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:

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

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

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

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 $H(x) = \log |\mathcal{X}|$.

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

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

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

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

**Theorem 3.1** (Chain rule for entropy).

$$H(x, y) = H(x) + H(y|x)$$

Similarly:

$$H(x, y|z) = H(x|z) + H(y|x, z)$$

(See [1] if interested in the proof.)
Definition 3.3 (Mutual information).

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

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:

\[ I(x, y) = \mathbb{H}(x) - \mathbb{H}(x|y) \]

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

Theorem 3.2 (Conditioning reduces entropy).

\[ \mathbb{H}(x|y) \leq \mathbb{H}(x) \]

Proof. \( 0 \leq I(x, y) = \mathbb{H}(x) - \mathbb{H}(x|y) \).

Definition 3.4 (Conditional mutual information).

\[ I(x, y|z) = \mathbb{H}(x|z) - \mathbb{H}(x|y, z) \]

Theorem 3.3 (Chain rule for mutual information).

\[ I(x, y, z) = I(x, y) + I(x, z|y) \]

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

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:

\[ p_{xyz}(x, y, z) = p_x(x)p_{y|x}(y|x)p_{z|y}(z|y) \]

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 Data processing inequality

The way to understand the following inequality in an intuitive fashion is as follows. Assume there is a random variable \( x \). Then we process \( x \) by possibly adding some additional randomness (e.g., noise) and/or losing some information (e.g., rounding) and we obtain a variable \( y \). Similarly, we process \( y \) by possibly adding some additional randomness (e.g., noise) and/or losing some information (e.g., rounding) and we obtain a variable \( z \). There is likely more information about \( x \) in \( y \) than in \( z \).
**Theorem 3.4** (Data processing inequality). If \( x \rightarrow y \rightarrow z \) then \( I(x, y) \geq I(x, z) \), or equivalently \( \mathbb{H}(x|y) \leq \mathbb{H}(x|z) \).

**Proof.** By using the chain rule for mutual information (Theorem 3.3), we have:

\[
I(x, y, z) = I(x, y) + I(x, z|y) = I(x, z) + I(x, y|z)
\]

Since \( I(x, z|y) = 0 \) and since \( I(x, y|z) \geq 0 \), we get:

\[
I(x, y) \geq I(x, z)
\]

The second statement is trivial since:

\[
\mathbb{H}(x) - \mathbb{H}(x|y) = I(x, y) \geq I(x, z) = \mathbb{H}(x) - \mathbb{H}(x|z)
\]

\( \Box \)

### 3 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 \( \bar{f} \) from some distribution of hypotheses. Then, a dataset \( S \) of \( n \) samples is produced, conditioned on the choice of \( \bar{f} \). The learner then infers \( \hat{f} \) from the dataset \( S \). The probability of error of the learner is given by \( \mathbb{P}[\hat{f} \neq \bar{f}] \).

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 3.5** (Fano’s inequality). For any estimator \( \hat{f} \) with \( k \) possible outcomes, such that \( \bar{f} \rightarrow S \rightarrow \hat{f} \), we have:

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

**Proof.** Define the following “error” random variable:

\[
w = \begin{cases} 
1 & \text{if } \hat{f} \neq \bar{f} \\
0 & \text{if } \hat{f} = \bar{f}
\end{cases}
\]

First, we analyze some quantities that will be useful later. Since conditioning reduces entropy (Theorem 3.2), we have that \( \mathbb{H}(w|\hat{f}) \leq \mathbb{H}(w) \leq \log 2 \). Note that \( \mathbb{H}(\bar{f}|w = 0, \hat{f}) = 0 \), since if \( w = 0 \) then \( \hat{f} = \bar{f} \). Also, note that since conditioning reduces entropy (Theorem 3.2) \( \mathbb{H}(\bar{f}|w = 1, \hat{f}) \leq \mathbb{H}(\bar{f}) \leq \log k \). By Definition 3.2, we have:

