University of Stockholm
MT7037-VT19 Statistical Information Theory
Final Examination

Marina Herrera Sarrias

March 24, 2019

1 Exercise 1

(a) Let X and Y be discrete random variables. Show that H ((X+Y)|X) = H(Y|X). Argue that if X and Y are independent then H(X+Y) $\geq$ H(Y).

We wish to prove that the conditional entropy of Z ($Z = g(X,Y) = X + Y$) is the same as the conditional entropy of Y ($Y = Y'$ where $Y' = Z - X$) given X.

As Z is a function of (X,Y) and the probability of Z taking a value is linked to the probability of X having one. Both probabilities are the same as Z will happen if and only if X happens.

$$p(Z = z|X = x) = p(Y = z - x|X = x) = p(Y = y'|X = x)$$

We will now prove the equality:

$$\begin{aligned} H(Z|X) &= \sum p(x)H(Z|X = x) \\ &= \sum_{x\in\chi} p(X = x) \sum_{z\in\mathcal{Z}} p(Z = z|X = x) \log p(Z = z|X = x) \\ &= \sum_{x\in\chi} p(X = x) \sum_{y'\in\mathcal{Y}'} p(Y = y'|X = x) \log p(Y = y'|X = x) \\ &= \sum p(x)H(Y|X = x) \\ &= H(Y|X) \end{aligned} \tag{1}$$

Referring to their entropy we can say that:

$$H(X,Y) = H(X) + H(Y|X) \geq H(g(X,Y)) = H(Z)$$

and consequently, knowing that conditioning reduces entropy:

$$H(Z) \geq H(Z|X) = H(Y|X)$$

In the case that X and Y are independent discrete random variables. We can say that the addition of independent random variables adds uncertainty:

$$H(Z) \geq H(Z|X) = H(Y|X) = H(Y)$$

or similarly,

$$H(Z) \geq H(Z|Y) = H(X|Y) = H(X)$$

So,

$$\max\{H(X), H(Y)\} \leq H(Z)$$

■

(b) Show that among all N - valued random variables (i.e. with values k = 1,2,...) with expected value μ ,the geometric distributed random variable with expected value μ has the maximum value of Shannon entropy. Reminder: The probability function in a geometric distribution is $p(k) = p(1-p)^{k-1}, k = 1, 2, ...$

Assuming that

$$P(X = k) = p(1-p)^{k-1}$$

Where k and $X \in \mathbb{N}_+$. Knowing that:

$$\sum_{k=1}^{+\infty}(k-1)(1-p)^{k-1} = \frac{1-p}{(1-(1-p))^2}, \qquad \sum_{k=1}^{+\infty}(1-p)^{k-1} = \frac{1}{(1-(1-p))}$$

The entropy of the geometric distribution is derived as:

$$\begin{aligned}
H(X) &= -\sum_{k=1}^{+\infty} p(1-p)^{k-1} \log p(1-p)^{k-1} \\
&= -\sum_{k=1}^{+\infty} p(1-p)^{k-1} \log p - \sum_{k=1}^{+\infty}(k-1)p(1-p)^{k-1} \log(1-p) \\
&= -[\sum_{k=0}^{+\infty} p(1-p)^{k} \log p - \sum_{k=0}^{+\infty} kp(1-p)^{k} \log(1-p)] \\
&= -\frac{p}{1-(1-p)} \log p - \frac{(1-p)}{(1-(1-p))^2} \log p(1-p) \\
&= -\log p - \frac{(1-p)}{p} \log p(1-p)
\end{aligned}$$

We are now going to solve the maximization problem:

$$\arg\max_{p} \Big(-\sum_{k=1}^{+\infty} p_k \log p_k \Big)$$

Subject to the constraints:

$$\sum_{k=1}^{+\infty} p_k = 1$$

$$\sum_{k=1}^{+\infty} k p_k = \mu$$

We will use the Lagrange multipliers to obtain the general form of the p_k distribution.

$$L(p_k, \lambda_1, \lambda_2) = -\Big(\sum_{k=1}^{+\infty} p_k \log p_k \Big) + \lambda_1 \Big(\sum_{k=1}^{+\infty} p_k - 1 \Big) + \lambda_2 \Big(\sum_{k=0}^{+\infty} k p_k - \mu \Big)$$

$$\begin{aligned}
\frac{\partial L}{\partial \lambda_1} &= 0 = \sum_{k=1}^{+\infty} p_k - 1 \\
&= \exp^{(-1+\lambda_1)} \sum_{k=1}^{+\infty} \exp^{(\lambda_2)k} \\
&= \exp^{(-1+\lambda_1)} \Big(\frac{1}{1-\exp^{\lambda_2}} - 1 \Big) = 1 \\
&= \exp^{(-1+\lambda_1)} \Big(\frac{\exp^{\lambda_2}}{1-\exp^{\lambda_2}} \Big) = 1 \\
\frac{\partial L}{\partial \lambda_2} &= 0 = \sum_{k=0}^{+\infty} k p_k - \mu \\
&= \exp^{(-1+\lambda_1)} \sum_{k=1}^{+\infty} k \exp^{(\lambda_2)k} \\
&= \exp^{(-1+\lambda_1)} \frac{\exp^{\lambda_2}}{(1-\exp^{\lambda_2})^2} = \mu \\
\frac{\partial L}{\partial p_k} &= 0 = -\log p_k - 1 + \lambda_1 + \lambda_2 k
\end{aligned}$$

Where,

$$\begin{aligned}\mu &= \frac{1}{1-\exp^{\lambda_2}}\\ \lambda_2 &= \log\left(\frac{\mu-1}{\mu}\right)\\ \lambda_1 &= \log\left(\frac{1}{\mu-1}\right)+1\end{aligned}$$

Finally, the p_k distribution is derived as:

$$\begin{aligned}p_k &= \exp^{(-1+\lambda_1+\lambda_2 k)}\\ &= \exp^{(-1+\lambda_1+\lambda_2)}\exp^{(k-1)\lambda_2}\\ &= \frac{1}{\mu}\exp^{(k-1)\lambda_2}\\ &= \frac{1}{\mu}\exp^{(k-1)\log\left(\frac{\mu-1}{\mu}\right)}\\ &= \left(\frac{1}{\mu}\right)\left(\frac{\mu-1}{\mu}\right)^{k-1}\end{aligned} \tag{2}$$

We also know that the mean of the geometric distribution is:

$$\begin{aligned}\mu = E[X] &= \sum_{k=1}^{+\infty} k(1-p)^{k-1}p\\ &= p\sum_{k=1}^{+\infty} k(1-p)^{k-1}\\ &= p\frac{1}{(1-(1-p))^2}\\ &= \frac{1}{p}\end{aligned} \tag{3}$$

If we now plug the results obtained in eq.(3) into eq.(2) we can verify that p_k is the geometric distribution i.e, the distribution that maximizes the entropy.

