Bloomfilter Probability Proof Visualized

First, define the variables below as follows:

- m - the number of bits to check
- n - the size of the hash output

Our induction hypothesis provides us the following lemma:

$$\sum_{inds \in [0..n]^k} (\frac{1}{n})^k (p \in inds \wedge ps \subseteq inds) \leq \underbrace{(1 - (1 - \frac{1}{n})^k)}_{P[\text{p is not inds}]} \times \underbrace{\sum_{inds \in [0..n]^k} \frac{1}{n}^k (ps \subseteq inds)}_{P[\text{ps is contained in inds}]}$$

Which can be roughly read as, the probability that the element p and the list ps will be found in a randomly drawn list is less than the product of the probability that p is found and the probability that ps is found.

Now, let's move on to prove the inductive step. Simplifying a bit, we obtain a goal of the following form.[1]

$$\sum_{inds \in [0..n]^{k+1}} (\frac{1}{n})^{k+1} (p \in inds \wedge ps \subseteq inds) \leq$$

$$(1 - (1 - \frac{1}{n})^{k+1}) \left((1 - \frac{m}{n}) \sum_{inds \in [0..n]^k} (\frac{1}{n})^k (ps \subseteq inds) + \sum_{ind \in ps} \sum_{inds \in [0..n]^k} \frac{1}{n}^{k+1} (ps \subseteq \{i\} \cup inds) \right)$$

Using the fact that:

$$\sum_{inds \in [0..n]+} \frac{1}{n}^k (ps \subseteq inds) \leq \sum_{inds \in [0..n]^k} \frac{1}{n}^k (ps \subseteq \{i\} \cup inds)$$

We can replace simplify the nested summation to the form:

$$\sum_{inds \in [0..n]^{k+1}} (\frac{1}{n})^{k+1} (p \in inds \wedge ps \subseteq inds) \leq$$

$$(1 - (1 - \frac{1}{n})^{k+1}) \left((1 - \frac{m}{n}) \sum_{inds \in [0..n]^k} (\frac{1}{n})^k (ps \subseteq inds) + \frac{m}{n} \sum_{inds \in [0..n]^k} (\frac{1}{n})^k (ps \subseteq inds) \right)$$

Factoring out the common term, we finaly obtain:

$$\sum_{inds \in [0..n]^{k+1}} (\frac{1}{n})^{k+1} (p \in inds \wedge ps \subseteq inds) \leq$$

$$(1 - (1 - \frac{1}{n})^{k+1}) \sum_{inds \in [0..n]^k} (\frac{1}{n})^k (ps \subseteq inds)$$

Expanding $1 - (1 - \frac{1}{n})^{k+1}$:

$$\sum_{inds \in [0..n]^{k+1}} (\frac{1}{n})^{k+1} (p \in inds \wedge ps \subseteq inds) \leq (1 - (1 - \frac{1}{n})^k + \frac{1}{n}(1 - \frac{1}{n})^k) \sum_{inds \in [0..n]^k} (\frac{1}{n})^k (ps \subseteq inds)$$

[1] the simplified expression on the RHS was obtained by splitting drawing a random list of length $k+1$ into drawing a single random element and drawing the remaining random list.

Then, applying the induction hypothesis:

$$\sum_{inds \in [0..n]^{k+1}} (\frac{1}{n})^{k+1}(p \in inds \wedge ps \subseteq inds) \leq \sum_{inds \in [0..n]^{k}} (\frac{1}{n})^{k}(p \in inds \wedge ps \subseteq inds) + C$$

Where for notational conciseness, C is:

$$C = \frac{1}{n}(1 - \frac{1}{n})^k \sum_{ind \in [0..n]^k} (\frac{1}{n})^k (ps \subseteq inds)$$

Focusing on the LHS:

$$\begin{aligned}
&\sum_{inds \in [0..n]^{k+1}} (\frac{1}{n})^{k+1}(p \in inds \wedge ps \subseteq inds) \\
&= \sum_{ind \in [0..n]} \sum_{inds \in [0..n]^k} (\frac{1}{n})^{k+1}(p \in \{ind\} \cup inds \wedge ps \subseteq \{ind\} \cup inds) \\
&\leq \sum_{ind \in [0..n]} \sum_{inds \in [0..n]^k} (\frac{1}{n})^{k+1}(ps \subseteq \{ind\} \cup inds) \\
&= \sum_{ind \in [0..n]} \frac{1}{n} \left(\sum_{inds \in [0..n]^k} (\frac{1}{n})^{k}(ps \subseteq \{ind\} \cup inds) \right)
\end{aligned}$$

Done.

$$\begin{aligned}
&\sum_{inds \in [0..n]^{k+1}} (\frac{1}{n})^{k+1}(p \in inds \wedge ps \subseteq inds) \\
&= \sum_{ind \in [0..n]} \sum_{inds \in [0..n]^k} (\frac{1}{n})^{k+1}(p \in \{ind\} \cup inds \wedge ps \subseteq \{ind\} \cup inds) \\
&= \frac{1}{n} \left(\sum_{inds \in [0..n]^k} (\frac{1}{n})^{k}(ps \subseteq inds) \right) + \left(\sum_{ind \neq p} \sum_{inds \in [0..n]^k} (\frac{1}{n})^{k+1}(p \in inds \wedge ps \subseteq \{ind\} \cup inds) \right) \\
&= \frac{1}{n} \left(\sum_{inds \in [0..n]^k} (\frac{1}{n})^{k}(ps \subseteq inds) \right) + \left(\sum_{ind \in ps} \sum_{inds \in [0..n]^k} (\frac{1}{n})^{k+1}(p \in inds \wedge ps \subseteq \{ind\} \cup inds) \right) + \\
&\qquad \left(\sum_{ind \notin p::ps} \sum_{inds \in [0..n]^k} (\frac{1}{n})^{k+1}(p \in inds \wedge ps \subseteq inds) \right) \\
&= \frac{1}{n} \left(\sum_{inds \in [0..n]^k} (\frac{1}{n})^{k}(ps \subseteq inds) \right) + \left(\sum_{ind \in ps} \sum_{inds \in [0..n]^k} (\frac{1}{n})^{k+1}(p \in inds \wedge ps \subseteq \{ind\} \cup inds) \right) + \\
&\qquad (1 - \frac{m+1}{n}) \left(\sum_{inds \in [0..n]^k} (\frac{1}{n})^{k}(p \in inds \wedge ps \subseteq inds) \right) \\
&\leq (1 - \frac{m}{n}) \left(\sum_{inds \in [0..n]^k} (\frac{1}{n})^{k}(ps \subseteq inds) \right) + \left(\sum_{ind \in ps} \sum_{inds \in [0..n]^k} (\frac{1}{n})^{k+1}(p \in inds \wedge ps \subseteq \{ind\} \cup inds) \right)
\end{aligned}$$