1 Introduction

Let $\Delta_K = \{\text{distributions defined over } [K]\}$. $n_{\text{learn}}^*$ is the minimum number of samples needed to learn a distribution $p \in \Delta_K$, while $n_{\text{test}}^*$ is the minimum number of samples to distinguish between $M$ distributions in $\Delta_K$. We have found the upper-bounds for $n_{\text{learn}}^*$ and $n_{\text{test}}^*$ so far. For the following two lectures, we will focus on the lower-bounds.

Previously we proved $O(K^{\epsilon_2})$ is an upper-bound for $n_{\text{learn}}^*$; however, we have no idea about the tightness of this upper-bound. In this lecture, we show the lower-bound for $n_{\text{learn}}^*$ is $\Omega(K^{\epsilon_2})$, which means both the upper-bound and lower-bound are tight.

2 Fano’s Inequality Revisited

2.1 General Case

Suppose we have a Markov chain $X \rightarrow Y \rightarrow \hat{X}$, where $X$ is a random variable over $[K]$, $Y$ are our observations or samples, and $\hat{X}$ is the estimate for $X$. Let $p_{\text{error}} = Pr(\hat{X} \neq X)$. Then Fano’s inequality says,

$$H(X \mid Y) \leq p_{\text{error}} \cdot \log K + \log 2$$

or equivalently,

$$I(X;Y) \geq H(X) - p_{\text{error}} \cdot \log K - \log 2$$

2.2 A special case

If $X$ is uniformly distributed over $[K]$, then $H(X) = \log K$. Substitute it into Fano’s inequality, and let $p_{\text{correct}} = 1 - p_{\text{error}}$,

$$I(X;Y) \geq p_{\text{correct}} \cdot \log K - \log 2$$

or

$$p_{\text{correct}} \leq \frac{I(X;Y) + \log 2}{\log K}$$

2.3 Fano 2.0

Testing Problem (Multiway classification):

i) Given $M$ distributions $\{p_1, p_2, \ldots, p_M\}$ in $\Delta_K$ which satisfy $\forall i, j \in [M], D(p_i, p_j) \leq \beta$

ii) sample $i^*$ uniformly from $[M]$

iii) generate samples $X_1, X_2, \ldots, X_n$ from $p_{i^*}$

iv) predict $\hat{i}$ such that $P(\hat{i} \neq i^*) < 0.1$
Model the testing problem as a Markov chain \( i^* \rightarrow X \rightarrow \hat{i} \), where \( X = \{X_1, X_2, \ldots, X_n\} \).

\[
I(i^*; X) = \sum_{i^* \in [M]} \Pr(i^*) \sum_X \Pr(X | i^*) \log \frac{\Pr(X | i^*)}{\Pr(X)}
= \sum_{i^* \in [M]} \frac{1}{M} \mathcal{D}(\Pr(X | i^*), \Pr(X))
= \sum_{i=1}^{M} \frac{1}{M} \mathcal{D}(\Pr(X | i^* = l), \Pr(X))
\]

where

\[
\Pr(X | i^*) = \prod_{j=1}^{n} p_{i^*}(X_j)
\]

\[
\Pr(X) = \sum_{i^* \in [M]} \Pr(X, i^*) = \sum_{i^* \in [M]} \frac{1}{M} \Pr(X | i^*) = \sum_{k=1}^{M} \frac{1}{M} \Pr(X | i^* = k)
\]

\[
\mathcal{D}(\Pr(X | i^* = l), \Pr(X)) \leq \sum_{k=1}^{M} \frac{1}{M} \mathcal{D}(\Pr(X | i^* = l), \Pr(X | i^* = k)) \quad \text{(convexity of } \mathcal{D})
= \sum_{k=1}^{M} \frac{1}{M} \sum_{j=1}^{n} \mathcal{D}(p(X_j), p(X_j)) \quad \text{(addictivity of } \mathcal{D})
\leq \sum_{k=1}^{M} \frac{1}{M} n\beta = n\beta
\]

thus,

\[
I(i^*; X) \leq \sum_{i=1}^{M} \frac{1}{M} n\beta = n\beta
\]

According to Fano’s inequality, we have

\[
p_{\text{correct}} \leq \frac{n\beta + \log 2}{\log M}
\]

For convenience, we call the above inequality Fano 2.0.

### 3 Learning is Harder than Testing

In this section, we show that \( n_{\text{learn}}^* \geq n_{\text{test}}^* \), which can be intuitively explained as 'Learning is harder than testing in terms of sample complexity'.

#### 3.1 Description of the Learning and Testing Problem

First, we give a short description of the learning and testing problem we would like to solve.

* Learning: learn \( \hat{p} \) such that w.p. > 0.9, \( d_{TV}(p, \hat{p}) < \epsilon \)
* Testing: suppose \( p_1, p_2, \ldots, p_M \) satisfy \( d_{TV}(p_i, p_j) > 3\epsilon \), \( \forall j \neq i \), the goal is to identify the right distribution

Note that the distributions we want to learn or identify are all defined over \([K]\).
3.2 Solving Testing through Learning

The method to prove $n_{\text{learn}} \geq n_{\text{test}}$ is to show that we can actually solve the testing problem through learning. Put it another way, $n_{\text{learn}}$ samples are sufficient for the testing problem, and as a result, $n_{\text{learn}} \geq n_{\text{test}}$.

