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:

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

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

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

**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).

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

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:

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

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

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

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

**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 \( x \to y \to z \) then \( I(x, y) \geq I(x, z) \), or equivalently \( H(x | y) \leq H(x | z) \).

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:

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

\[\square\]

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 \( \overline{f} \) from some distribution of hypotheses. Then, a dataset \( S \) of \( n \) samples is produced, conditioned on the choice of \( \overline{f} \). The learner then infers \( \hat{f} \) from the dataset \( S \). The probability of error of the learner is given by \( \mathbb{P}[\hat{f} \neq \overline{f}] \). 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 \( \hat{f} \), such that \( \overline{f} \to S \to \hat{f} \), we have:

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

Proof. Define the following “error” random variable:

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

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

\[
H(\overline{f} | w, \hat{f}) = \mathbb{P}[w = 0] H(\overline{f} | w = 0, \hat{f}) + \mathbb{P}[w = 1] H(\overline{f} | w = 1, \hat{f}) \\
\leq \mathbb{P}[w = 1] H(\overline{f}) \\
= \mathbb{P}[\hat{f} \neq \overline{f}] H(\overline{f})
\]
Finally, since \( w \) is a deterministic function of \( \mathcal{Y} \) and \( \hat{f} \), then \( \mathbb{H}(w|\mathcal{Y}, \hat{f}) = 0 \).

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

\[
\mathbb{H}(w, \mathcal{Y}|\hat{f}) = \mathbb{H}(\mathcal{Y}|\hat{f}) + \mathbb{H}(w|\mathcal{Y}, \hat{f}) = \mathbb{H}(w|\hat{f}) + \mathbb{H}(\mathcal{Y}|w, \hat{f})
\]

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

\[
\mathbb{H}(\mathcal{Y}|\hat{f}) = \mathbb{H}(w|\hat{f}) + \mathbb{H}(\mathcal{Y}|w, \hat{f}) - \mathbb{H}(w|\mathcal{Y}, \hat{f})
\]

\[
\leq \log 2 + \mathbb{P}[\hat{f} \neq \mathcal{Y}] \mathbb{H}(\mathcal{Y}) - 0
\]

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

\[
\mathbb{H}(\mathcal{Y}|S) \leq \log 2 + \mathbb{P}[\hat{f} \neq \mathcal{Y}] \mathbb{H}(\mathcal{Y})
\]

By rearranging the terms, we prove our claim. \( \square \)

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

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

**Proof.** By property of the mutual information, we have \( \mathbb{H}(\mathcal{Y}|S) = \mathbb{H}(\mathcal{Y}) - \mathbb{I}(\mathcal{Y}, S) \).

Since \( \mathcal{Y} \) is chosen uniformly at random from \( k \) possible outcomes, then \( \mathbb{H}(\mathcal{Y}) = \log k \) and we prove our claim. \( \square \)

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 \( \mathbb{I}(\mathcal{Y}, S) \) 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 \( \mathbb{I}(\mathcal{Y}, S) \). 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 \( P = p(\cdot) \) and \( Q = q(\cdot) \) respectively. The KL divergence is defined as:

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

$$\operatorname{KL}(P_{xy} \| Q_{xy}) = \operatorname{KL}(P_x \| Q_x) + \operatorname{KL}(P_y \| Q_y)$$ (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:

$$I(x, y) = \operatorname{KL}(P_{xy} \| P_x P_y) = \int_{x \in X, y \in Y} p_{xy}(x, y) \log \frac{p_{xy}(x, y)}{p_x(x) p_y(y)} dx dy$$

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

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

$$= \sum_{T \in F} \int_S p_T(T) p_{S|T}(S) \log \frac{p_T(T) p_{S|T}(S)}{p_T(T) p_S(S)} dS$$

$$= \frac{1}{k} \sum_{T \in F} \int_S p_{S|T}(S) \log \frac{p_{S|T}(S)}{p_S(S)} dS$$

$$= \frac{1}{k} \sum_{T \in F} \operatorname{KL}(P_{S|T} \| P_S)$$

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

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

$$I(T, S) \leq \frac{1}{k^2} \sum_{T \in F} \sum_{T' \in F} \operatorname{KL}(P_{S|T} \| P_{S|T'})$$ (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 $x \sim \mathcal{N}(\mu, \Sigma)$.

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

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

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

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

(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 $I(\mathcal{F}, S)$ in order to apply Fano’s inequality.

Theorem 3.6. Assume that nature picks a “true” hypothesis $\mathcal{F}$ 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 $\mathcal{F}$. 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 $n \leq \log \frac{k}{2} - \log 2$, then learning fails, i.e.,

$$\Pr[\hat{f} \neq \mathcal{F}] \geq 1/2$$

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 $(\forall j) f_j : \mathcal{X} \rightarrow \mathcal{Y}$.

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 $\mathcal{F}$ 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 $\mathcal{F}$.

We assume that $\Pr[y = +1|\mathcal{F} = f_j] = \Pr[y = -1|\mathcal{F} = f_j] = 1/2$. We also assume that $x_{y = +1, \mathcal{F} = f_j} \sim \mathcal{N}(\mu^{(j)}, I)$ and $x_{y = -1, \mathcal{F} = f_j} \sim \mathcal{N}(-\mu^{(j)}, I)$ where $\mu^{(j)}_i = 1[i = j]$. That is, if $\mathcal{F} = f_j$ 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 \( P_{xy} \) and \( Q_{xy} \) where \( p_y(+1) = p_y(-1) = q_y(+1) = q_y(-1) = 1/2 \), we have:

\[
\begin{align*}
\mathbb{KL}(P_{xy} || Q_{xy}) &= \sum_{y \in \{-1,+1\}} \int_{x \in X} p_y(y)p_{x|y}(x) \log \frac{p_y(y)p_{x|y}(x)}{q_y(y)q_{x|y}(x)} \, dx \\
&= \frac{1}{2} \left( \int_{x \in X} p_{x|y=+1}(x) \log \frac{p_{x|y=+1}(x)}{q_{x|y=+1}(x)} \, dx + \int_{x \in X} p_{x|y=-1}(x) \log \frac{p_{x|y=-1}(x)}{q_{x|y=-1}(x)} \, dx \right) \\
&= \frac{1}{2} \left( \mathbb{KL}(P_{x|y=+1} || Q_{x|y=+1}) + \mathbb{KL}(P_{x|y=-1} || Q_{x|y=-1}) \right) \\
&= \frac{1}{2} \left( \mathbb{KL}(P_{x|y=+1} || Q_{x|y=+1}) + \mathbb{KL}(P_{x|y=-1} || Q_{x|y=-1}) \right)
\end{align*}
\]

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:

\[
I(\hat{f}, S) \leq \frac{1}{k^2} \sum_{k=1}^{k} \sum_{j=1}^{k} \mathbb{KL}(P_{S|f_{j}} || P_{S|f_{j}'})
\]

\[
= \frac{n}{k^2} \sum_{k=1}^{k} \sum_{j=1}^{k} \mathbb{KL}(P_{x|y|f_{j}} || P_{x|y|f_{j}'})
\]

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

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

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

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

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

\[
= \frac{n(k^2 - k)}{k^2}
\]

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

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

References