ECE 6980
Algorithmic and Information-Theoretic Methods in Data Science

Instructor: Jayadev Acharya — Lecture #7
Scribe: Chamsi Hssaine — 18th September, 2017

1 Introduction

In the last lecture, we covered Le Cam's two-point theorem.
In this lecture, we will continue the proof of the lower bound for uniformity testing.
We claimed that, because we are sampling $X_1, \ldots, X_n$ independently, it is enough to look at $N_1, \ldots, N_K$ (these are called *sufficient statistics*), the number of times each symbol i appears.
Recall, we constructed $2^{K/2}$ distributions by generating a string $\bar{Z} = Z_1 \ldots Z_{K/2}$, with $Prob(Z_i = 1) = Prob(Z_i = -1) = 1/2$, where Z_i's are independent. Note that there are $2^{K/2}$ such strings.
Given $\bar{Z}$, we construct a distribution p over K elements such that:

$$\begin{cases} p_{\bar{Z}}(2i-1) = \frac{1+5\epsilon Z_i}{K} \\ p_{\bar{Z}}(2i) = \frac{1-5\epsilon Z_i}{K} \end{cases}$$

Note that $d_{TV}(p_{\bar{Z}}, u) = 5\epsilon, \forall \bar{Z}$.
We will show that we need $\Omega(\frac{\sqrt{K}}{\epsilon^2})$ samples if we sample uniformly at random from the set of $\bar{Z}$.

2 Lower bound for uniformity testing

Theorem 1. *$\Omega(\frac{\sqrt{K}}{\epsilon^2})$ is a lower bound for uniformity testing.*

Proof. Throughout the rest of the proof, we assume Poisson sampling.
We first consider the case where $p = u$, so $N_i \sim Poi(n \cdot \frac{1}{K})$. Then:

$$Prob_{p=u}(N_1 = n_1, \ldots, N_K = n_K) = e^{-n} \prod_{i=1}^{K} \frac{(n/K)^{n_i}}{n_i!}.$$

In general,we have:

$$Prob(N_1 = n_1, \ldots, N_K = n_K) = \frac{1}{2^{K/2}} \sum_{\bar{Z}} Prob_{p_{\bar{Z}}}(N_1 = n_1, \ldots, N_K = n_K).$$

We leave it as an exercise to the reader to show that we also have:

$$Prob(N_1 = n_1, \ldots, N_K = n_K) = \prod_{i=1}^{K/2} Prob(N_{2i-1} = n_{2i-1}, N_{2i} = n_{2i}).$$

Let's first compute $Prob(N_1 = n_1, N_2 = n_2)$. By conditioning on Z_1, we get that:

$$\begin{aligned}
Prob(N_1 = n_1, N_2 = n_2) =& \frac{1}{2} \cdot e^{-n\frac{1+5\epsilon}{K}} \frac{\left(n(1+5\epsilon)/K\right)^{n_1}}{n_1!} \cdot e^{-n\frac{1-5\epsilon}{K}} \frac{\left(n(1-5\epsilon)/K\right)^{n_2}}{n_2!} \\
&+ \frac{1}{2} \cdot e^{-n\frac{1-5\epsilon}{K}} \frac{\left(n(1-5\epsilon)/K\right)^{n_1}}{n_1!} \cdot e^{-n\frac{1+5\epsilon}{K}} \frac{\left(n(1+5\epsilon)/K\right)^{n_2}}{n_2!} \\
=& \frac{1}{2} e^{-2n/K} \frac{(n/K)^{n_1}(n_K)^{n_2}}{n_1! n_2!} \left((1+5\epsilon)^{n_1}(1-5\epsilon)^{n_2} + (1-5\epsilon)^{n_1}(1+5\epsilon)^{n_2} \right).
\end{aligned}$$

We can extend this to obtain an expression for $Prob(N_1 = n_1, \ldots, N_K = n_K)$:

$$Prob(N_1 = n_1, \ldots, N_K = n_K) = e^{-n} \left[\prod_{i=1}^{K} \frac{(n/K)^{n_i}}{n_i!} \right] \cdot \left[\prod_{i=1}^{K/2} \frac{1}{2} \Big((1+5\epsilon)^{n_1}(1-5\epsilon)^{n_2} + (1-5\epsilon)^{n_1}(1+5\epsilon)^{n_2} \Big) \right]$$

Note that if you set $\epsilon = 0$, then you obtain exactly the expression for when $p = u$.

