Problem Set 5 — *Due Friday, November 29, before class starts*

For the Exercise Sessions on Nov 15 and Nov 22

Last name	First name	SCIPER Nr	Points

Problem 1: Poisson Sampling

Assume that we have given a distribution p on $\mathcal{X} = \{1, \cdots, k\}$. Let X^n denote a sequence of n iid samples. Let $T_i = T_i(X^n)$ be the number of times symbol i appears in X^n. Then

$$\mathbb{P}\{T_i = t_i\} = \binom{n}{t_i} p_i^{t_i} (1 - p_i)^{n - t_i}.$$

Note that the random variables T_i are *dependent*, since $\sum_i T_i = n$. This dependence can sometimes be inconvenient.

There is a convenient way of getting around this problem. This is called *Poisson* sampling. Let N be a random variable distributed according to a Poisson distribution with mean n. Let X^N be then an iid sequence of N variables distributed according to p.

Conditioned on $N = n$, what is the induced distribution of the Poisson sampling scheme?

Now, show that:

1. $T_i(X^N)$ is distributed according to a Poisson random variable with mean $p_i n$.
2. The $T_i(X^N)$ are independent.

Problem 2: Add-β Estimator

The add-β estimator $q_{+\beta}$ over $[k]$, assigns to symbol i a probability proportional to its number of occurrences plus β, namely,

$$q_i \stackrel{\text{def}}{=} q_i(X^n) \stackrel{\text{def}}{=} q_{+\beta,i}(X^n) \stackrel{\text{def}}{=} \frac{T_i + \beta}{n + k\beta}$$

where $T_i \stackrel{\text{def}}{=} T_i(X^n) \stackrel{\text{def}}{=} \sum_{j=1}^n \mathbf{1}(X_j = i)$. Prove that for all $k \geq 2$ and $n \geq 1$,

$$\min_{\beta \geq 0} r_{k,n}^{l_2^2}(q_{+\beta}) = r_{k,n}^{l_2^2}(q_{+\sqrt{n}/k}) = \frac{1 - \frac{1}{k}}{(\sqrt{n} + 1)^2}$$

Furthermore, $q_{+\sqrt{n}/k}$ has the same expected loss for every distribution $p \in \Delta_k$.

Problem 3: Uniformity Testing

Let us reconsider the problem of testing against uniformity. In the lecture we saw a particular *test statistics* that required only $O(\sqrt{k}/\epsilon^2)$ samples where ϵ was the ℓ_1 distance.

Let us now derive a test from scratch. To make things simple let us consider the ℓ_2^2 distance. Recall that the alphabet is $\mathcal{X} = \{1, \cdots, k\}$, where k is known. Let U be the uniform distribution on $\mathcal{X}$, i.e., $u_i = 1/k$. Let P be a given distribution with components p_i. Let X^n be a set of n iid samples. A pair of samples (X_i, X_j), $i \neq j$, is said to *collide* if $X_i = X_j$, if they take on the same value.

1. Show that the expected number of collisions is equal to $\binom{n}{2}\|p\|_2^2$.
2. Show that the uniform distribution minimizes this quantity and compute this minimum.
3. Show that $\|p - u\|_2^2 = \|p\|_2^2 - \frac{1}{k}$.

   NOTE: In words, if we want to distinguish between the uniform distribution and distributions P that have an ℓ_2^2 distance from U of at least ϵ, then this implies that for those distributions $\|p\|_2^2 \geq 1/k + \epsilon$. Together with the first point this suggests the following test: compute the number of collisions in a sample and compare it to $\binom{n}{2}(1/k + \epsilon/2)$. If it is below this threshold decide on the uniform one. What remains is to compute the variance of the collision number as a function of the sample size. This will tell us how many samples we need in order for the test to be reliable.
4. Let $a = \sum_i p_i^2$ and $b = \sum_i p_i^3$. Show that the variance of the collision number is equal to

$$\binom{n}{2} a + \binom{n}{2}\left[\binom{n}{2} - \left(1 + \binom{n-2}{2}\right)\right] b + \binom{n}{2}\binom{n-2}{2} a^2 - \binom{n}{2}^2 a^2$$
$$= \binom{n}{2}\left[2b(n-2) + a(1 + a(3-2n))\right]$$

   by giving an interpretation of each of the terms in the above sum.

   NOTE: If you don't have sufficient time, skip this step and go to the last point.

   For the uniform distribution this is equal to

$$\binom{n}{2}\frac{(k-1)(2n-3)}{k^2} \leq \frac{n^2}{2k}.$$

   NOTE: You don't have to derive this from the previous result. Just assume it.
5. Recall that we are considering the ℓ_2^2 distance which becomes generically small when k is large. Therefore, the proper scale to consider is $\epsilon = \kappa/k$. Use the Chebyshev inequality and conclude that if we have $\Theta(\sqrt{k}/\kappa^2)$ samples then with high probability the empirical number of collisions will be less than $\binom{n}{2}(1/k + \kappa/(2k))$ assuming that we get samples from a uniform distribution.

NOTE: The second part, namely verifying that the number of collisions is with high probability no smaller than $\binom{n}{2}(1/k + \kappa/(2k))$ when we get $\Theta(\sqrt{k}/\kappa^2)$ samples from a distribution with ℓ_2^2 distance at least κ/k away from a uniform distribution follows in a similar way.

HINT: Note that if p represents a vector with components p_i then $\|p\|_1 = \sum_i |p_i|$ and $\|p\|_2^2 = \sum_i p_i^2$.