Lecture 3: Information Theory
3.1 Entropy

**Def.** The entropy $H(X)$ of a random variable $X$ with density $p(x)$ is defined as

$$H(X) = - \int_S p(x) \ln p(x) dx,$$

where $S$ is the support of $p(x)$ (that is, the set where $p(x)$ is not zero).

Entropy depends only on the density $p(x)$ and thus entropy is sometime written as $H(p)$ rather than $H(X)$. 
3.1 Entropy

Example. Let $X$ is a Gaussian with density $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$.

\[
H(p) = -\int p \ln p \, dx
= -\int p \left[-\frac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2}\right] \, dx
= \frac{E[X^2]}{2\sigma^2} + \frac{1}{2} \ln 2\pi\sigma^2
= \frac{1}{2} + \frac{1}{2} \ln 2\pi\sigma^2
= \frac{1}{2} \ln 2\pi e\sigma^2
\] (1)

Note. For a n-dimensional Gaussian $X$ with mean zero and covariance $K$, $H(p) = \frac{1}{2} \ln (2\pi e)^m |K|$ where $|K|$ is the determinant of $K$. 
3.2 Joint and Conditional Entropy

**Def.** The entropy of a set $X_1, X_2, ..., X_n$ of random variables with density $p(x_1, x_2, ..., x_n)$ is defined as

$$H(p(x_1, x_2, ..., x_n)) = - \int p(x_1, x_2, ..., x_n) \ln p(x_1, x_2, ..., x_n) dx_1 \cdots dx_n.$$  

**Def.** If $X$ and $Y$ have a joint density $p(x, y)$, the conditional entropy $H(X|Y)$ is defined as

$$H(X|Y) = - \int p(x, y) \ln p(x|y) dx dy.$$
3.2 Joint and Conditional Entropy

**Def.** The entropy of a set $X_1, X_2, \ldots, X_n$ of random variables with density $p(x_1, x_2, \ldots, x_n)$ is defined as

$$H(p(x_1, x_2, \ldots, x_n)) = -\int p(x_1, x_2, \ldots, x_n) \ln p(x_1, x_2, \ldots, x_n) \, dx_1 \cdots dx_n.$$ 

**Def.** If $X$ and $Y$ have a joint density $p(x, y)$, the conditional entropy $H(X|Y)$ is defined as

$$H(X|Y) = -\int p(x, y) \ln p(x|y) \, dx \, dy.$$ 

**Q.** Why not $-\int p(x|y) \ln p(x|y) \, dx \, dy$?
3.2 Joint and Conditional Entropy

**Def.** The entropy of a set $X_1, X_2, ..., X_n$ of random variables with density $p(x_1, x_2, ..., x_n)$ is defined as

$$H(p(x_1, x_2, ..., x_n)) = - \int p(x_1, x_2, ..., x_n) \ln p(x_1, x_2, ..., x_n) dx_1 \cdots dx_n.$$

**Def.** If $X$ and $Y$ have a joint density $p(x, y)$, the conditional entropy $H(X|Y)$ is defined as

$$H(X|Y) = - \int p(x, y) \ln p(x|y) dx dy.$$

**Q.** Why not $- \int p(x|y) \ln p(x|y) dx dy$?

**Fact.** $H(X|Y) = H(X, Y) - H(Y)$
3.3 Relative Entropy and Mutual Information

**Def.** The relative entropy (or Kullback-Leibler distance) $D(p, q)$ between two densities $p$ and $q$ is defined by

$$D(p, q) = \int p \ln \frac{p}{q} dx$$

$D$ is a measure of the inefficiency of assuming that the distribution is $q$ when the true distribution is $p$.

**Def.** The mutual information $I(X, Y)$ between two random variables with joint density $p(x, y)$ is defined as

$$I(X, Y) = \int p(x, y) \ln \frac{p(x, y)}{p(x)p(y)} dx dy.$$ 

**Note.**

$I(X, Y) = D(p(x, y), p(x)p(y)) = H(X) + H(Y) - H(X, Y)$. 
3.3 Relative Entropy and Mutual Information

**Example.** Let \((X, Y)\) is a Gaussian with mean \((0, 0)\) and a covariance \(K = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\).

\[
H(X) = H(Y) = \frac{1}{2} \ln(2\pi e) \quad \text{and} \quad H(X, Y) = \frac{1}{2} \ln(2\pi e) 2(1 - \rho^2).
\]

Therefore
\[
I(X, Y) = H(X) + H(Y) - H(X, Y) = -\frac{1}{2} \ln(1 - \rho^2).
\]

If \(\rho = 0\), \(X\) and \(Y\) are independent and the mutual information is 0.