We could also prove it using the Gibb's inequality eq.(5), which is just the difference of the Kullback–Leibler divergence eq.(4). Knowing that the random variable X distributed as $p_k \neq 0$ and for the random variable Y

distributed as q_k, where $k \in \mathbb{Z}^+$

$$\begin{aligned} \mathbb{D}_{KL}(Y|X) &= \sum_{k=1}^{+\infty} q_k \log \frac{q_k}{p_k} \geq 0 \\ &= \sum_{k=1}^{+\infty} q_k \log q_k - \sum_{k=1}^{+\infty} q_k \log p_k \end{aligned} \tag{4}$$

$$-\sum_{k=1}^{+\infty} q_k \log q_k \leq -\sum_{k=1}^{+\infty} q_k \log p_k \tag{5}$$

and knowing that:

$$H(Y) = -\sum_{k=1}^{+\infty} q_k \log q_k$$

$$\begin{aligned} -\sum_{k=1}^{+\infty} q_k \log p_k &= -\sum_{k=1}^{+\infty} q_k(\log p + (1-k)\log(1-p)) \\ &= -\log p \sum_{k=1}^{+\infty} q_k - \sum_{k=1}^{+\infty} q_k(1-k)\log(1-p)) \\ &= -\log p - \sum_{k=1}^{+\infty} q_k(1-k)\log(1-p) \\ &= -\log p - \log(1-p)\sum_{k=1}^{+\infty} q_k(1-k) \\ &= -\log p - \log(1-p)\sum_{k=1}^{+\infty} q_k + \log(1-p)\sum_{k=1}^{+\infty} kq_k \\ &= -\log p - \log(1-p) + \log(1-p)\sum_{k=1}^{+\infty} kq_k \\ &= -\log p - \log(1-p) + \log(1-p)\mathbb{E}[Y] \\ &= -\log p + \log(1-p)(\mathbb{E}[Y] - 1) \\ &= -\log p - \frac{(1-p)}{p}\log p(1-p) = \mathbb{H}(X) \end{aligned}$$

$$\mathbb{H}(Y) \leq \mathbb{H}(X)$$

$\blacksquare$

(c) Show that
(i) H(X_1,X_2,X_3) $\leq \frac{1}{2}$ [H(X_1,X_2) +H(X_2,X_3) +H(X_1,X_3)]

(ii)H(X_1,X_2,X_3) $\geq \frac{1}{2}$ [H(X_1,X_2 |X_2)+H(X_2,X_3|X_1)+H(X_1,X_3|X_2)]

Using the chain rule of entropy we can prove that a collection of random variables are the sum of the conditional entropies, such that:

$$H(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, ..., X_1)$$

We also know that as conditioning reduces entropy:

$$H(X_n) \geq H(X_n|X_1) \geq H(X_n|X_{n-1}, X_1)$$

(i)

The left hand side of the inequality is:

$$\begin{aligned} H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3|X_1) \\ &= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) \end{aligned} \tag{6}$$

While , the right hand side is:

$$\begin{aligned} H(X_1, X_2) + H(X_2, X_3) + H(X_1, X_3) &= H(X_1) + H(X_2|X_1) + H(X_2) \\ &= +H(X_3|X_2) + H(X_1) + H(X_3|X_1) \\ &= 2H(X_1) + H(X_2|X_1) + H(X_2) \\ &= +H(X_3|X_2) + H(X_3|X_1) \end{aligned} \tag{7}$$

Now, referring to the results obtained in eq.(7) we could also say that:

$$\begin{aligned} &\frac{1}{2}[H(X_2) + H(X_2|X_1) + \\ &H(X_3|X_2) + H(X_3|X_1)] \geq +\frac{1}{2}[H(X_2|X_1) + H(X_2|X_1) + \\ &\qquad H(X_3|X_2, X_1) + H(X_3|X_2, X_1)] \\ &\qquad \geq H(X_2|X_1) + H(X_3|X_2, X_1) \end{aligned} \tag{8}$$

and, because of the above, we can finally prove that:

$$H(X_1, X_2, X_3) \leq H(X1) + \frac{1}{2}[H(X_2) + H(X_2|X_1) + H(X_3|X_2) + H(X_3|X_1)]$$

$\blacksquare$

(ii)

The left hand side of the inequality is the same as in eq.(6) while the right hand side is as follows:

$$\begin{aligned}
&= \frac{1}{2}[H(X_1, X_2|X_3) + H(X_2, X_3|X_1) + H(X_1, X_3|X_2)] \\
&= \frac{1}{2}[H(X_1|X_3) + H(X_2|X_3, X_1) + H(X_2|X_1) \\
&+ H(X_3|X_2, X_1) + H(X_1|X_2) + H(X_3|X_2, X_1)] \qquad (9) \\
&= H(X_3|X_2, X_1) + \frac{1}{2}[H(X_1|X_3) + H(X_2|X_3, X_1) \\
&+ H(X_2|X_1) + H(X_1|X_2)]
\end{aligned}$$

Following the same intuition as in eq.(8) we can say about the results obtained in eq.(9) that:

$$\begin{aligned}
&\frac{1}{2}[H(X_1|X_3) + H(X_2|X_3, X_1) \\
&\quad +H(X_2|X_1) + H(X_1|X_2)] \leq \frac{1}{2}[H(X_2|X_1) + H(X_2|X_1) + H(X_1) + H(X_1)] \\
&\quad \leq H(X_1) + H(X_2|X_1)
\end{aligned}$$

We can now verify that:

$$\begin{aligned}
H(X_1, X_2, X_3) \geq H(X_3|X_2, X_1) + \frac{1}{2}[H(X_1|X_3) + H(X_2|X_3, X_1) \\
+ H(X_2|X_1) + H(X_1|X_2)]
\end{aligned}$$

$\blacksquare$

(d) Suppose that (X, Y, Z) are jointly normal distributed and that $X \to Y \to Z \to$ forms a Markov chain. Let X and Y have the correlation coefficient $\rho 1$ and let Y and Z have the correlation coefficient $\rho 2$. Find mutual information $I(X; Z)$.

