Lecture 3: Fano's Inequality and Applications

1 Introduction

So far, we have focused on problems concerning the *sufficiency* of sample sizes. For instance, we might ask for the minimum number of samples n required for an algorithm like Empirical Risk Minimization (ERM) to work with a certain guarantee.

Now, we shift our perspective to ask a different question: what happens if n is not sufficient? We want to determine the *necessary* number of samples for *any* learning algorithm to succeed. This involves establishing lower bounds on the sample complexity.

The general setting we consider is as follows:

1. Nature picks a "true" hypothesis $\tilde{f}$ from a finite hypothesis class $\mathcal{H}$.
2. A dataset S of size n is generated, conditioned on the choice of $\tilde{f}$.
3. A learner observes the dataset S and produces an estimate $\hat{f} \in \mathcal{H}$. This process defines a Markov Chain: $\tilde{f} \to S \to \hat{f}$.

Our goal is to understand the conditions under which the probability of making a mistake, $\mathbb{P}\left[\hat{f} \neq \tilde{f}\right]$, is high, regardless of the learning algorithm used.

2 Information Theory Basics

2.1 Entropy

Definition 1 (Entropy)**.** *The entropy of a discrete random variable X with support $\mathcal{X}$ and probability mass function $p(x)$ is defined as:*

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log(p(x)) \,.$$

Entropy measures the average uncertainty of a random variable. It has the following properties:

1. $H(X) \geq 0$.
2. $H(X) \leq \log |\mathcal{X}|$.

Proof of property 2. We use Jensen's inequality, which states that for a concave function ϕ, we have

$\mathbb{E}[\phi(Y)] \leq \phi(\mathbb{E}[Y])$. The logarithm function is concave.

$$
\begin{aligned}
H(X) = -\sum_{x\in\mathcal{X}} p(x)\log(p(x)) &= \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{1}{p(x)}\right) = \mathbb{E}_{X\sim p}\left[\log\left(\frac{1}{p(X)}\right)\right] \\
&\leq \log\left(\mathbb{E}_{X\sim p}\left[\frac{1}{p(X)}\right]\right) \quad \text{(by Jensen's inequality)} \\
&= \log\left(\sum_{x\in\mathcal{X}} p(x)\frac{1}{p(x)}\right) = \log\left(\sum_{x\in\mathcal{X}} 1\right) = \log|\mathcal{X}| \, .
\end{aligned}
$$

□

Definition 2 (Conditional Entropy)**.** *The conditional entropy of a random variable Y given X is defined as:*

$$
\begin{aligned}
H(Y|X) = \sum_{x\in\mathcal{X}} &\mathcal{P}_X(x) H(Y|X=x) \\
&= -\sum_{x\in\mathcal{X}} \mathcal{P}_X(x) \sum_{y\in\mathcal{Y}} \mathcal{P}_{Y|X}(y|x)\log \mathcal{P}_{Y|X}(y|x) \\
&= -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \mathcal{P}_{XY}(x,y)\log \mathcal{P}_{Y|X}(y|x) \, .
\end{aligned}
$$

Fact 1 (Chain Rule for Entropy)**.** *The joint entropy of two random variables X and Y can be expressed as:*

$$H(X,Y) = H(X) + H(Y|X) \, .$$

Similarly, for three variables: $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$.

2.2 Mutual Information

Definition 3 (Mutual Information)**.** *The mutual information between two random variables X and Y is defined as:*

$$I(X,Y) = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \mathcal{P}_{XY}(x,y)\log\frac{\mathcal{P}_{XY}(x,y)}{\mathcal{P}_X(x)\mathcal{P}_Y(y)} \, .$$

Mutual information measures the reduction in uncertainty about one random variable given knowledge of another. It has the following key properties:

1. $I(X,Y) \geq 0$.
2. $I(X,Y) = 0$ if and only if X and Y are independent.
3. $I(X,Y) = H(X) - H(X|Y)$.

Fact 2 (Conditioning Reduces Entropy)**.** *For any two random variables X and Y, we have:*

$$H(X|Y) \leq H(X) \, .$$

This follows directly from the properties $I(X,Y) = H(X) - H(X|Y)$ and $I(X,Y) \geq 0$.