If \(\rho = \pm 1\), \(X\) and \(Y\) are perfectly correlated and the mutual information is infinite.

Note. \(X\) and \(Y\) are Gaussian and thus zero correlation implies independence.
3.3 Relative Entropy and Mutual Information

**Example.** Let \((X, Y)\) is a Gaussian with mean \((0, 0)\) and a covariance \(K = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\).

\[ H(X) = H(Y) = \frac{1}{2} \ln(2\pi e) \quad \text{and} \quad H(X, Y) = \frac{1}{2} \ln(2\pi e)^2 (1 - \rho^2). \]

Therefore \(I(X, Y) = H(X) + H(Y) - H(X, Y) = -\frac{1}{2} \ln(1 - \rho^2). \) If \(\rho = 0\), \(X\) and \(Y\) are independent and the mutual information is 0. If \(\rho = \pm 1\), \(X\) and \(Y\) are perfectly correlated and the mutual information is infinite.

**Note.** \(X\) and \(Y\) are Gaussian and thus zero correlation implies independence.
3.4 Properties of entropy, relative entropy, and mutual information

**Theorem.**

\[ D(p, q) \geq 0 \]

with equality iff \( p = q \) almost everywhere.

**Proof.**

\[
\begin{align*}
-D(p, q) &= \int p \ln \frac{q}{p} \, dx \\
&\leq \ln \int p \frac{q}{p} \, dx \quad \text{from Jensen’s inequality} \\
&= \ln \int g \\
&\leq \ln 1 = 0.
\end{align*}
\]

(2)

**Corollary.** \( I(X, Y) \geq 0 \) with equality iff \( X \) and \( Y \) are independent.
Corollary. \( H(X|Y) \leq H(X) \) with equality iff \( X \) and \( Y \) are independent.
Corollary. $H(X|Y) \leq H(X)$ with equality iff $X$ and $Y$ are independent. That is, collecting data decreases uncertainty (yay!).
3.4 Properties of entropy, relative entropy, and mutual information

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

\[ H(X_1, X_2, ..., X_n) = \sum H(X_i | X_1, X_2, ..., X_{i-1}). \]

**Proof.** Homework.

**Corollary.**

\[ H(X_1, X_2, ..., X_n) \leq \sum H(X_i) \]

**Hadamard’s inequality.** If \( X \) is a Gaussian distribution with mean 0 and a covariance \( K \), we have

\[ |K| \leq \prod_{i=1}^{n} K_{ii} \]

where \( |K| \) is the determinant of \( K \).
3.4 Properties of entropy, relative entropy, and mutual information

In Lecture 1, we have seen that the probability density maximizing entropy with a given mean and a variance is Gaussian. Now we show the following general result.

**Theorem.** Let the random vector $X \in \mathbb{R}^n$ have zero mean and covariance $K$. Then

$$H(X) \leq \frac{1}{2} \ln(2\pi e)^n |K|,$$

with equality iff $X$ is Gaussian is the covariance $K$ and mean zero. $|K|$ is the determinant of $K$.

**Proof.** Let $g(x)$ be any density satisfying $\int g(x)x_i x_j dx_i dx_j = K_{ij}$ for all $i, j$. Let $\phi_K$ be the density of the Gaussian $N(0, K)$. Then

$$0 \leq D(g, \phi_K) = \int g \ln(g/\phi_K) = -h(g) - \int g \ln \phi_K = -h(g) - \int \phi_K \ln \phi_K = -h(g) + h(\phi_K).$$ (3)
3.4 Properties of entropy, relative entropy, and mutual information

**Theorem.** (Estimation error) For any one-dimensional random variable $X$ and estimator $\hat{X}$,

$$E[(X - \hat{X})^2] \geq \frac{1}{2\pi e} e^{2H(X)},$$

with equality iff $X$ is Gaussian and $\hat{X}$ is the mean of $X$.

**Proof.** Let $\hat{X}$ be any estimator of $X$. Then

$$E[(X - \hat{X})^2] \geq \min_{\hat{X}} E[(X - \hat{X})^2]$$
$$= E[(X - E[X])^2]$$
$$= Var(X)$$
$$\geq \frac{1}{2\pi e} e^{2H(X)}. \quad (4)$$
Homework

- Draw $n$ values of the standard normal random variable, $X$.
- When $Y = X^2$, calculate $D(X, Y)$ using the sample. If you use a histogram in a sense, change the number of bins and check the change of the relative entropy.
- Compare the relative entropy with an analytic solution.