1 Information Theory

First, we provide some information theory background.

**Definition 3.1 (Entropy).** The entropy of a discrete random variable $x$ of support $\mathcal{X}$ and probability mass function $p$ is defined as:

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

A basic property of the entropy of a discrete random variable $x$ is that:

$$0 \leq \mathbb{H}(x) \leq \log |\mathcal{X}|$$

In fact, the entropy is maximal for the discrete uniform distribution. That is, $(\forall x \in \mathcal{X}) p(x) = 1/|\mathcal{X}|$, in which case $\mathbb{H}(x) = \log |\mathcal{X}|$.

**Definition 3.2 (Conditional entropy).** The conditional entropy of $y$ given $x$ is defined as:

$$\mathbb{H}(y|x) = \sum_{v \in \mathcal{X}} p_x(v) \mathbb{H}(y|x = v)$$

$$= - \sum_{v \in \mathcal{X}} p_x(v) \sum_{y \in \mathcal{Y}} p_{y|x}(y|v) \log p_{y|x}(y|v)$$

$$= - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p_{xy}(x, y) \log p_{y|x}(y|x)$$

The conditional entropy can be expressed in terms of the entropy:

$$\mathbb{H}(y|x) = \mathbb{H}(x, y) - \mathbb{H}(x)$$

**Definition 3.3 (Mutual information).**

$$\mathbb{I}(x, y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p_{xy}(x, y) \log \frac{p_{xy}(x, y)}{p_x(x)p_y(y)}$$
A basic property of the mutual information of random variables $x$ and $y$ is that:

$$I(x, y) \geq 0$$

Furthermore, the mutual information can be expressed in terms of the entropy:

$$I(x, y) = H(x) - H(x|y)$$

Note that random variables $x$ and $y$ are independent if and only if $I(x, y) = 0$.

**Definition 3.4** (Conditional mutual information).

$$I(x, y|z) = H(x|z) - H(x|y, z)$$

**Definition 3.5** (Markov chain). Random variables $x$, $y$ and $z$ are said to form a Markov chain $x \rightarrow y \rightarrow z$ if and only if their joint probability distribution can be written as:

$$p_{xyz}(x, y, z) = p_x(x)p_y(x|y)p_z(z|y)$$

Equivalently, random variables $x$, $y$ and $z$ are said to form a Markov chain $x \rightarrow y \rightarrow z$ if and only if $x$ and $z$ are conditionally independent given $y$, and thus $I(x, z|y) = 0$.

## 2 Fano’s inequality

Fano’s inequality allows to provide information-theoretic lower bounds on the sample complexity. The setting for the analysis is as follows. Nature picks a “true” hypothesis $f$ from some distribution of hypotheses. Then, a dataset $S$ of $n$ samples is produced, conditioned on the choice of $f$. The learner then infers $\hat{f}$ from the dataset $S$. The probability of error of the learner is given by $P[\hat{f} \neq f]$. By lower-bounding this probability of error, one can find the necessary number of samples for learning. (Analyses as in the previous lecture allows to find a sufficient number of samples.)

**Theorem 3.1** (Fano’s inequality). For any random variable $\hat{f}$ with $k$ possible outcomes, such that $\overline{f} \rightarrow S \rightarrow \hat{f}$, we have:

$$P[\hat{f} \neq f] \geq \frac{H(\overline{f}|S) - \log 2}{\log k}$$

(See [1] if interested in the proof.)

**Corollary 3.1** (Fano’s inequality). For any random variable $\hat{f}$ with $k$ possible outcomes, such that $\overline{f} \rightarrow S \rightarrow \hat{f}$, where $\overline{f}$ is chosen by nature uniformly at random (also from $k$ possible outcomes), we have:

$$P[\hat{f} \neq f] \geq 1 - \frac{I(\overline{f}, S) + \log 2}{\log k}$$
Proof. By property of the mutual information, we have \( \mathbb{H}(\mathcal{F}|S) = \mathbb{H}(\mathcal{F}) - I(\mathcal{F}, S). \) Since \( \mathcal{F} \) is chosen uniformly at random from \( k \) possible outcomes, then \( \mathbb{H}(\mathcal{F}) = \log k \) and we prove our claim.

