Problem Set 4

[Your name] and [student ID]
MAT1841-2021

Problem 1 [Thorup-inspired] (30 points). Let $U = \mathbb{F}_q^*$ be the set of variable-length strings in the Galois Field $\mathbb{F}_q$. In class, we learned how to construct universal hash families from $U \to \mathbb{F}_q$ and k-independent hash families from $\mathbb{F}_q \to \mathbb{F}_q$.

Construct an efficient hash function that is with good probability k-independent when restricted to a finite subset $S \subset U$, where $|S| = n$. What is the failure probability of your scheme?

Problem 2 [BHK 6.7] (30 points). Consider an algorithm that uses a random hash function and gives an estimate $\hat{x}$ of the true value x of some variable. Suppose that $\frac{x}{4} \leq \hat{x} \leq 4x$ with probability at least 0.6. The probability of the estimate is with respect to the choice of the hash function. How would you improve the probability that $\frac{x}{4} \leq \hat{x} \leq 4x$ to 0.8?

Problem 3 (40 points). The MinHash sketch measures the Jaccard index (resemblance) of two sets by storing the minimum hash value of each set; the probability that the minimum hash values are the same is precisely the Jaccard index. Obviously, to get a reasonable error, you will want to repeat the process multiple times (roughly k times to get $O\sqrt{k}$ additive error).

MinHash has been applied to biological sequences to measure similarity by measuring the Jaccard index of the set of 'k-mers' (length-k substrings) of a sequence. For example, the string AACCGGTT has 4-mers AACC, ACCG, CCGG, CGGT, GGTT.

Write a Python function with call signature

```
def approximate_jaccard(A, B, k):
   '''A and B are Python strings
      k is an integer specifying the k-mer length
      ans is a float
   '''
   ...
   return ans
```

The Python function should take two strings A and B, and compute the Jaccard index of their k-mer sets to an error of 10% with 95% probability. I will be running your code on real bacterial sequences, so be sure your code is scalable. i.e. I'll be unhappy if it crashes my computer. Some marks will be given/taken off for efficiency.