\[
\mathbb{H}(\bar{f}|w, \hat{f}) = \mathbb{P}[w = 0] \mathbb{H}(\bar{f}|w = 0, \hat{f}) + \mathbb{P}[w = 1] \mathbb{H}(\bar{f}|w = 1, \hat{f}) \leq \mathbb{P}[w = 1] \log k
\]

\[
\mathbb{P}[\hat{f} \neq \bar{f}] \geq \frac{\mathbb{H}(\bar{f}|S) - \log 2}{\log k}
\]
Finally, since $w$ is a deterministic function of $\overline{f}$ and $\hat{f}$, then $\mathbb{H}(w|\overline{f}, \hat{f}) = 0$.

By using the chain rule for entropies (Theorem 3.1), in two different ways, we have:

$$\mathbb{H}(w, \overline{f}|\hat{f}) = \mathbb{H}(\overline{f}|\hat{f}) + \mathbb{H}(w|\overline{f}, \hat{f})$$

and thus from the conclusions at the beginning of the proof, we have:

$$\mathbb{H}(\overline{f}|\hat{f}) = \mathbb{H}(\overline{f}|w, \hat{f}) - \mathbb{H}(w|\overline{f}, \hat{f})$$

$$\leq \log 2 + \mathbb{P}[\hat{f} \neq \overline{f}] \log k - 0$$

By the data-processing inequality (Theorem 3.4) and since $\overline{f} \rightarrow S \rightarrow \hat{f}$ is a Markov chain, we have $\mathbb{H}(\overline{f}|S) \leq \mathbb{H}(\overline{f}|\hat{f})$ and thus the above implies:

$$\mathbb{H}(\overline{f}|S) \leq \log 2 + \mathbb{P}[\hat{f} \neq \overline{f}] \log k$$

By rearranging the terms, we prove our claim. \hfill \Box

Corollary 3.1 (Fano’s inequality). For any estimator $\hat{f}$ 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:

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

Proof. By property of the mutual information, we have $\mathbb{H}(\overline{f}|S) = \mathbb{H}(\overline{f}) - \mathbb{I}(\overline{f}, S)$. Since $\overline{f}$ is chosen uniformly at random from $k$ possible outcomes, then $\mathbb{H}(\overline{f}) = \log k$ and we prove our claim. \hfill \Box

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 $\mathbb{I}(\overline{f}, S)$ is small and of order $n$.

4 Upper Bounds on the Mutual Information

One key step in the application of Fano’s inequality is to upper-bound the mutual information $\mathbb{I}(\overline{f}, S)$. 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 $\mathcal{X}$. Assume that there are two probability density functions $p$ and $q$, which define two probability distributions $\mathcal{P} = p(\cdot)$ and $\mathcal{Q} = q(\cdot)$ respectively. The KL divergence is defined as:

$$\mathbb{KL}(\mathcal{P}||\mathcal{Q}) = \int_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \, dx$$
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:

\[
\text{KL}(P_{xy} \| Q_{xy}) = \text{KL}(P_{x} \| Q_{x}) + \text{KL}(P_{y} \| Q_{y}) \tag{1}
\]

(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:

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

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

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

\[
\mathbb{I}(\bar{f}, S) = \sum_{\bar{f} \in F} \int_{S} p_{\bar{f}, S}(\bar{f}, S) \log \frac{p_{\bar{f}, S}(\bar{f}, S)}{p_{\bar{f}}(\bar{f})p_{S}(S)} \, dS
\]

\[
= \sum_{\bar{f} \in F} \int_{S} p_{\bar{f}}(\bar{f})p_{S|\bar{f}}(S) \log \frac{p_{\bar{f}}(\bar{f})p_{S|\bar{f}}(S)}{p_{\bar{f}}(\bar{f})p_{S}(S)} \, dS
\]

\[
= \frac{1}{k} \sum_{\bar{f} \in F} \int_{S} p_{S|\bar{f}}(S) \log \frac{p_{S|\bar{f}}(S)}{p_{S}(S)} \, dS
\]

\[
= \frac{1}{k} \sum_{\bar{f} \in F} \text{KL}(P_{S|\bar{f}} \| P_{S})
\]

In the above, we use the distribution \( P_{S|\bar{f}} = p_{S|\bar{f}}(\cdot) \) as well as the distribution \( P_{S} = p_{S}(S) = \frac{1}{k} \sum_{\bar{f} \in F} p_{S|\bar{f}}(S) \).