The mutual information $I(X; Z)$ is the reduction in the uncertainty of X due to the knowledge of Z and it is represented as:

$$\begin{aligned} I(X;Z) &= H(X) + H(Z) - H(X,Z) \\ &= H(X) + H(X|Z) \end{aligned} \tag{10}$$

Furthermore the entropy of a normal distribution is derived as:

$$h(X) = -\int_{-\infty}^{\infty} \phi(X) \log \phi(X) dx$$

Where the density function $\phi(X)$ is:

$$\phi(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{11}$$

and then, plugging eq.(11) in eq.(10) we obtain:

$$\begin{aligned} h(X) &= -\int_{-\infty}^{\infty} \phi(X) \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}\right) dx \\ &= -\int_{-\infty}^{\infty} \phi(X) \log \frac{1}{\sqrt{2\pi\sigma^2}} dx + \log(e) \int_{-\infty}^{\infty} \phi(X) (\frac{-(x-\mu)^2}{2\sigma^2}) dx \\ &= -\int_{-\infty}^{\infty} \phi(X) \log \frac{1}{\sqrt{2\pi\sigma^2}} dx - \log(e) \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x-\mu)^2 \phi(X) dx \\ &= \frac{1}{2} \log\bigl(2\pi\sigma^2\bigr) - \log e \frac{1}{2} \\ &= \frac{1}{2} [\log 2\pi\sigma^2 - \log e] \\ &= -\frac{1}{2} \log[2\pi e \sigma^2] \end{aligned}$$

Since (X, Y, Z) are jointly normal distributed X and Z will also be. Their covariance matrix $\Sigma_{x,z}$ is defined as:

$$\Sigma_{x,z} = \begin{bmatrix} {\sigma_x}^2 & \sigma_x \sigma_z \rho_{xz} \\ \sigma_x \sigma_z \rho_{xz} & {\sigma_z}^2 \end{bmatrix}$$

Now, going back to the mutual information $I(X;Z)$ defined on eq.(10) we obtain:

$$\begin{aligned} I(X;Z) &= \frac{1}{2}\log\bigl(2\pi e {\sigma_x}^2\bigr) + \frac{1}{2}\log\bigl(2\pi e {\sigma_z}^2\bigr) - \frac{1}{2}\log\bigl(2\pi e^2 |\Sigma_{x,z}|\bigr) \\ &= \frac{1}{2}\log\bigl(2\pi e {\sigma_x}^2\bigr) + \frac{1}{2}\log\bigl(2\pi e {\sigma_z}^2\bigr) - \log(2\pi e) - \frac{1}{2}|\Sigma_{x,z}| \end{aligned}$$

therefore;

$$I(X;Z) = -\frac{1}{2}|\Sigma_{x,z}|$$

Where,

$$\begin{aligned} |\Sigma_{x,z}| &= {\sigma_x}^2{\sigma_z}^2 - {\sigma_x}^2{\sigma_z}^2{\rho_{xz}}^2 \\ &= {\sigma_x}^2{\sigma_z}^2(1-{\rho_{xz}}^2) \end{aligned}$$

If we now assume that $(X,Y,Z) \sim \mathcal{N}(0,1)$:

$$I(X;Z) = -\frac{1}{2}\log\bigl(1-{\rho_{xz}}^2\bigr) \tag{12}$$

The Markov chain $X \rightarrow Y \rightarrow Z$ implies that the conditional distribution of Z depends only on Y and is conditionally independent of X. The joint probability of the random variables (X,Y,Z) is:

$$p(x,y,z) = p(x)p(y|x)p(z|y)$$

and it implies conditional independence as:

$$\begin{aligned} p(x,z|y) &= \frac{p(x,y,z)}{p(y)} \\ &= \frac{p(x)p(x|y)p(z|y)}{p(y)} \\ &= \frac{p(x,y)p(z|y)}{p(y)} \\ &= \frac{p(x|y)p(y)p(z|y)}{p(y)} \\ &= p(x|y)p(z|y) \end{aligned} \tag{13}$$

Then the correlation coefficient ρ_{xz} is obtained as:

$$\begin{aligned}
\rho_{xz} &= \frac{E[XZ] - E[X]E[Z]}{\sigma_x \sigma_z} \\
&= \frac{E[XZ]}{\sigma_x \sigma_z} \\
&= \frac{E[E[XZ|Y]]}{\sigma_x \sigma_z} \\
&= \frac{E[E[X|Y]E[Z|Y]]}{\sigma_x \sigma_z} \\
&= \frac{E[\frac{E[Y]-E[X]cov(X,Y)}{\sigma_y} + \frac{cov(X,Y)}{\sigma_y}Y][\frac{E[Y]-E[Z]cov(Z,Y)}{\sigma_y} + \frac{cov(Z,Y)}{\sigma_y}Y]}{\sigma_x \sigma_z} \\
&= \frac{E[\frac{cov(X,Y)}{\sigma_y}Y][\frac{cov(Z,Y)}{\sigma_y}Y]}{\sigma_x \sigma_z} \\
&= \frac{E[\sigma_x \rho_{xy} Y][\sigma_z \rho_{zy} Y]}{\sigma_x \sigma_z} \\
&= \rho_{xy}\rho_{zy}E[Y^2] \\
&= \rho_{xy}\rho_{zy}
\end{aligned}$$

Plugging the results into eq.(12), we can now conclude that:

$$\begin{aligned}
I(X;Z) &= -\frac{1}{2}\log\bigl(1 - {\rho_{xy}}^2 {\rho_{zy}}^2\bigr) \\
&= -\frac{1}{2}\log\bigl(1 - {\rho_1}^2 {\rho_2}^2\bigr)
\end{aligned}$$

$\blacksquare$

2 Exercise 2

(a) Define the transfer entropy, starting from Schreiber's definition in Phys. Rev. Lett., 85, 461, 2000, in terms of mutual information and Shannon entropy.