Definition 4 (Conditional Mutual Information)**.** *The mutual information between X and Y conditioned on a third variable Z is:*

$$I(X, Y|Z) = H(X|Z) - H(X|Y, Z) .$$

Fact 3 (Chain Rule for Mutual Information)**.**

$$I(X, (Y, Z)) = I(X, Y) + I(X, Z|Y) .$$

Proof of fact 3. The definition of mutual information is $I(X, Y) = H(X) - H(X|Y)$. The definition of conditional mutual information is $I(X, Y|Z) = H(X|Z) - H(X|Y, Z)$.

$$\begin{aligned}
\text{RHS} &= I(X, Y) + I(X, Z|Y) \\
&= [H(X) - H(X|Y)] + [H(X|Y) - H(X|Y, Z)] \quad \text{(substituting the definitions)} \\
&= H(X) - H(X|Y) + H(X|Y) - H(X|Y, Z) \\
&= H(X) - H(X|Y, Z) \quad \text{(canceling terms)} \\
&= I(X, (Y, Z)) = \text{LHS} \quad \text{(by the definition of mutual information)}
\end{aligned}$$

□

3 Markov Chains and the Data Processing Inequality

Definition 5 (Markov Chain)**.** *Random variables X, Y, Z are said to form a Markov Chain, denoted $X \to Y \to Z$, if their joint probability distribution can be written as:*

$$\mathcal{P}_{XYZ}(x, y, z) = \mathcal{P}_X(x)\mathcal{P}_{Y|X}(y|x)\mathcal{P}_{Z|Y}(z|y) .$$

This is equivalent to the statement that X and Z are conditionally independent given Y, which implies $I(X, Z|Y) = 0$.

Theorem 1 (Data Processing Inequality)**.** *If $X \to Y \to Z$ form a Markov Chain, then*

$$I(X, Z) \leq I(X, Y) .$$

Intuitively, no amount of processing on Y (to get Z) can increase the information that Y contains about X.

Proof. By the chain rule for mutual information, we have two ways to expand $I(X, (Y, Z))$:

$$\begin{aligned}
I(X, (Y, Z)) &= I(X, Y) + I(X, Z|Y) \\
&= I(X, Z) + I(X, Y|Z)
\end{aligned}$$

Since $X \to Y \to Z$ is a Markov chain, we know $I(X, Z|Y) = 0$. Therefore:

$$I(X, Y) = I(X, Z) + I(X, Y|Z)$$

Because mutual information is non-negative, $I(X, Y|Z) \geq 0$. Thus, we conclude that $I(X, Y) \geq I(X, Z)$. □

4 Fano's Inequality

Fano's inequality provides a lower bound on the probability of error for any estimator in a classification problem. It connects the probability of error with the conditional entropy of the true hypothesis given the data.

Theorem 2 (Fano's Inequality)**.** *Consider the learning setup where Nature picks a true hypothesis $\tilde{f} \in \mathcal{H}$, data S is generated conditioned on $\tilde{f}$, and a learner produces an estimate $\hat{f}$ from S. For any such estimator $\hat{f}$, the probability of error $\mathcal{P}_e = \mathbb{P}\left[\hat{f} \neq \tilde{f}\right]$ is bounded. A simplified and useful version of the inequality is:*

$$\mathcal{P}_e \log|\mathcal{H}| \geq H(\tilde{f}|S) - \log(2) .$$

Corollary 1 (Simplified Fano's Inequality)**.** *Let the true hypothesis $\tilde{f}$ be chosen uniformly at random from the hypothesis class $\mathcal{H}$. Then $H(\tilde{f}) = \log|\mathcal{H}|$. Using the relation $H(\tilde{f}|S) = H(\tilde{f}) - I(\tilde{f}, S)$, we can rearrange the inequality to get:*

$$\mathbb{P}\left[\hat{f} \neq \tilde{f}\right] \geq 1 - \frac{I(\tilde{f}, S) + \log 2}{\log|\mathcal{H}|} .$$

This form is particularly useful. To get a high probability of error (i.e., a lower bound close to 1), we need to show that the mutual information $I(\tilde{f}, S)$ is small compared to $\log|\mathcal{H}|$.

