_j Y_{i,j}$$

根据期望的线性性质以及 $E\left(Y_{i,j}\right)=1/k$, 得到

$$E\left[X_i\right]=X_i=\sum_{j \in [n] \backslash \{a\}} \frac{f_j}{k}=\frac{\|f\|_1-f_a}{k}=\frac{\|f_{-a}\|_1}{k}$$

因为 $f_j \geqslant 0$, 所以 $X_i \geqslant 0$, 运用 Markov 不等式, 因为本题三个哈希函数并不相互独立，且 $k=8$, 所以得到如下的尾概率不等式:

$$P\left[X_i \geqslant \epsilon\|f\|_1\right] \leqslant P\left[X_i \geqslant \epsilon\left\|f_{-a}\right\|_1\right] \leqslant \frac{\left\|f_{-a}\right\|_1}{k\epsilon\left\|f_{-a}\right\|_1} \quad =\frac{1}{8\epsilon}$$

c

主要从哈希函数、d 和 ω 的选取方面阐述即可。