The random time series processes $\mathbb{X}_t^{(k)}$ and $\mathbb{Y}_t^{(l)}$ are a join of sequential processes that evolve over time t, with a history length (lags/time steps) of k for the target variable and j for the source variable, in which most of the cases $l = 1$ or $l = k$.

$$\mathbb{X}_t^{(k)} = (X_t, X_{t-1}, ..., X_{t-k+1})$$
$$\mathbb{Y}_t^{(l)} = (Y_t, Y_{t-1}, ..., Y_{t-l+1})$$

The transfer entropy $T^{(k,l)}_{\mathbb{Y}\to\mathbb{X}}(t)$ is defined as the difference between the entropy of the state of X_t conditioned on its own past $\mathbb{X}^k_{t-1}$ and its entropy conditioned in addition to the past of $\mathbb{Y}^l_{t-1}$. It therefore measures how much information the source variable $\mathbb{Y}^l_{t-1}$ provides about the state transitions in the target variable $\mathbb{X}^k_{t-1}$. It measures the deviation in bits from the Markov property of conditional independence.

$$T^{(k,l)}_{\mathbb{Y}\to\mathbb{X}}(t) = H(X_t|\mathbb{X}^k_{t-1}) - H(X_t|\mathbb{X}^k_{t-1}, \mathbb{Y}^l_{t-1})$$

It can also be formulated as the conditional mutual information: $I(X_t, \mathbb{Y}_{t-1}{}^{(l)}|\mathbb{X}_{t-1}{}^{(k)})$ which is the reduction of uncertainty by the process $\mathbb{Y}_{t-1}{}^{(l)}$ in the state X_t conditioned to its own past. i.e, how much information the source variable $\mathbb{X}_{t-1}{}^{(k)}$ adds in predicting the next state of the target variable. The current state of X_t will depend probabilistically on the state of Y_{t-1} at the previous time step $t-1$.

(b) Why is mutual information not a good measure for an information transfer?

Mutual information neither contains dynamical nor directional information. Directional in the sense that for: $I(X_t : Y_t)$ the exchange between the two systems is a symmetric measure of statically shared information. i.e, we do not know if there is a direction. We will only know that if the mutual information is greater than zero the two variables are not independent. While in a directional system there is only information transfer in one direction

$Y \to X$ as the information transfer in $X \to Y$ will be zero.

Dynamical, in the sense that mutual information is a property of static probability distributions and a dynamic system is described as the evolution of a set of state variables, i.e a system that evolves over time. Conditioning on the past can make a measure to be directional, and dynamic, which is the case of the transfer entropy. The Cellular Automata is a good example of a directional and dynamic system.

(c) Does Schreiber's definition of transfer entropy coincide with the definition of mutual information?

Yes, it coincides with the definition of conditional mutual information, it can be proved as following:

$$T^{(k,l)}_{\mathbb{Y}\to\mathbb{X}}(t) = \sum_{X_t, {\mathbb{X}_{t-1}}^{(k)}, {\mathbb{Y}_{t-1}}^{(l)}} p(x_t, {\mathbb{x}_{t-1}}^{(k)}, {\mathbb{y}_{t-1}}^{(l)}) \log \frac{p(x_t|{\mathbb{x}_{t-1}}^{(k)}, {\mathbb{y}_{t-1}}^{(l)})}{p(x_t|{\mathbb{x}_{t-1}}^{(k)})} \tag{14}$$

Where:

$$\begin{aligned} p(x_t|{\mathbb{x}_{t-1}}^{(k)}, {\mathbb{y}_{t-1}}^{(l)}) &= \frac{p(x_t, {\mathbb{x}_{t-1}}^{(k)}, {\mathbb{y}_{t-1}}^{(l)})}{p({\mathbb{x}_{t-1}}^{(k)}, {\mathbb{y}_{t-1}}^{(l)})} \\ &= \frac{p(x_t, {\mathbb{y}_{t-1}}^{(l)}|{\mathbb{x}_{t-1}}^{(k)})p({\mathbb{x}_{t-1}}^{(k)})}{p({\mathbb{x}_{t-1}}^{(k)}, {\mathbb{y}_{t-1}}^{(l)})} \\ &= \frac{p(x_t, {\mathbb{y}_{t-1}}^{(l)}|{\mathbb{x}_{t-1}}^{(k)})}{\frac{p({\mathbb{x}_{t-1}}^{(k)}, {\mathbb{y}_{t-1}}^{(l)})}{p({\mathbb{x}_{t-1}}^{(k)})}} \\ &= \frac{p(x_t, {\mathbb{y}_{t-1}}^{(l)}|{\mathbb{x}_{t-1}}^{(k)})}{p({\mathbb{y}_{t-1}}^{(l)}|{\mathbb{x}_{t-1}}^{(k)})} \end{aligned} \tag{15}$$

Plugging the results of eq.(15) into eq.(14) we obtain:

$$T^{(k,l)}_{\mathbb{Y}\to\mathbb{X}}(t) = \sum_{X_t, {\mathbb{X}_{t-1}}^{(k)}, {\mathbb{Y}_{t-1}}^{(l)}} p(x_t, {\mathbb{x}_{t-1}}^{(k)}, {\mathbb{y}_{t-1}}^{(l)}) \log \frac{p(x_t, {\mathbb{y}_{t-1}}^{(l)}|{\mathbb{x}_{t-1}}^{(k)})}{p(x_t|{\mathbb{x}_{t-1}}^{(k)})p({\mathbb{y}_{t-1}}^{(l)}|{\mathbb{x}_{t-1}}^{(k)})}$$

