Assignment 7: Hashing and Sketching

This final problem set of the quarter explores hash functions and sketches. Once you've finished, you'll be all set to start working on the final project!

Working in Pairs

We suggest working on this problem set in pairs. If you work in a pair, you should jointly submit a single assignment, which will be graded out of 10 points. If you work individually, the problem set will be graded out of 9 points, but we won't award extra credit if you earn more than 9 points.

Due Wednesday, May 21 at 2:15PM at the start of lecture.

Problem One: Subcritical Galton-Watson Processes (2 Points)

For a given subcritical Galton-Watson process with underlying variable ξ, let X_n denote the number of nodes alive in the process after n generations. In lecture, we claimed that if $E[\xi] < 1$, then $E[X_n] = E[\xi]^n$. Prove this. For reference, the inductive definition of X_n is

$$X_0 = 1 \qquad X_{n+1} = \sum_{i=1}^{X_n} \xi_{i,n}$$

Here, $\xi_{i,n}$ is an i.i.d. copy of ξ. *(Hint: Use conditional expectation)*

Problem Two: Count-Min Sketches and Zipfian Distributions (7 Points)

Part of the beauty of the count-min sketch is that it makes no assumptions about the underlying distribution of the data when obtaining its bounds. However, if we start making stronger assumptions about how the input data are distributed, we can show that the count-min sketch produces tighter bounds and, therefore, needs less space to guarantee its bounds.

A *Zipfian* distribution is a probability distribution over a (possibly infinite) set $S = \{ x_1, x_2, x_3, \ldots \}$. The Zipfian distribution is parameterized by a value z and has probability

$$P(x_i) = \alpha i^{-z}$$

Here, α is a normalization constant chosen so that the probability mass sums to one. You may find it useful to note that when $z > 1$, the normalization constant α satisfies $\alpha \leq z - 1$.

Notice that this distribution makes x_1 the most probable element, then x_2, then x_3, etc.

Many realistic data sets obey a Zipfian distribution, such as the frequency of words in natural language. In the case where $z > 1$, the Zipfian distribution is highly biased toward the most-frequently-occurring elements.

i. **(1 Point)** Let X be sampled from a Zipfian distribution where $z > 1$. This distribution is sufficiently skewed that the likelihood of not choosing one of the k most frequent elements is extremely small. Prove for all $k \geq 1$ that $P[X > k] \leq k^{1-z}$. *(Hint: Express the probability as an summation, then upper-bound that summation with an integral.)*

ii. **(1 Point)** Suppose that n elements are drawn from a Zipfian distribution where $z > 1$. Prove that the expected number of times that element x_i is drawn is $n\alpha i^{-z}$.

Suppose we are processing a data stream that comes from Zipfian distribution with parameter $z > 1$. Let w be the width of each array in the count-min sketch and d the number of rows in the sketch.

iii. **(2 Points)** Let $\hat{\boldsymbol{a}}_i$ be the estimate of $\boldsymbol{a}_i$ given by an arbitrary row of a count-min sketch. Prove that the probability that the $w/3$ most-frequent elements of $\mathscr{U}$ don't collide with x_i is at least 2/3 and that if these elements don't collide with x_i, then $E[\hat{\boldsymbol{a}}_i - \boldsymbol{a}_i] \leq w^{-z}\, 3^{z-1}\, \|\boldsymbol{a}\|_1$.

iv. **(1 Point)** Show that there is a constant $0 < p < 1$ such that with probability at least p, the frequency estimate for an element x_i in a particular row of the count-min sketch is less than or equal to $\boldsymbol{a}_i + w^{-z}\, 3^z\, \|\boldsymbol{a}\|_1$.

v. **(2 Points)** Show that with appropriate choices of w and d in terms of ε and δ, a count-min sketch can be constructed for a Zipfian distribution with $z > 1$ in $O((1 / \varepsilon)^{1/z} \log (1 / \delta))$ space such that any frequency estimation, with probability $1 - \delta$, overestimates the total frequency by at most $\varepsilon \|f\|_1$. In other words, we can obtain the same accuracy bounds as a normal count-min sketch, but with reduced space usage.

Problem Three: Course Feedback (1 Point)

We want this course to be as good as it can be and would really appreciate your feedback on how we're doing. For a free point, please take a few minutes to answer the course feedback questions available at https://docs.google.com/forms/d/1a-oev-0prNsKJp0-P0NKhkyPNTwYcPOt1mydo4YEOTw/viewform. If you submit this problem set in a group, **please have each group member fill this out individually**.