Practical session 7

Exercise 1 (Equality of long strings)
Suppose that Alice and Bob have long binary strings a and b respectively and they want to check if the strings are equal. A trivial solution would be for Alice to transmit a to Bob and for Bob to then check whether $a = b$. We would like to have a solution with lower communication complexity, that is, an algorithm in which less bits are transmitted from Alice to Bob (or vice versa). For a binary string s let $H(s)$ be the natural number whose binary representation is s. Consider the following communication protocol.

Algorithm 0.1.
Alice's input is $a \in \{0,1\}^n$ and Bob's input is $b \in \{0,1\}^n$, for some large $n \in \mathbb{N}$.

1. *Given a positive integer $M < 2^n$, Alice draws a prime number $p \leq M$ uniformly at random.*
2. *Alice calculates $a' = H(a) \mod p$. She then transmits a' and p to Bob.*
3. *Bob outputs YES if $H(b) \mod p = a'$ and NO otherwise.*

For $n \in \mathbb{N}$, let $\pi(n)$ denote the number of prime numbers less than n. We will use some known facts from Number Theory without proof.

Fact 0.2.
$$\pi(n) = (1 + o(1))\frac{n}{\log n}.$$

Fact 0.3. *For sufficiently large n and for every $m \leq 2^n$, it holds that the number of primes that divide m is less than $\pi(n)$.*

For sufficiently large n, use the above facts to upper bound the probability that Bob's output is incorrect.

Solution
If $a = b$, then for all prime numbers p, it holds that $H(a) \equiv H(b) \mod p$ and thus the output is always correct in this case. Assume then that $a \neq b$. Then Bob's output is incorrect if and only if p divides $k := |H(a) - H(b)|$. Since $k \leq 2^n$, it follows by Fact 0.3 that the number of primes that divides k is at most $\pi(n)$. Therefore

$$\Pr(\text{Bob's output is incorrect}) = \frac{|\{p < M : p \text{ is prime and } p|k\}|}{\pi(M)} \leq \frac{\pi(n)}{\pi(M)} = (1+o(1)) \cdot \frac{n}{M} \cdot \frac{\log M}{\log n},$$

where the last equality holds by Fact 0.2. Note that the total number of bits transmitted is $\lceil \log_2 a' \rceil + \lceil \log_2 p \rceil \leq 2\lceil \log_2 M \rceil$. Therefore the choice of M is a tradeoff between the error probability of the algorithm and its communication complexity.