And knowing that the mutual information $I(X_t, {\mathbb{Y}_{t-1}}^{(L)}|{\mathbb{X}_{t-1}}^{(k)})$ is:

$$I(X_t, {\mathbb{Y}_{t-1}}^{(L)}|{\mathbb{X}_{t-1}}^{(k)}) = \sum_{X_t, {\mathbb{X}_{t-1}}^{(k)}, {\mathbb{Y}_{t-1}}^{(l)}} p(X_t, {\mathbb{X}_{t-1}}^{(k)}, {\mathbb{Y}_{t-1}}^{(l)}) \log \frac{p(x_t, {\mathbb{y}_{t-1}}^{(l)}|{\mathbb{x}_{t-1}}^{(k)})}{p(x_t|{\mathbb{x}_{t-1}}^{(k)})p({\mathbb{y}_{t-1}}^{(l)}|{\mathbb{x}_{t-1}}^{(k)})}$$

We can now verify that:

$$T^{(k,l)}_{\mathbb{Y}\to\mathbb{X}}(t) = I(X_t, {\mathbb{Y}_{t-1}}^{(l)}|{\mathbb{X}_{t-1}}^{(k)})$$

■

(d) What is the meaning of local transfer entropy?

The local transfer entropy describes the information contained in the source variable y_{t-1} about the next state x_t at a time t and that is not contained in $x^{(k)}_{t-1}$, as this is a local measure it is defined at each time t. The average over all the points will be the transfer entropy of the system.

The local transfer entropy gives information about the dynamics (transition probabilities) of information transfer in time. Insights that the transfer entropy can not access, as it is an average. It can also be defined as a local conditional mutual information: $i({\mathbb{y}_{t-1}}^{(l)}; x_t|{\mathbb{x}_{t-1}}^{(k)})$.

(e) Is the Granger causality concept the same as transfer entropy? Explain possible similarities and differences.

According to the paper by Barnett et al., the concepts od Granger causality and transfer entropy are entirely equivalent for Gaussian processes, up to certain conditions and under a parametric formulation. It is important to stress in the fact that they will only measure the same thing for Gaussian processes.

Both concepts numerical equivalence will depend on the method used to estimate the transfer entropy in the sample. i.e, if conditional entropies are estimated directly from sample probability distributions, results will vary,therefore the Granger causality can be considered after the introduction of constraints in the data, i.e Granger parametric formulation.

Some of this equivalence formulations are: Granger causality and transfer entropy for discrete variables with jointly multivariate Gaussian distribution are equivalent up to a factor of two,this as the Granger transfer entropy general formulation only deals with univariate variables, so this is just the result of the formulation extension by adding an extra conditioning factor. On the other hand, for discrete random variables with jointly multivariate Weinman exponential distribution, the Granger causality and transfer entropy will be the same up to a factor of 1. And for discrete random variables with jointly multivariate log-normal distribution, the Granger causality and transfer entropy will also be the same, up to a factor of $\frac{1}{2}$

The possible advantages that either measure could have are strongly related with the data and its underlying stochastic generative process. For instance, For highly non-linear data where there is too much variability, this can lead to invalidate the residuals-regressors un-correlation assumption (eq.(5) in the paper), same for a given case in which the regressors used for modeling do not have sufficient explanatory power, all of these could lead to compromise the statistical inference. So, indifferently that our data satisfies, the conditions for formulation a linear model this does not mean that the results will be significant, and that the parsimonious principal will hold, and therefore there will be occasions in which it will be better a non parametric method (TE).

(f) To complete this part you need to read the paper by Lizier and Prokopenko in Eur. Phys. J. B 73, 605-615, 2010.
i) Does information transfer have the same meaning as information flow in the opinion of the authors of the paper? Explain possible similarities and differences.
ii) What is the difference between interventional and standard conditional probabilities?
iii) Explain the difference between local transfer entropy and local information flow.
iv)What are the advantages of the definition of local information flow according to the authors of the paper.

(*i*)

The thesis of the paper is that both concepts, are different, and should be considered separately.

The information flow is principally related to identifying up to what extent the source variable can influence on predicting the next state of the target variable, while the predictive information transfer refers to how the state of the source variable helps on predicting the next state of the target variable. i.e, in which amount the source variable contributes to the prediction.

Both measures are also done differently, information transfer is quantified by transfer entropy, which only takes into account local values (observed conditional correlations), while information flow is quantified by introducing certain intervention or deviation, so we can actually learn in which extent the source variable influences the state of the target variable.

The argument for defining local information transfer and local information flow is the same however the meaning in local information flow is different. Also, unlike information transfer, information flow is not an average of local values.

(ii)

Following the paper's definition, considering two variables s and a an interventional conditional probability $p(a|\tilde{s})$ considers the distribution of a resulting from imposing the value of $\tilde{s}$, while its correlation alters $p(a|s)$ generally by $p(a)$, meaning that imposing a value on s, has no effect. I will illustrate this with a simple example. Lets say we have a disease s ("a weird disease"), which is correlated with a ("shortness of breath"), in the sense that s can cause a. If we ask about the probability of suffering from shortness of breath given that we intervened having this weird disease. The intervention wont have any effect on the relevant causal link, so it will be the same as the conditional probability:

$$p(a|\tilde{s}) = P(a|s)$$

Now, we also know that, s and a are solely caused by g ("certain bacterial infection"). If the system of s is intervened to set the value of this variable, so the question will be now: what is the probability of suffering from shortness of breath if I eliminate having the weird disease using a magic potion? such that:

$$P(a|\tilde{s}) = P(a)$$

This will be the same as the prior probability of suffering from shortness of breath. This, as having the weird disease has been disconnected from suffering from shortness of breath, thus it does not provide any information about it.

(iii)

The local information flow is defined in a similar manner as the local transfer entropy. However the local information flow has a different meaning. Let's say that given the observation $(y, x, \boldsymbol{s})$, we study the local cause effect of y on x given the imposition $\tilde{s}$ of the same observation. The local information flow will finally be the average over the product of interventional conditional probabilities $p(s)p(y|\tilde{\boldsymbol{s}})p(x|\tilde{y}, \tilde{\boldsymbol{s}})$.

As there is a possibility that not all of the tuples $\{y, x, s\}$ will actually be observed, we still need to take into account all the unobserved tuples as they still have a valuable weight. Therefore we should not rely only in the observed values, we should take into account all possible combinations so we can understand the causation in our system.

