**ECE 6980**
**An Algorithmic and Information-Theoretic Toolbox for Massive Data**

Instructor: Jayadev Acharya — Lecture #3
Scribe: Ibrahim Issa — September 1st, 2016

Please send errors to acharya@cornell.edu

We showed last time that the ML estimator can learn distributions in $\Delta_k$ (up to $\epsilon - d_{TV}$) using $k/\epsilon^2$ samples. We also showed that is this tight (up to a constant) for Bernoulli distributions, i.e., any estimator needs at least $c/\epsilon^2$ samples to learn distributions in $\Delta_2$, for some constant $c$.

# 1 Lower Bound for $\Delta_k$

**Theorem 1.**

$$n^\star(\Delta_k, \epsilon) \geq c\frac{k}{\epsilon^2}, \tag{1}$$

*for some constant $c$ (independent of $k$ and $\epsilon$).*

**Remark 2.** *Recall that $n^\star(\Delta_k, \epsilon)$ is the minimum number of samples needed to learn distributions in $\Delta_k$ up to $\epsilon - d_{TV}$ with error probability at most 1/4. The definition can be found in the previous lecture.*

To prove the lower bound for $\Delta_2$, we exhibited two distributions (in particular $Ber(1/2)$ and $Ber(1/2+\epsilon)$) that are hard to distinguish: that is, if we knew the samples were coming from one of these two distributions, and we only needed to *test* which is the true one, we needed $\Omega(k/\epsilon^2)$ samples — which gives a lower bound on complexity of the harder problem of learning.

We will adopt the same strategy for proving lower bounds on $\Delta_k$. We will exhibit a collection of distributions $\{P_1, P_2, \cdots, P_M\} \subset \Delta_k$ that are hard to distinguish in a testing problem. More precisely, let $\{P_1, P_2, \cdots, P_M\}$ satisfy:

1. $d_{TV}(Pi, Pj) > 2\epsilon$, for all $i \neq j$,
2. $D(P_i||P_j) < \beta$, for all $(i, j)$,

where $\beta$ is small, say $\beta = c\epsilon^2$.

## 1.1 Testing Problem

Consider the following testing problem.

1. Pick a distrubtion $P \in \{P_1, P_2, \cdots, P_M\}$ uniformly at random. Let $i^\star$ be the chosen index.
2. Generate $n$ samples from $P_{i^\star}$.
3. Decide $\hat{i}$.

Let $n_T$ be the minimum number of samples needed for a success probability $> 3/4$. Clearly $n_T \leq n^\star(\Delta_k, \epsilon)$.

By *Fano's inequality*:

$$\Pr(\text{error}) > 1 - \frac{n\beta + \log 2}{\log(M-1)}. \tag{2}$$

So, (roughly) we need $n\beta/\log(M) >$ some constant, i.e., $n > \log(M)/\beta$.

To get our desired result, we need $M = \exp\{ck\}$. To build such a collection of distrbutions (that is large and such that any two distributions are far apart), we borrow ideas from Coding Theory where the goal is to build large codes that also have large minimum distance.

Let $\mathcal{C} \subset \{0,1\}^k$ be binary code. The minimu distance is defined as:

$$d_{\min}(\mathcal{C}) := \min_{\bar{a},\bar{b}\in\mathcal{C}} d_H(\bar{a},\bar{b}), \tag{3}$$

where $d_H$ is the Hamming distance.

**Claim:** Can construct $\mathcal{C} \subset \{0,1\}^k$ such that

1. $|\mathcal{C}| > 2^{k/8}$,
2. $d_{min}(\mathcal{C}) > k/8$,
3. For any $\bar{a} \in \mathcal{C}$ , $|\{i : \bar{a}_i = 0\}| = k/2$.

*Proof sketch:* The number of sequence of weight $k/2$ is $\begin{pmatrix} k \\ k/2 \end{pmatrix} \geq 2^k/k$. Pick a sequence of weight $k/2$ and throw away all sequence that are too close in Hamming distance. There are at most (roughly) $2^{7k/8}$ such sequences. Thus, can repeat $2^{k/8}$ times.

Now, let $U$ be the uniform distribution over $\{1, 2, \cdots, K\}$. We will use the code to perturb this distribution. In particular, let $p^+ = (1 + 16\epsilon)/k$, and $p^- = (1 - 16\epsilon)/k$. Then, with each codeword $\bar{c}$, we assign the following probability distribution:

$$P_{\bar{c}}(i) = \begin{cases} p^+, & \text{if } \bar{c}(i) = 0 \\ p^-, & \text{if } \bar{c}(i) = 1. \end{cases} \qquad i = 1, 2, \cdots, k. \tag{4}$$

It is straightforward to check that is a valid probability distribution (this comes from the weight of the codewords being $k/2$). Moreover,

$$d_{TV}(P_{\bar{a}}, P_{\bar{b}}) = d_H(\bar{a},\bar{b}) \times 16\epsilon/k > (k/8)(16\epsilon/k) = 2\epsilon. \tag{5}$$

*Exercise:* Show that $D(P_{\bar{a}}, P_{\bar{b}}) < c'\epsilon^2$, for some constant $c'$.

The rest of the lecture was a review of basic information theory. These can be found in Elements of Information Theory (Cover and Thomas) which is available online through the Cornell library.