WST501: Exercise List 1

Problem 1 (Mergability of the bloom filter). Let S_1 and S_2 be two sets where the elements come from the same universe U. Let $F(S_1)$ and $F(S_2)$ be the bloom filters on S_1 and S_2 respectively. Recall that a bloom filter is a bit array of length l constructed using a set of hash functions from U to $[l]$, where $[l]$ denotes the set of integers $\{0, ..., l-1\}$. Assume that $F(S_1)$ and $F(S_2)$ have the same length l, and are constructed with the same set of hash functions.

Now, consider $F = F(S_1) \text{ AND } F(S_2)$, where the AND operator produces a bit array of length as l by taking the conjunction of each pair of corresponding bits. Prove that F is exactly the bloom filter on $S_1 \cup S_2$.

Problem 2 (Mergability of the FM-sketch). Let S_1 and S_2 be two bags where the elements come from the same universe U. Let $FM(S_1)$ and $FM(S_2)$ be the FM-sketches on S_1 and S_2 respectively. Recall that each FM-sketch is constructed using a hash function from U to $[2^w]$, where w is set to $\log_2 U$ in our context. Suppose that $FM(S_1)$ and $FM(S_2)$ are built using the same hash function. Describe an algorithm to obtain an FM-sketch on $S_1 \cup S_2$ from $FM(S_1)$ and $FM(S_2)$ in constant time.

Problem 3. In Theorem 4 of the notes of Lecture 3, we made an assumption that $\epsilon < p$. Intuitively, if $\epsilon > p$, the sampling problem should be easier, because the permissible error is large, compared to the real value p. In this problem, we will confirm this intuition.

We need the following variant of the Chernoff bound:

Theorem 1. *Let $X_1, ..., X_k$ be k independent random variables such that, for each $i \in [1, k]$, X_i equals 1 with probability p, and 0 with probability $1 - p$. Let $X = \sum_{i=1}^{k} X_i$ and $\mu = kp$. For any $\alpha \geq 1$, it holds that:*

$$\mathbf{Pr}\big[X \geq (1+\alpha)\mu\big] \quad \leq \quad e^{\frac{-(1+\alpha)\mu}{6}}.$$

Utilize the above theorem to prove the following extension of Theorem 4: Let δ be any value satisfying $0 < \delta < 1$. With $k = 6 \ln \frac{1}{\delta}$, the probability that $|b/k - p| \leq \epsilon$ is at least $1 - \delta$ when $\epsilon > p$. Note that the number k of samples is not even related to ϵ.