In summary, the local information flow does not make reference to a certain observation given a moment in time t, if not, it is related to how the general configuration of $(y, x, \boldsymbol{s})$ is attributed to the given observation at the given time step.

(iv)

Although both measures are complementary, the main advantage could rely on the fact that the measure for transfer entropy does not detect all causal effects that information flow does. so, in order to measure information transfer, the transfer entropy should only be applied to causal information sources, for a given target variable. Otherwise it will just be a measure of correlations.

3 Exercise 3

To understand better the concepts of multi-information and its decomposition, this exercise requires you to work out some details in the paper by Schneidman et al., Phys. Rev. Lett., 91, 238701, 2003 discussed in the class.

(b) Can you derive $Eq.(8)$ of the paper and correct the typos? Hint: Use the Venn diagram. This exercise shows that 1) although the area-information correspondence of the Venn diagram does not hold in the 3-variable case, the Venn diagram is still useful when deriving relations between different information quantities; 2) mistakes exist in published papers, so always be critical in reading them.

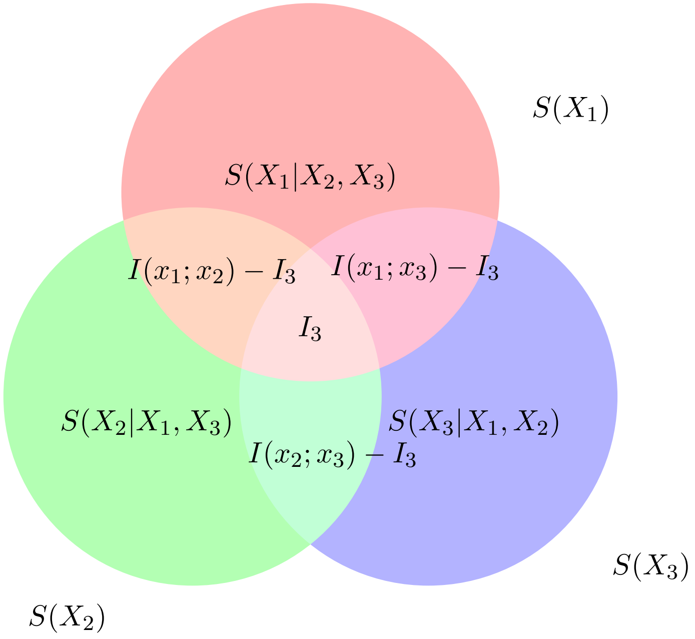

Figure 1: Venn diagram of triplet information

From the Venn diagram in Figure(1) we can derive the value of I_3 as follow:

$$\begin{aligned}
I_3 &= S(X_1,X_2,X_3) - S(X_1|X_2,X_3) - S(X_2|X_1,X_3) - S(X_3|X_2,X_1) \\
&- [I(X_1;X_3) - I_3] - [I(X_2;X_3) - I_3] - [I(X_1;X_2) - I_3] \\
&= S(X_1,X_2,X_3) - S(X_1|X_2,X_3) - S(X_2|X_1,X_3) - S(X_3|X_2,X_1) \\
&- I(X_1,X_2|X_3) - I(X_1,X_3|X_2) - I(X_2,X_3|X_1) \\
&= S(X_1,X_2,X_3) - [S(X_1,X_2,X_3) - S(X_2,X_3)] - [S(X_1,X_2,X_3) - S(X_1,X_3)] \\
&- [S(X_1,X_2,X_3) - S(X_2,X_1)] - I(X_1,X_2|X_3) - I(X_1,X_3|X_2) - I(X_2,X_3|X_1) \\
&= S(X_1,X_2,X_3) - 3S(X_1,X_2,X_3) + S(X_2,X_3) + S(X_1,X_3) + S(X_2,X_1) \\
&+ 3S(X_1,X_2,X_3) - 2S(X_2,X_3) - 2S(X_2,X_1) - 2S(X_1,X_3) + S(x_1) + S(x_2) + S(x_3) \\
&= S(X_1,X_2,X_3) - S(X_2,X_3) - S(X_1,X_3) - S(X_1,X_2) + S(X_1) + S(X_2) + S(X_3) \\
&= \sum_i S(X_i) - \sum_{i<j} S(X_i,X_j) + S(X_1,X_2,X_3)
\end{aligned} \tag{16}$$

The results obtained in eq.(16) agree with the one derived in the first line of $eq(8)$ in the paper. I will now proceed to prove it referring to the mutual information where:

$$\sum_{i<j} I(X_i,X_j) - I(X_1,X_2,X_3) = [I(X_1,X_3) + I(X_2,X_3)] - I(\{X_1,X_2\};X_3)$$

We also know that:

$$\begin{aligned}
I(\{X_1,X_2\};X_3) &= \sum_{X_1,X_2,X_3} p(X_1,X_2,X_3) \log \frac{p(X_1,X_2,X_3)}{p(X_1,X_2)p(X_3)} \\
&= H(X_3) - H(X_3|X_1,X_2) \\
&= H(X_3) + H(X_1,X_2) - H(X_1,X_2,X_3)
\end{aligned}$$

$$\begin{aligned}
\sum_{i<j} I(X_i,X_j) - I(X_1,X_2,X_3) &= [I(X_1,X_3) + I(X_2,X_3)] - [H(X_3) + H(X_1,X_2) - H(X_1,X_2,X_3)] \\
&= [H(X_1) + H(X_3) - H(X_1,X_3)] + [H(X_2) + H(X_3) \\
&- H(X_2,X_3)] - H(X_3) - H(X_1,X_2) + H(X_1,X_2,X_3) \\
&= [H(X_1) + H(X_2) + H(X_3) - H(X_1,X_3) - H(X_2,X_3) \\
&- H(X_1,X_2) + H(X_1,X_2,X_3) \\
&= \sum_i S(X_i) - \sum_{i<j} S(X_i,X_j) + S(X_1,X_2,X_3)
\end{aligned}$$

We can now verify that:

$$\sum_i S(X_i) - \sum_{i<j} S(X_i, X_j) + S(X_1, X_2, X_3) = \sum_{i<j} I(X_i, X_j) - I(X_1, X_2, X_3)$$

And point out that there is an error of signs in the paper, where the right formulation is : $\sum_{i<j} I(X_i, X_j) - I(X_1, X_2, X_3)$ instead of $I(X_1, X_2, X_3) - \sum_{i<j} I(X_i, X_j)$.