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

\[
\mathbb{I}(\bar{f}, S) \leq \frac{1}{k^2} \sum_{\bar{f} \in F} \sum_{\bar{f}' \in F} \text{KL}(P_{S|\bar{f}} \| P_{S|\bar{f}'}) \tag{2}
\]

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

5 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:

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

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

Let the distributions \( N_1 = N(\mu_1, \Sigma_1) \) and \( N_2 = N(\mu_2, \Sigma_2) \), then:

\[
\text{KL}(N_1 \| N_2) = \frac{1}{2} \left( \text{tr}(\Sigma_2^{-1} \Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) - k + \log \det \Sigma_2 / \det \Sigma_1 \right)
\]

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

\[
\text{KL}(N_1 \| N_2) = \frac{1}{2} \| \mu_2 - \mu_1 \|^2
\]  

(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 \( I(\mathbf{f}, S) \) in order to apply Fano’s inequality.

**Theorem 3.6.** Assume that nature picks a “true” hypothesis \( \mathbf{f} \) from some distribution of hypotheses with support \( \mathcal{F} \) where \( |\mathcal{F}| = k \). Then, a dataset of \( n \) samples is produced, conditioned on the choice of \( \mathbf{f} \). The learner then infers \( \hat{\mathbf{f}} \) from the dataset. Under the same setting as in Theorem 2.1, there exists a specific prediction problem and data distribution such that if \( n \leq \log k / 2 - \log 2 \), then learning fails, i.e.,

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

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

**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 \( \mathbf{f} \) 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 \( \mathbf{f} \).

We assume that \( P[y = +1| \mathbf{f} = f_j] = P[y = -1| \mathbf{f} = f_j] = 1/2 \). We also assume that \( x|y = +1, \mathbf{f} = f_j \sim N(\mu_j^{(i)}, I) \) and \( x|y = -1, \mathbf{f} = f_j \sim N(-\mu_j^{(i)}, I) \) where \( \mu_j^{(i)} = 1[i = j] \). That is, if \( \mathbf{f} = f_j \) then every hypothesis \( f \in \mathcal{F} - \{f_j\} \)
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:

$$\mathbb{H}(f, S) \leq \frac{1}{k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} \mathbb{H}(P_{S|f_j} \| P_{S|f'_j})$$

$$= \frac{n}{k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} \mathbb{H}(P_{x|y} | f_j \| P_{x|y} | f'_j)$$

$$= \frac{n}{2k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} \left( \mathbb{H}(P_{x|y} | f_j, y=+1 \| P_{x|f'_j, y=+1} P_{x|f'_j, y=-1} \| P_{x|f'_j, y=-1}) \right)$$

$$= \frac{n}{2k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} \left( \mathbb{H}(\mathcal{N}(\mu^{(j)}, I) \| \mathcal{N}(\mu^{(j')}, I)) + \mathbb{H}(\mathcal{N}(\mu^{(j)}, I) \| \mathcal{N}(\mu^{(j')}, I)) \right)$$

$$= \frac{n}{2k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} \left( \frac{1}{2} ||\mu^{(j)} - \mu^{(j')}||^2 + \frac{1}{2} ||\mu^{(j)} - \mu^{(j')}||^2 \right)$$

$$= \frac{n}{2k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} ||\mu^{(j)} - \mu^{(j')}||^2$$

$$= \frac{n}{k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} 1[j \neq j']$$

$$= \frac{n(k^2 - k)}{k^2} \leq n$$

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

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

\section*{References}