**Algorithm:** Let $p_{i^*}$ be the chosen distribution in testing problem. We first estimate $p_{i^*}$ by some learning algorithm. Denote the estimated distribution as $\hat{p}$. Then we output $\arg\min_{j \in [M]} d_{TV}(p_j, \hat{p})$ as the solution for the testing problem.

**Proof of Correctness:** Learning algorithm ensures that w.p. > 0.9, $d_{TV}(p_{i^*}, \hat{p}) < \epsilon$. For any $j \neq i^*$, we have

$$d_{TV}(p_j, \hat{p}) \geq d_{TV}(p_j, p_{i^*}) - d_{TV}(\hat{p}, p_{i^*}) \geq 3\epsilon - \epsilon = 2\epsilon$$

thus we can conclude that $i^* = \arg\min_{j \in [M]} d_{TV}(p_j, \hat{p})$, which means that w.p. > 0.9, we find the right distribution.

4 Design a Hardest Possible Testing Problem

As is shown in previous section, $n_{\text{test}}$ forms a lower-bound for $n_{\text{learn}}$. We would therefore like to design a hardest possible testing problem so as to achieve as tighter a lower-bound for $n_{\text{learn}}$ as possible.

In Fano 2.0, if $\beta = c\epsilon^2$, then

$$\frac{n_{\text{test}}\epsilon^2 + \log 2}{\log M} > p_{\text{correct}} \geq 0.9$$

which means,

$$n_{\text{test}} > \frac{0.9 \log M - \log 2}{c\epsilon^2}$$

Here comes the **Intuition:** If we can design a testing problem with $M = 2^{c_1K}$, then we will achieve a lower-bound $\Omega(\frac{K}{\epsilon^2})$ for $n_{\text{test}}$, which is also a lower-bound for $n_{\text{learn}}$. Previously, we showed that $n_{\text{learn}}$ has an upper-bound $O(\frac{K}{\epsilon^2})$. As a consequence, the lower-bound and upper-bound for $n_{\text{learn}}$ will both be tight as they are equal.

The question remains to be how to design a set of distributions $\{p_1, p_2, \ldots, p_M\}$ which satisfy

$$\begin{cases} p_1, p_2, \ldots, p_M \text{ are defined over } [K] \\ d_{TV}(p_i, p_j) > 3\epsilon, \forall i \neq j \\ D(p_i, p_j) < c\epsilon^2, \forall i \neq j \\ M \text{ is exponential to } K \end{cases}$$

We borrow ideas from coding theory to accomplish our design goal. A code $\mathcal{C}$ is a subset of $\{0, 1\}^K$. $c \in \mathcal{C}$ is a binary string of length $K$. Let $c[i]$, $1 \leq i \leq K$ be its $i^{th}$ bit. The distance of a code is defined as the minimum Hamming distance between two codewords, namely

$$d(\mathcal{C}) \triangleq \min_{c, d \in \mathcal{C}} d_H(c, d)$$

$$\triangleq$$
Claim: There exists a \( C \) such that
\[
\begin{align*}
|C| &= 2^{K/2} \\
d(C) &> \frac{K}{8} \\
\forall c \in C, \left| \{ i : c[i] = 0 \} \right| &= \frac{K}{2}
\end{align*}
\]

The proof of this claim is left to reader. Next, we design a mechanism to map every codeword \( c \in C \) to a distribution \( p_c \) over \([K]\). The mapping is \( \forall i \in [K] \),
\[
p_c(i) = \begin{cases} 
\frac{1+30\epsilon}{K} & \text{if } c[i] = 1 \\
\frac{1-30\epsilon}{K} & \text{if } c[i] = 0
\end{cases}
\]

Clearly, \( p_c \) is a valid distribution as there are \( \frac{K}{2} \) zeros in \( c \). Additionally, for \( c, d \in C, c \neq d \), we have
\[
d_{TV}(p_c, p_d) \geq \frac{1}{2} d_H(c, d) \cdot \frac{60\epsilon}{K} > \frac{1}{2} \frac{K}{8} \frac{60\epsilon}{K} > 3\epsilon
\]

The proof of \( D(p_a, p_b) < c\epsilon^2 \), for some constant \( c \) is left to the reader as an exercise. Therefore, we have designed a set of distributions with all desired properties, which then can be used to form a hardest possible testing problem.

5 Appendix

1. Convexity of \( D \)
   Let \( p^{(1)}, p^{(2)}, q^{(1)}, q^{(2)} \) be 4 distributions over the same domain, and \( 0 \leq \lambda \leq 1 \). Let \( p = \lambda p^{(1)} + (1-\lambda) p^{(2)}, q = \lambda q^{(1)} + (1-\lambda) q^{(2)} \), then we have
   \[
   D(p, q) \leq \lambda D(p^{(1)}, q^{(1)}) + (1-\lambda) D(p^{(2)}, q^{(2)})
   \]

2. Additivity of \( D \)
   Let \( P, Q \) be two joint distributions with independent marginals defined over domain \( \mathcal{X} \times \mathcal{Y} \), namely \( p(x, y) = p(x)p(y), q(x, y) = q(x)q(y) \), then we have
   \[
   D(p(x, y), q(x, y)) = D(p(x), q(x)) + D(p(y), q(y))
   \]

Remark: The proof for the above two properties is not hard, so I omit it here for simplicity. Also note that they can be easily generalized to the case of \( n \) distributions.