The key in using Fano’s inequality is to define a hypothesis class \( \mathcal{F} \) for which \( k = |\mathcal{F}| \) is large, while the mutual information \( I(\mathcal{F}, S) \) is small and of order \( n \).

3 Upper Bounds on the Mutual Information

One key step in the application of Fano’s inequality is to upper-bound the mutual information \( I(\mathcal{F}, S) \). Next, we revise some important definitions and inequalities from information theory.

**Definition 3.6 (Kullback-Leibler (KL) divergence).** Assume that a random variable \( x \) has support \( X \). Assume that there are two probability density functions \( p \) and \( q \), which define two probability distributions \( P = p(\cdot) \) and \( Q = q(\cdot) \) respectively. The KL divergence is defined as:

\[
\text{KL}(P||Q) = \int_{x \in X} p(x) \log \frac{p(x)}{q(x)} dx
\]

One important property of the KL divergence for independent random variables is the following. Let \( P_{xy} = p_{xy}(\cdot) \) and \( P_x P_y = p_x(\cdot)p_y(\cdot) \). Assume that \( x \) and \( y \) are independent, and thus \( P_{xy} = P_x P_y \) and likewise, assume that \( Q_{xy} = Q_x Q_y \). We have:

\[
\text{KL}(P_{xy}||Q_{xy}) = \text{KL}(P_x||Q_x) + \text{KL}(P_y||Q_y)
\]

(The proof of the above might be left for homework very soon.)

Let \( P_{xy} = p_{xy}(\cdot) \) and \( P_x P_y = p_x(\cdot)p_y(\cdot) \). We can define the mutual information as:

\[
I(x, y) = \text{KL}(P_{xy}||P_x P_y)
\]

\[
= \int_{x \in X, y \in Y} p_{xy}(x, y) \log \frac{p_{xy}(x, y)}{p_x(x)p_y(y)} dx dy
\]

By well-known identities \( p_{T, S}(\mathcal{F}, S) = p_T(\mathcal{F})p_s(S) \) and \( p_s(S) = \sum_{\mathcal{F} \in \mathcal{F}} p_{T,S}(\mathcal{F}, S) \), and since \( \mathcal{F} \) follows a uniform distribution \( p_T(\mathcal{F}) = 1/k \), we have:

\[
I(\mathcal{F}, S) = \sum_{\mathcal{F} \in \mathcal{F}} \int_S p_{T,S}(\mathcal{F}, S) \log \frac{p_{T,S}(\mathcal{F}, S)}{p_T(\mathcal{F})p_S(S)} dS
\]

\[
= \sum_{\mathcal{F} \in \mathcal{F}} \int_S p_T(\mathcal{F})p_S(S) \log \frac{p_{T,S}(\mathcal{F}, S)}{p_T(\mathcal{F})p_S(S)} dS
\]

\[
= \frac{1}{k} \sum_{\mathcal{F} \in \mathcal{F}} \int_S p_S(S) \log \frac{p_{S}(S)}{p_S(S)} dS
\]
\[ = \frac{1}{k} \sum_{f \in F} \text{KL}(\mathcal{P}_{S|\overline{f}} || \mathcal{P}_S) \]

In the above, we use the distribution \( \mathcal{P}_{S|\overline{f}} = \mathcal{P}_S(S) \) as well as the distribution \( \mathcal{P}_S = \frac{1}{k} \sum_{f \in F} \mathcal{P}_{S|\overline{f}}(S) \).

Furthermore, from the convexity of the KL divergence, we can show that:

\[ \mathbb{I}(\overline{f}, S) \leq \frac{1}{k^2} \sum_{f \in F} \sum_{f' \in F} \text{KL}(\mathcal{P}_{S|f} || \mathcal{P}_{S|f'}) \quad (2) \]

(The proof of the above might be left for homework very soon.)

