Count-min sketch: extension to negative counters

Check the analysis seen in class, and discuss how to allow $F[i]$ to change by arbitrary values read in the stream. Namely, the stream is a sequence of pairs of elements, where the first element indicates the item i whose counter is to be changed, and the second element is the amount v of that change (v can vary in each pair). In this way, the operation on the counter becomes $F[i] = F[i] + v$, where the increment and decrement can be now seen as $(i, 1)$ and $(i, -1)$.

SOLUTION

The frequency of the item i is represented by $F[i] \rightarrow T[j, h_j(i)] = F[i] + X_{ji}$, where X_{ji} represent the garbage introduce by the other counter. If we just increment the counter the latter quantity is going to be positive, and then we can take the $min_j T[j, h_j(i)]$ to approximate $F[i]$. Instead, if we have also decrement, it could happen that $X_{ij} < 0$ therefore the method for the min is not going to work. In this case, we consider the absolute value of X_{ji}(i.e. $|X_{ji}|$) and to approximate $F[i]$ we use $median_j T[j, h_j(i)]$. Now let's proof that, with probability $1 - \delta^{1/4}$ holds:

$$F[i] - 3\epsilon\|F\| \leq \hat{F}[i] \leq F[i] + 3\epsilon\|F\|$$

First, let's do some consideration. The value of $\hat{F}[i] = median_j T[j, h_j(i)]$, and $T[j, h_j(i)] = F[i] + |X_{ji}|$. The first inequality(i.e. $F[i] \leq \hat{F}[i]$) holds because we took the absolute value of X_{ji}. Now we shall prove that $Pr[\hat{F}[i] > F[i] + 3\epsilon\|F\|]$. Taken j such that $j = median_j T[j, h_j(i)]$, than we have:

$$\begin{aligned}
&Pr[\hat{F}[i] \leq F[i] + 3\epsilon\|F\|]\\
&Pr[F[i] - 3\epsilon\|F\| \leq F[i] + |X_{ji}| \leq F[i] + 3\epsilon\|F\|]\\
&Pr[|X_{ji}| \leq 3\epsilon\|F\|]
\end{aligned}$$

From what we have seen in class and by the property of the absolute value, we have $E[|X_{ji}|] \leq E[X_{ji}] = \frac{\epsilon}{e}\|F\|$. Then we can apply the Markov inequality and since universal hash are pairwise independence we have:

$$\begin{aligned}
&Pr[|X_{ji}| > 3\epsilon\|F\|]\\
<&\frac{E[|X_{ji}|]}{3\epsilon\|F\|}\\
<&\frac{\frac{\epsilon}{e}\|F\|}{3\epsilon\|F\|}\\
=&\frac{1}{3e} < \frac{1}{8}
\end{aligned}$$

Let's now define the condition variable $Y = \sum_{j=0}^{r} Y_j$ which tell us the number of element i (column) that have a garbage $|X_{ji}|$ greater that $\epsilon\|F\|$.

$$Y_j = \begin{cases} 1 & \text{IF } |X_{ji}| > 3\epsilon\|F\| \text{ WITH } p < \frac{1}{8} \\ 0 & \text{OTHERWISE} \end{cases}$$

The median of $\hat{F}[i]$ is going to be a good approximation if we don't have more that $\frac{r}{2}$ rows such that $|X_{ji}| > 3\epsilon\|F\|$ (that is we want $Y < \frac{r}{2}$).
Therefore, to calculate the probability of error, we calculate the probability that $Pr[Y \geq \frac{r}{2}]$. Here, we can use the Chernoff's Bound We set $(1 + \delta)\mu = \frac{r}{2}$, where $\mu = E(Y) = rp$. Therefore, we have:

$$P[X \geq \frac{r}{2}] <$$