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:

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

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]

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

Noticing that the second additive term is simply a marginalization of the internal distribution, we can eliminate the nested sum:

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

As $\sum_{ind \in [0..n]^k} \frac{1}{n}^k (tail\ ps \subseteq inds) \leq \sum_{ind \in [0..n]^k} \frac{1}{n}^k (ps \subseteq inds)$, we can remove the tail operation and factor out the sum from the addition, simplifying to the form[2]:

$$\sum_{inds \in [0..n]^{k+1}} (\frac{1}{n})^{k+1} (p \in inds \land ps \subseteq inds) \leq$$
$$(1 - (1 - \frac{1}{n})^{k+1})(2 - \frac{m}{n}) \left(\sum_{inds \in [0..n]^k} (\frac{1}{n})^k (ps \subseteq inds) \right)$$

As $(1 - (1 - \frac{1}{n})^k) \leq (1 - (1 - \frac{1}{n})^{k+1})$, we can reduce the upper bound and apply the induction hypothesis:

$$\sum_{inds \in [0..n]^{k+1}} (\frac{1}{n})^{k+1} (p \in inds \land ps \subseteq inds) \leq$$
$$(2 - \frac{m}{n}) \sum_{inds \in [0..n]^k} (\frac{1}{n})^k (p \in inds \land ps \subseteq inds)$$

[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.

[2] this is also why this property is not independent (I think...)

The remainder of the proof is trivial[3], in that the LHS will simplify down to an expression of the form $c \times \sum_{inds \in [0..n]^k} \cdots$, where $c \leq 1$, thus obviously $c \leq 2 - \frac{m}{n}$.

[3] at least conceptually, unfortunately probably not mechanically