4 Application: Empirical Risk Minimization with a Finite Hypothesis Class

Here we will prove a negative result in a setting similar to Theorem 2.1. First, some necessary definitions.

**Definition 3.7.** The multivariate normal distribution of a random vector \( x \in \mathbb{R}^k \) with mean \( \mu \in \mathbb{R}^k \) and (symmetric and positive definite) covariance \( \Sigma \in \mathbb{R}^{k \times k} \) is defined by the probability density function:

\[ p(x) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \]

For shortness, we write \( x \sim \mathcal{N}(\mu, \Sigma) \).

Let the distributions \( \mathcal{N}_1 = \mathcal{N}(\mu_1, \Sigma_1) \) and \( \mathcal{N}_2 = \mathcal{N}(\mu_2, \Sigma_2) \), then:

\[ \text{KL}(\mathcal{N}_1 || \mathcal{N}_2) = \frac{1}{2} \left( \text{tr}(\Sigma_2^{-1} \Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) - k + \log \frac{\det \Sigma_2}{\det \Sigma_1} \right) \]

Note that when \( \Sigma_1 = \Sigma_2 = I \), the KL divergence becomes:

\[ \text{KL}(\mathcal{N}_1 || \mathcal{N}_2) = \frac{1}{2} \| \mu_2 - \mu_1 \|^2 \quad (3) \]

(The proof of the above might be left for homework very soon.)

Next, we show our negative result. As we mentioned before, our main goal will be to upper-bound the mutual information \( \mathbb{I}(\overline{f}, S) \) in order to apply Fano’s inequality.

**Theorem 3.2.** Assume that nature picks a “true” hypothesis \( \overline{f} \) from some distribution of hypotheses with support \( F \) where \( |F| = k \). Then, a dataset of \( n \) samples is produced, conditioned on the choice of \( \overline{f} \). The learner then infers \( \hat{f} \) from the dataset. Under the same setting as in Theorem 2.1, there exists a
specific prediction problem and data distribution such that if \( n \leq \frac{\log k}{2} - \log 2 \), then learning fails, i.e.,

\[
\mathbb{P}[\hat{f} \neq f] \geq 1/2
\]

for any mechanism (or algorithm) that a learner could use for picking \( \hat{f} \).

**Proof.** Recall that in Theorem 2.1, we assume that \( \mathcal{F} \) is a finite set of hypotheses, i.e., \( \mathcal{F} = \{f_1 \ldots f_k\} \) where \( k < +\infty \) and \((\forall j) \ f_j : \mathcal{X} \to \mathcal{Y}\).

Here, we further assume that \( \mathcal{X} = \mathbb{R}^k \) and \( \mathcal{Y} = \{-1, +1\} \) and that \( f_j(x) \) is the sign of the \( j \)-th element of the \( k \)-dimensional vector \( x \), i.e., \( f_j(x) = \text{sgn}(x_j) \).

(For clarity, we are now using a super-index for the sample index and a sub-index for the vector entry.) Assume that nature picks a “true” hypothesis \( f \) uniformly at random from \( \mathcal{F} \). Then, a dataset \( S = (x^{(1)}, y^{(1)}, \ldots x^{(n)}, y^{(n)}) \) of \( n \) samples is produced, conditioned on the choice of \( f \).

We assume that \( \mathbb{P}[y = +1|f = f_j] = \mathbb{P}[y = -1|f = f_j] = 1/2 \). We also assume that \( x|y = +1, f = f_j \sim \mathcal{N}(\mu^{(j)}(i), 1) \) and \( x|y = -1, f = f_j \sim \mathcal{N}(-\mu^{(j)}(i), 1) \) where \( \mu^{(j)}(i) = 1[i = j] \). That is, if \( f = f_j \) then every hypothesis \( f \in \mathcal{F} \) \(-f_j\) has on average a 50% risk, since the sign of a normal random variable with mean zero is either \(+1\) or \(-1\) with 50% probability.