On the other hand, representing three terms entropies using the $Venn$ diagram can be misleading. Mostly as entropies can be confused with probabilties, and by the misconception that all regions correspond to positive quantities, which could be not case. (it is for a two entropy representation) ■

(c) Show that the connected information of order k in $Eq.$(6) of the paper can be written as a relative entropy. When does it equal to zero?

The connected information of order k (subset of k elements of N variables) is the information gain when we go from knowing only the marginals of order $k-1$ to knowing also the marginals of order k. It is defined as:

$$I_C^{(k)}(\{x_i\}) = S[\tilde{P}^{(k-1)}(\{x_i\})] - S[\tilde{P}^{(k)}(\{x_i\})]$$

Considering N variables $\{x_i\}$, where $i = 1, 2, 3$

$$\begin{aligned}
I_C^{(2)}(\{x_i\}) &= S[\tilde{P}^{(1)}(\{x_i\})] - S[\tilde{P}^{(2)}(\{x_i\})] \\
&= S[\tilde{p}^{(1)}(x_1, x_2, x_3)] - S[\tilde{p}^{(2)}(x_1, x_2, x_3)] \\
&= S[p(x_1)p(x_2)p(x_3)] - S[p(x_1, x_2, x_3)] \\
&= \sum p(x_1, x_2, x_3) \log p(x_1, x_2, x_3) - \sum p(x_1, x_2, x_3) \log p(x_1)p(x_2)p(x_3) \\
&= \sum p(x_1, x_2, x_3) \frac{\log p(x_1, x_2, x_3)}{p(x_1)p(x_2)p(x_3)} \\
&= D_{KL}(p(x_1, x_2, x_3)||(p(x_1)p(x_2)p(x_3))) \\
&= D_{KL}\tilde{P}^{(2)}(\{x_i\})||\tilde{P}^{(1)}(\{x_i\})) \\
&= D_{KL}(\tilde{P}^{(k)}(\{x_i\})||\tilde{P}^{(k-1)}(\{x_i\})) \geq 0
\end{aligned}$$

If the result of the Kullback-Leiber divergence is zero, this means that both distributions are the same, $\tilde{P}^{(k)}(\{x_i\}) = \tilde{P}^{(k-1)}(\{x_i\})$. i.e, there is independence among the variables in the model. ■

4 Exercise 4

(a) Plot the data points on the x-y plane to see how they look like and prepare the element-to-element distance matrix $d(\vec{x}_i, \vec{x}_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ $(i, j = 1, 2, ..., 300)$ that will be the input for the clustering algorithm. Also assume each data point carries the same weight, i.e., $p(\vec{x}_i) = 1/300$ for all i.

As it is already known, the data set is composed by 300 data points from which the first 100 have a different normal distribution than the remaining 200.

After plotting the data points without making any difference, if we had no previous information of the data points distribution, at first glance in Figure(2) we can say that it is clear that there are two clusters.

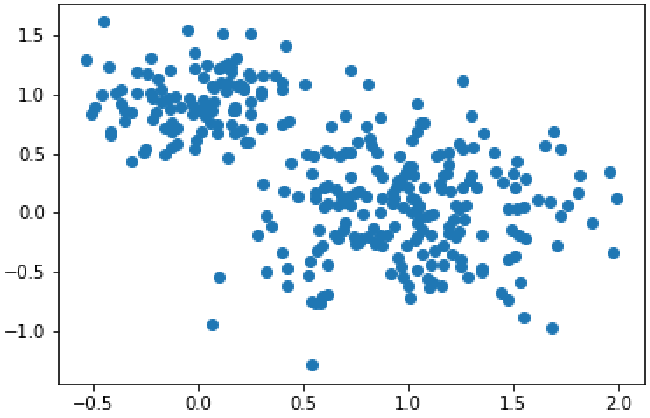

Figure 2: Data set Scatter plot

After dividing the data set by distribution we can now verify from Figure(3) that there are indeed two clusters.

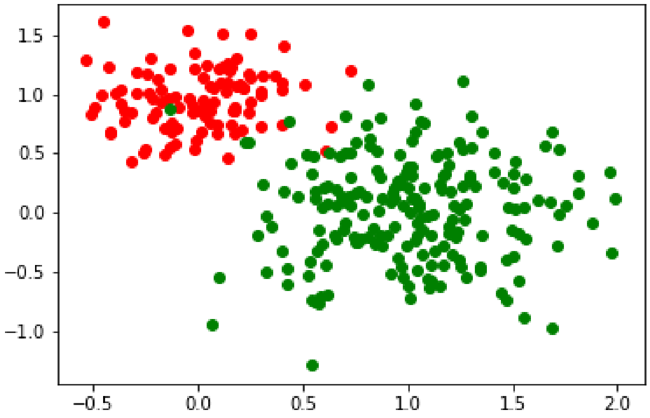

Figure 3: Data set Scatter plot by distribution

Unfortunately, in real applications is not always the case in which we will be aware of our data distribution. So we will work based on the fact that we do not know how our data points are distributed.

(b) Write a code (no restriction on the program language) to implement the Blahut-Arimoto algorithm discussed in the class to evaluate the clustering membership probability,, with fixed number of clusters, and compression distortion trade-off parameter. Your code should implement a multiple run each starting with random initial conditions. Note: Your source code should include clear comments/documentations to describe what is evaluating. We may later randomly ask a few students, especially those without clear documentations, to demonstrate how their code works.

I have created a small program for running the Blahut-Arimoto algorithm. A pdf with the code is enclosed. I have also included an html file with the jupyter notebook I used for testing my code.

(c), (d) Run your program to construct the information curves for $N_c = 2$, 3 and 4. Hint: choose different values of β in between 1 to 50.

After running the program and constructing the information curves Figure(4) and Figure(5) show some results.