5 A Lower Bound on Sample Complexity

Our goal is to construct a learning problem where any algorithm must fail. To do this using Fano's inequality, we need to find an upper bound on the mutual information $I(\tilde{f}, S)$.

5.1 Kullback-Leibler (KL) Divergence

Definition 6 (KL Divergence)**.** *Let $\mathcal{P}$ and $\mathcal{Q}$ be two probability distributions on the same support $\mathcal{X}$, with probability mass functions $p(x)$ and $q(x)$ respectively. The KL-divergence between $\mathcal{P}$ and $\mathcal{Q}$ is defined as:*

$$KL(\mathcal{P}||\mathcal{Q}) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} .$$

Fact 4 (Additivity of KL Divergence)**.** *Let $\mathcal{P}_{XY}$ and $\mathcal{Q}_{XY}$ be product distributions, i.e., $\mathcal{P}_{XY} = \mathcal{P}_X \mathcal{P}_Y$ and $\mathcal{Q}_{XY} = \mathcal{Q}_X \mathcal{Q}_Y$. Then:*

$$KL(\mathcal{P}_{XY}||\mathcal{Q}_{XY}) = KL(\mathcal{P}_X||\mathcal{Q}_X) + KL(\mathcal{P}_Y||\mathcal{Q}_Y) .$$

5.2 Bounding Mutual Information via KL Divergence

The mutual information can be expressed in terms of KL divergence. When $\tilde{f}$ is uniform over $\mathcal{H}$:

$$\begin{aligned} I(\tilde{f}, S) &= \sum_{\tilde{f} \in \mathcal{H}} \sum_{S} \mathcal{P}_{\tilde{f},S}(\tilde{f}, S) \log \frac{\mathcal{P}_{\tilde{f},S}(\tilde{f}, S)}{\mathcal{P}_{\tilde{f}}(\tilde{f})\, \mathcal{P}_S(S)} \\ &= \frac{1}{|\mathcal{H}|} \sum_{f \in \mathcal{H}} \sum_{S} \mathcal{P}_{S|f}(S) \log \frac{\mathcal{P}_{S|f}(S)}{\mathcal{P}_S(S)} \\ &= \frac{1}{|\mathcal{H}|} \sum_{f \in \mathcal{H}} \mathrm{KL}\big(\mathcal{P}_{S|f} \,\|\, \mathcal{P}_S\big). \end{aligned}$$

Using the convexity of KL-divergence and expanding the $\mathcal{P}_S$ in the denominator, one can further show that:

$$I(\tilde{f}, S) \leq \frac{1}{|\mathcal{H}|^2} \sum_{f \in \mathcal{H}} \sum_{f' \in \mathcal{H}} \mathrm{KL}\big(\mathcal{P}_{S|f} \,\|\, \mathcal{P}_{S|f'}\big).$$

Since the dataset S consists of n i.i.d. samples, by the additivity of KL divergence, this becomes:

$$I(\tilde{f}, S) \leq \frac{n}{|\mathcal{H}|^2} \sum_{f \in \mathcal{H}} \sum_{f' \in \mathcal{H}} \mathrm{KL}\big(\mathcal{P}_{X,Y|f} \,\|\, \mathcal{P}_{X,Y|f'}\big).$$

5.3 The Main Result

Theorem 3. *Nature picks a "true" hypothesis $\tilde{f}$ uniformly at random from $\mathcal{H}$. Then a dataset of n iid samples is generated conditioned on the choice of $\tilde{f}$. The learner infers $\hat{f}$ from the data. There exists a specific learning problem and data distribution such that if the number of samples n satisfies:*

$$n < \frac{\log(|\mathcal{H}|)/2 - \log 2}{48\epsilon^2}$$

for a fixed $\epsilon \in (0, 1/8)$, the learning fails, i.e. $\mathbb{P}\left[\hat{f} \neq \tilde{f}\right] \geq \frac{1}{2}$, for any mechanism that a learner could use for picking $\hat{f}$.