Let u^{*n} denote the process obtained from picking n samples independently from the uniform distribution, and $p^{*n}_{\bar{Z}}$ the process obtained from picking n samples from the distribution constructed, given $\bar{Z}$.

Recall from Assignment 1 that the optimal testing scheme gives us a probability of error $P^*_e = \frac{1}{2} - \frac{1}{2} d_{TV}(p, q)$ for arbitrary distribution p, q. Assume we want $P^*_e \leq 0.1$, then we need $d_{TV}(p^{*n}_{\bar{Z}}, u^{*n}) \geq 0.8$, and consequently $\chi^2(p^{*n}_{\bar{Z}}, u^{*n}) \geq 1.28$ (see Appendix for explanation).

We compute $\chi^2(p^{*n}_{\bar{Z}}, u^{*n}) = \mathbb{E}_{p_{\bar{Z}^{*n}}} \left[\frac{p^{*n}_{\bar{Z}}}{u^{*n}} \right] - 1$.

$$\begin{aligned}
\frac{p^{*n}_{\bar{Z}}(N_1 = n_1, \ldots, N_K = n_K)}{u^{*n}(N_1 = n_1, \ldots, N_K = n_K)} =& \prod_{i=1}^{K/2} \frac{1}{2} \Big((1+5\epsilon)^{n_1}(1-5\epsilon)^{n_2} + (1-5\epsilon)^{n_1}(1+5\epsilon)^{n_2} \Big) \\
\implies \mathbb{E}_{p_{\bar{Z}^{*n}}} \left[\frac{p^{*n}_{\bar{Z}}}{u^{*n}} \right] =& \mathbb{E}_{p_{\bar{Z}^{*n}}} \left[\prod_{i=1}^{K/2} \frac{1}{2} \Big((1+5\epsilon)^{n_1}(1-5\epsilon)^{n_2} + (1-5\epsilon)^{n_1}(1+5\epsilon)^{n_2} \Big) \right] \\
=& \prod_{i=1}^{K/2} \mathbb{E}_{p_{\bar{Z}^{*n}}} \left[\frac{1}{2} \Big((1+5\epsilon)^{n_1}(1-5\epsilon)^{n_2} + (1-5\epsilon)^{n_1}(1+5\epsilon)^{n_2} \Big) \right] \\
=& \prod_{i=1}^{K/2} \frac{1}{2} \cdot \frac{1}{2} \Bigg(e^{(5\epsilon)\cdot n(1+5\epsilon)/K} \cdot e^{(-5\epsilon)\cdot n(1-5\epsilon)/K} + e^{(-5\epsilon)\cdot n(1+5\epsilon)/K} \cdot e^{(5\epsilon)\cdot n(1-5\epsilon)/K} \\
&+ e^{(5\epsilon)\cdot n(1-5\epsilon)/K} \cdot e^{(-5\epsilon)\cdot n(1+5\epsilon)/K} + e^{(-5\epsilon)\cdot n(1-5\epsilon)/K} \cdot e^{(5\epsilon)\cdot n(1+5\epsilon)/K} \Bigg) \\
=& \left(\frac{e^{50n\epsilon^2/K} + e^{-50n\epsilon^2/K}}{2} \right)^{K/2} \\
\implies \chi^2(p^{*n}_{\bar{Z}}, u^{*n}) =& \left(\frac{e^{50n\epsilon^2/K} + e^{-50n\epsilon^2/K}}{2} \right)^{K/2} - 1,
\end{aligned}$$

where the third equality follows from the fact that each Z_i is drawn independently, so we can pull the product outside of the expectation.
By Taylor series expansion, we have that $\frac{e^x+e^{-x}}{2} \leq e^{x^2/2}$. We use this to obtain the upper bound on the $\chi^2(p_{\bar{Z}}^{*n}, u^{*n})$:

$$\begin{aligned} 1 < \chi^2(p_{\bar{Z}}^{*n}, u^{*n}) &\leq e^{1250n^2\epsilon^4/K} - 1 \\ \iff \frac{n^2\epsilon^4}{K} &> \frac{\log 2}{1250} \\ \iff n &> \frac{\sqrt{K}}{\epsilon^2}\sqrt{\frac{\log 2}{1250}} \end{aligned}$$