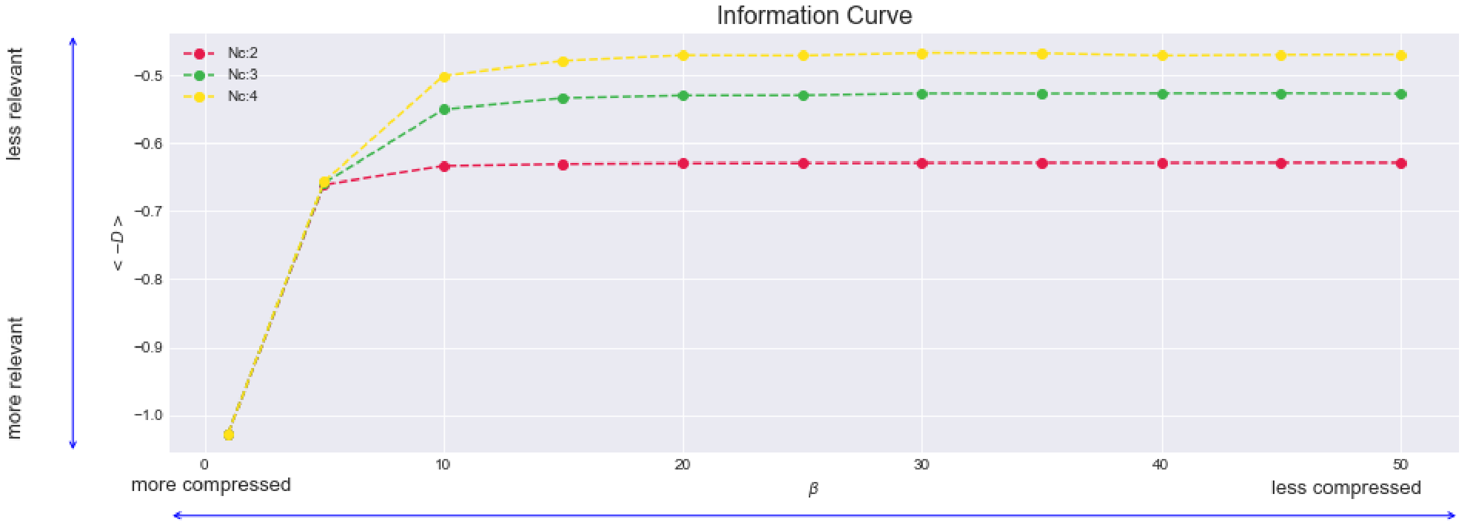

Figure 4: Information curve for β values from 0 to 50

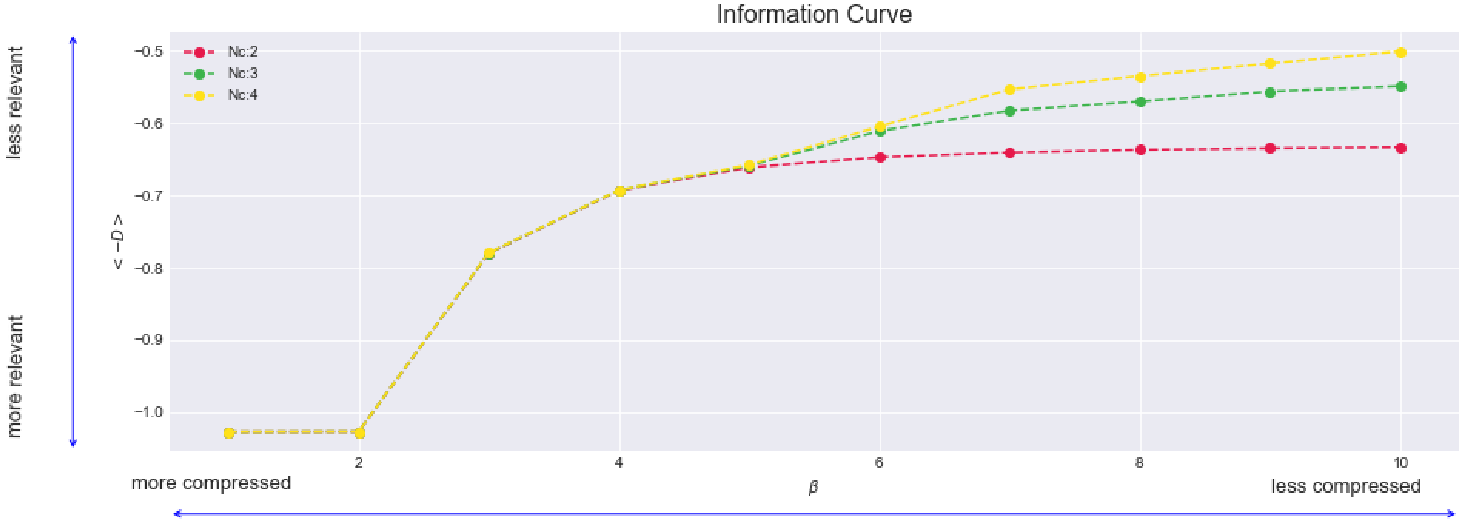

Figure 5: Information curve for β values from 1 to 10

Form Figure(4) and Figure(5) what we are trying to achieve is to find the model that compresses the information the most, while decreasing the distortion, i.e keeping relevant our model, and preserving as much information as possible by reducing the distortion.
On the x-axis we can find β, which represents the Lagrange multiplier attached to the constrained meaningful information. By changing the value of the parameter β we can actually explore the trade-off between compression and the preserved meaningful information.

On the y-axis we find the expected distortion rate, By decreasing the distortion we increment the relevance. The information curve will increase in both directions, and we will like to select the model in the predictive point.

In Figure(5) we can observe that for the given values of β from zero until five there is a fixed value of distortion rate for all clusters, which in this case we will always prefer the model with smallest number of clusters. i.e, $N_c = 2$.

It can also be noticed in Figure(4) that up to certain point ($\beta = 5$) each model takes a different direction, and as the number of clusters increases, the compression rate decreases and the distortion rate increases.
as $\beta \to \infty$ the distortion rate remains constant for all values of β. (starting from $\beta = 20$)

We can now conclude that the best model is given using 2 clusters.