1 Concepts to Study

Foundational Probability Concepts + Concentration Bounds
- Linearity of expectation and variance.
- Markov’s inequality, Chebyshev’s inequality (should know from memory).
- Union bound (should know from memory).
- General idea of higher moment inequalities.
- Chernoff and Bernstein bounds (don’t need to memorize the exact bounds, but should be able to apply if given).
- General idea of law of large numbers and central limit theorem.
- Technique of breaking random variables into sums of indicator random variables.
- Averaging to reduce error.
- Median trick.

Random Hashing and Related Algorithms
- Random hash functions.
- Definitions of 2-universal and pairwise independent hash functions (should have memorized).
- Application of random hashing to load balancing.
- Hashing for Distinct Elements. Understand the ‘idealized’ algorithm where we hash to real numbers. Don’t need to understand details of HyperLogLog.
- Bloom Filters. Don’t need to have formulas memorized.
- MinHash for Jaccard similarity.
- Idea of locality sensitive hashing. How it is used for similarity search (with hash signatures and repeated tables). Idea of s-curve tuning (don’t need to memorize formula).

Other
- Frequent elements problem definition and setup.
- High level idea of Boyer-Moore and Misra-Gries, but don’t need to know in detail.
- Count-min sketch and analysis.
- The Johnson-Lindenstrauss Lemma. Don’t need to memorize, but should understand and be able to apply if given.
- Do not need to be able to recreate the JL proof, but should understand the ideas behind it.
2 Practice Questions

Work in progress. Check back to see if more questions have been added.

Probability, Expectation, Variance:

1. Exercises 2.1, 2.3, 2.4, 2.28, 2.41 of Foundations of Data Science (https://www.cs.cornell.edu/jeh/book.pdf)

2. Show that for any $X$, $E[X^2] \geq E[X]^2$.


   **Hint:** use part (3).

5. For the statements below, indicate if they are always true, sometimes true, or never true. Give a sentence explaining why.
   (a) $Pr[X = s \cap Y = t] > Pr[X = s]$. ALWAYS  SOMETIMES  NEVER
   (b) $Pr[X = s \cup Y = t] \leq Pr[X = s] + Pr[Y = t]$. ALWAYS  SOMETIMES  NEVER
   (c) $Pr[X = s \cap Y = t] = Pr[X = s] \cdot Pr[Y = t]$. ALWAYS  SOMETIMES  4 NEVER

Concentration Inequalities:

1. Let $X_1, \ldots, X_n$ be the number of visitors to a website on $n$ consecutive days. These are independent and identically distributed random variables. We have $E[X_i] = 20,000$ and $Var[X_i] = 100,000,000$.
   (a) Give an upper bound on the probability that on day $i$, more than 40,000 visitors hit the website.
   (b) Let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ be the average number of visitors over $n$ days. What are $E[\bar{X}]$ and $Var[\bar{X}]$?
   (c) Give an upper bound on the probability that $\bar{X} \geq 25,000$, for $n = 100$.

2. Assume there are 1000 registered users on your site $u_1, \ldots, u_{1000}$, and in a given day, each user visits the site with some probability $p_i$. The event that any user visits the site is independent of what the other users do. Assume that $\sum_{i=1}^{1000} p_i = 500$.
   (a) Let $X$ be the number of users that visit the site on the given day. What is $E[X]$.
   (b) Apply a Chernoff bound to show that $Pr[X \geq 600] \leq .01$.

Random Hashing Algorithms:

1. Exercises 6.1, 6.2, 6.6, 6.7, 6.10, 6.19, 6.22, 6.23 of Foundations of Data Science

2. Consider a hash function mapping $m$-bit strings to a single bit – $h : \{0,1\}^m \to \{0,1\}$. We generate $h$ by selecting a random position $i$ from 1, \ldots, $m$. Then let $h(x) = x(i)$, the value of $x$ at position $i$. Note that after $i$ is chosen, it remains fixed, when we apply $h$ to different inputs.
   (a) Given $x, y \in \{0,1\}^m$ with hamming distance $\|x - y\|_0$ (i.e., $x$ and $y$ have different bit values in $\|x - y\|_0$ positions), what is $Pr[h(x) = h(y)]$. 


(b) Is $h$ a locality sensitive hash function?

(c) Let $m$ be the number of all possible 5-singles in a document (i.e., all possible strings of 5 English words). If $x$ and $y$ are indicator vectors of the 5-shingles in two different documents, why do we expect them to be very sparse (i.e., each only have a few bits set to 1)?

(d) Why might MinHash and Jaccard similarity be more useful in the situation of (c) than the hash function $h$ and Hamming distance.

3. Use a Chernoff bound to show that if we hash $n$ items into a table with $n$ buckets, with probability $\geq 1 - \delta$, the maximum number of items in a single bucket is upper bounded by $O(\log n/\delta)$. 