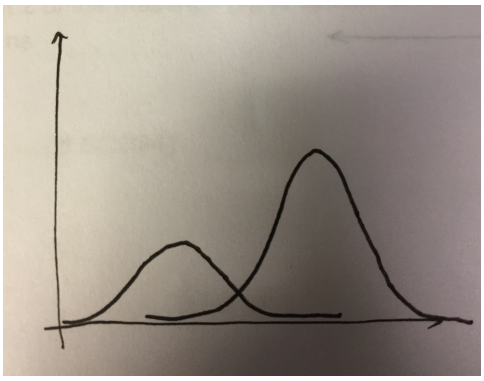
□

3 Mixture Models

We now change gears to discuss mixture models, used to characterized more complex distributions than those we've seen thus far. The high-level explanation is that there are K underlying distributions, one is chosen according to some probability distribution, and samples are generated according to the chosen distribution.

3.1 Motivation

The motivation for mixture models arose around 1895, with Karl Pearson's crab experiment. In his experiment, he collected 1,000 crabs and took the ration of their height and their weight, expecting to observe a normal distribution. However, what he observed was a bimodal distribution, resembling the figure below:



Mixture models have more recently been used for tasks such as digit recognition, and have found applications in such things as housing prices.

3.2 Setup

- K distributions $p_1, \ldots, p_K$
- K corresponding weights $w_1, \ldots, w_K$, $w_i \geq 0, \sum_i w_i = 1$
- $p(x) = \sum_{i=1}^{K} w_i p_i(x)$

- **Problem:** You don't get to see which distribution p_i each observation was drawn from.

3.3 Gaussian Mixture Model

We consider the most simple case, a mixture of K Gaussian distributions, with $d = 1$. We have two goals:

1. Goal 1: Given n_1 samples, output $(\hat{w}_1, \hat{p}_1), \ldots, (\hat{w}_K, \hat{p}_K)$ such that $d_{TV}(\sum_{i=1}^K \hat{w}_i \hat{p}_i, \sum_{i=1}^K w_i p_i) < \epsilon$.
2. Goal 2: Given n_2 samples, output $\hat{p}$ such that $d_{TV}(p, \hat{p}) < \epsilon/2$.

Clearly, accomplishing goal 2 is easier than accomplishing goal 1. However, it turns out that the sample complexity of these two goals is the same (up to order ϵ). It is the *time* complexity of goal 1 that is much greater than that of goal 2.To be more precise, the time complexity of goal 2 is *linear* in the sample size, whereas for goal 1, there is no known algorithm that does better than solving it in time that is *exponential in K*.

Claim 1. $n_1 \leq n_2$

Proof. Let n_2 be such that $d_{TV}(p, \hat{p}) < \epsilon/2$.
Once we have our distribution $\hat{p}$ the problem simply becomes approximating $\hat{p}$ through a mixture of K Gaussians. That is, we can construct all possible mixtures of K Gaussians by going through all possible tuples $(w_i, p_i), i \in [K]$, and find the mixture which best approximates $\hat{p}$.
Let p^* be that mixture, with $d_{TV}(p^*, \hat{p}) < \epsilon/2$. By the triangle inequality, $d_{TV}(p^*, p) < \epsilon$. □

We do not show the following claim:

Claim 2. *The sample complexity of goals 1 and 2 is* $\tilde{\Theta}(\frac{K}{\epsilon^2})$.

Appendix

We used the following facts in our proof of the lower bound for uniformity testing:

- For any two distributions p, q: $\chi^2(p, q) \geq \mathcal{D}(p, q) \geq 2 \cdot d_{TV}(p, q)^2$.

- An alternative expression for $\chi^2(p, q)$:

$$\begin{aligned}\chi^2(p, q) &= \sum_x \frac{(p(x) - q(x))^2}{q(x)} \\ &= \sum_x \left[\frac{p(x)^2}{q(x)}\right] - 1 \\ &= \mathbb{E}_p\left[\frac{p(x)}{q(x)}\right] - 1\end{aligned}$$

- Suppose $X \sim Poi(\lambda)$, and $a > 0$. Then, $\mathbb{E}[a^X] = e^{\lambda(a-1)}$.