Proof. We construct a "hard" learning problem.

1. **Constructing the hypothesis class**: Let the input space be $\mathcal{Z} = \{z_1, \ldots, z_d\}$. For each vector $\tau \in \{-1, 1\}^d$, define a hypothesis $h_\tau : \mathcal{Z} \to \{-1, 1\}$ such that $h_\tau(z_i) = \tau_i$. The hypothesis class is $\mathcal{H} = \{h_\tau : \tau \in \{-1, 1\}^d\}$, so $|\mathcal{H}| = 2^d$.
2. **Defining the data distribution**: Nature picks a τ uniformly at random, setting the true hypothesis $\tilde{f} = h_\tau$. For each sample (x_i, y_i) in the dataset S:
   - x_i is chosen uniformly from $\mathcal{Z}$.
   - The label y_i is a noisy version of the true label $\tilde{f}(x_i)$. Specifically, if $x_i = z_j$,

$$\mathbb{P}\left[y_i = y | x_i = z_j, \tilde{f} = h_\tau\right] = \begin{cases} \frac{1}{2} + 2\epsilon & \text{if } y = \tau_j \\ \frac{1}{2} - 2\epsilon & \text{if } y = -\tau_j \end{cases}.$$

3. **Bounding KL divergence**: For any two distinct hypotheses $f, f' \in \mathcal{H}$, we compute the KL divergence between their corresponding data distributions. Since $P_{X|f}$ is uniform and independent of f, the KL divergence simplifies:

$$\begin{aligned} KL(\mathcal{P}_{X,Y|f} \| \mathcal{P}_{X,Y|\bar{f}}) &= \sum_{y \in \{-1,1\}} \sum_{x \in \mathcal{Z}} P_{x,y|f} \log \frac{P_{x,y|f}}{P_{x,y|\bar{f}}} \\ &= \sum_{y \in \{-1,1\}} \sum_{x \in \mathcal{Z}} P_{x|f} P_{y|x,f} \log \frac{P_{y|x,f}}{P_{y|x,\bar{f}}} \\ &= \frac{1}{d} \sum_{y \in \{-1,1\}} \sum_{x \in \mathcal{Z}} P_{y|x,f} \log \frac{P_{y|x,f}}{P_{y|x,\bar{f}}} \\ &\leq \frac{d}{d} \sum_{y \in \{-1,1\}} P_{y|x=\tilde{x},f} \log \frac{P_{y|x=\tilde{x},f}}{P_{y|x=\tilde{x},\bar{f}}} \end{aligned}$$

The last inequality is due to the fact that $P_{y|x=\tilde{x},f}(\cdot)$ and $P_{y|x=\tilde{x},\bar{f}}(\cdot)$ differ only when $f(\tilde{x}) \neq \bar{f}(\tilde{x})$ for some $\tilde{x} \in \mathcal{Z}$. Also, notice that this can only happen for a maximum of d samples in $\mathcal{Z}$. Furthermore, due to our construction:

$$\begin{aligned} \sum_{y \in \{-1,1\}} P_{y|x=\tilde{x},f}(\cdot) \log \frac{P_{y|x=\tilde{x},f}(\cdot)}{P_{y|x=\tilde{x},\bar{f}}(\cdot)} &= \left(\frac{1}{2} + 2\epsilon\right) \log \left(\frac{\frac{1}{2} + 2\epsilon}{\frac{1}{2} - 2\epsilon}\right) + \left(\frac{1}{2} - 2\epsilon\right) \log \left(\frac{\frac{1}{2} - 2\epsilon}{\frac{1}{2} + 2\epsilon}\right) \\ &\leq 48\epsilon^2 \quad \text{for } \epsilon \in (0, 1/8) \,. \end{aligned}$$