Note that for any pair of distributions \( \mathcal{P}_{xy} \) and \( \mathcal{Q}_{xy} \) where \( p_y(+1) = p_y(-1) = q_y(+1) = q_y(-1) = 1/2 \), we have:

\[
\mathbb{KL}(\mathcal{P}_{xy} \| \mathcal{Q}_{xy}) = \sum_{y \in \{-1,+1\}} \int p_x(y)p_{x|y}(x) \log \frac{p_y(y)p_{x|y}(x)}{q_y(y)q_{x|y}(x)} \, dx,
\]

\[
= \frac{1}{2} \left( \int p_{x|y=+1}(x) \log \frac{p_x|y=+1(x)}{q_x|y=+1(x)} \, dx + \int p_{x|y=-1}(x) \log \frac{p_x|y=-1(x)}{q_x|y=-1(x)} \, dx \right)
\]

\[
= \frac{1}{2} (\mathbb{KL}(\mathcal{P}_{x|y=+1} \| \mathcal{Q}_{x|y=+1}) + \mathbb{KL}(\mathcal{P}_{x|y=-1} \| \mathcal{Q}_{x|y=-1})) \tag{4}
\]

By eq. (2), by eq. (1) since \( S \) is a dataset of \( n \) independent samples \((x^{(i)}, y^{(i)})\) for \( i = 1 \ldots n \), by eq. (4), and by eq. (3) since \( x|y \) is normally distributed, we have:

\[
\mathbb{I}(\mathcal{F}, S) \leq \frac{1}{k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} \mathbb{KL}(\mathcal{P}_{S|f_j} \| \mathcal{P}_{S|f_{j'}})
\]

\[
= \frac{n}{k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} \mathbb{KL}(\mathcal{P}_{x,y|f_j} \| \mathcal{P}_{x,y|f_{j'}})
\]

\[
= \frac{n}{2k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} \left( \mathbb{KL}(\mathcal{P}_{x|f_j,y=+1} \| \mathcal{P}_{x|f_{j'},y=+1}) + \mathbb{KL}(\mathcal{P}_{x|f_j,y=-1} \| \mathcal{P}_{x|f_{j'},y=-1}) \right)
\]

\[
= \frac{n}{2k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} \left( \mathbb{KL}(\mathcal{N}(\mu^{(j)}, 1) \| \mathcal{N}(\mu^{(j')}, 1)) + \mathbb{KL}(\mathcal{N}(\mu^{(j)}, 1) \| \mathcal{N}(\mu^{(j')}, 1)) \right)
\]
\[
\begin{align*}
\sum_{j=1}^{k} \sum_{j'=1}^{k} & \left( \frac{1}{2} \| \mu^{(j)} - \mu^{(j')} \|^2 + \frac{1}{2} \| \mu^{(j)} - \mu^{(j')} \|^2 \right) \\
= \frac{n}{2k^2} & \sum_{j=1}^{k} \sum_{j'=1}^{k} \| \mu^{(j)} - \mu^{(j')} \|^2 \\
= \frac{n}{k^2} & \sum_{j=1}^{k} \sum_{j'=1}^{k} I[j \neq j'] \\
= \frac{n(k^2 - k)}{k^2} \\
\leq n
\end{align*}
\]

By Corollary 3.1 and assuming a probability of error of at least 1/2:

\[
P[\hat{f} \neq f] \geq 1 - \frac{I(f, S) + \log 2}{\log k} \geq 1 - \frac{n + \log 2}{\log k} = \frac{1}{2}
\]

By solving for \( n \) in the above, we obtain that if \( n \leq \frac{\log k}{2} - \log 2 \), then we have that \( P[\hat{f} \neq f] \geq 1/2 \). \( \square \)

References