4. **Bound mutual information**: Plugging this into our bound for $I(\tilde{f}, S)$:

$$\begin{aligned} I(\tilde{f}, S) &\leq \frac{1}{|\mathcal{H}|^2} \sum_{f \in \mathcal{H}} \sum_{\bar{f} \in \mathcal{H}} KL(\mathcal{P}_{S|f} \| \mathcal{P}_{S|\bar{f}}) \\ &= \frac{n}{|\mathcal{H}|^2} \sum_{f \in \mathcal{H}} \sum_{\bar{f} \in \mathcal{H}} KL(\mathcal{P}_{X,Y|f} \| \mathcal{P}_{X,Y|\bar{f}}) \leq \frac{n}{|\mathcal{H}|^2} \sum_{f,f' \in \mathcal{H}} 48\epsilon^2 = n \cdot 48\epsilon^2 \,. \end{aligned}$$

5. **Apply Fano's inequality**:

$$\begin{aligned} \mathbb{P}\left[\hat{f} \neq \tilde{f}\right] &\geq 1 - \frac{I(\tilde{f}, S) + \log 2}{\log |\mathcal{H}|} \\ &\geq 1 - \frac{48n\epsilon^2 + \log 2}{d \log 2} \,. \end{aligned}$$

We want this probability to be at least $\frac{1}{2}$.

$$1 - \frac{48n\epsilon^2 + \log 2}{d \log 2} > \frac{1}{2} \implies \frac{d \log 2}{2} > 48n\epsilon^2 + \log 2 \implies n < \frac{d \log(2)/2 - \log 2}{48\epsilon^2}$$

Since $\log |\mathcal{H}| = d \log 2$, this proves the theorem.

□

This result provides a fundamental limit on learnability, showing that if the sample size is too small relative to the complexity of the hypothesis class (measured by $\log |\mathcal{H}|$) and the difficulty of the problem (inversely related to ϵ), no algorithm can be guaranteed to succeed.

5.4 Connection to PAC Learning

The previous subsection demonstrated that it is possible to construct instances where the probability that $\hat{f} \neq \tilde{f}$ remains bounded away from zero. However, this fact alone does not immediately imply a poor generalization (or risk) bound in the PAC framework. In this subsection, we provide an informal proof-sketch to connect the earlier result to the PAC setting.

We assume that nature picks the $\tilde{f} = h_\tau$ as the "true" hypothesis. We denote the corresponding induced joint distribution on (x, y) as $\mathcal{P}^\tau_{X,Y}$. One can compute the risk of any hypothesis $h \in \mathcal{H}$ as:

$$
\begin{aligned}
R(h) &= \mathbb{E}_{(x,y)\sim\mathcal{P}^\tau_{X,Y}}\left[\mathbb{1}(h(x) \neq y)\right] \\
&= \sum_{y\in\{-1,1\}} \sum_{x\in\mathcal{Z}} \mathcal{P}^\tau_{X,Y}(x,y)\mathbb{1}(h(x) \neq y) \\
&= \sum_{y\in\{-1,1\}} \sum_{i=1}^{d} \mathcal{P}^\tau_{X}(z_i)\mathcal{P}^\tau_{Y|X=z_i}(y)\mathbb{1}(h(z_i) \neq y) \\
&= \frac{1}{d}\sum_{i=1}^{d} \Big(\mathcal{P}^\tau_{Y|X=z_i}(y=\tau_i)\mathbb{1}(h(z_i) \neq \tau_i) + \mathcal{P}^\tau_{Y|X=z_i}(y=-\tau_i)\mathbb{1}(h(z_i) = \tau_i)\Big) \\
&= (\frac{1}{2} - 2\epsilon) + \frac{4\epsilon}{d}\sum_{i=1}^{d} \mathbb{1}(h(z_i) \neq \tau_i)
\end{aligned}
$$

Notice that $R(h)$ is minimized by picking $h = h_\tau$. Furthermore, one can also show that $\sum_{i=1}^{d} \mathbb{1}(h(z_i) \neq \tau_i) = \Omega(d)$ in expectation.

Therefore, for any $h \neq h_\tau$,

$$
R(h) > R(h_\tau) + \Omega(\epsilon)\ .
$$

Disclaimer: These notes have not been scrutinized with the level of rigor usually applied to formal publications. Readers should verify the results before use.