ECE 587 / STA 563: Lecture 2 – Measures of Information

Information Theory
Duke University, Fall 2024

History: Written by Galen Reeves (-2023). Updated by Henry Pfister (2024-).
Last Modified: September 16, 2024

Outline of lecture:

2.1 Quantifying Information

- How much "information" do the answers to the following questions provide?

  (i) Will it rain today in Durham? (two possible answers)

  (ii) Will it rain today in Death Valley? (two possible answers)

  (iii) What is today's winning lottery number? (for the Mega Millions Jackpot, 5 white balls numbered between 1 and 70 are chosen followed by one gold ball numbered between 1 and 25. This gives $25\binom{70}{5} = 302,575,350$ combinations.)

- The amount of "information" is linked to the number of possible answers. In 1928, Ralph Hartley gave the following definition:

$$\text{Hartley Information} = \log \# \text{ answers}$$

- Hartley's measure of information is additive. The number of possible answers for two questions corresponds to the *product* of the number of answers for each question. Taking the logarithm turns the product into a sum.

- **Example:** Two questions

  ◦ What is today's winning lottery number?

$$\log_2(302575350) \approx 28(\text{bits})$$

- What are the winning lottery numbers for today and tomorrow?

$$\log_2(302575350 \times 302575350) = \log_2(302575350) + \log_2(302575350) \approx 56(\text{bits})$$

- But Hartley's information does not distinguish between likely and unlikely answers (e.g. rain in Durham vs. rain in Death Valley).
- In 1948, Shannon introduced measures of information that depend on the *probabilities* of the answers.

2.2 Entropy and Mutual Information

In thermodynamics, the word *entropy* was coined by German physicist and mathematician Rudolf Julius Emanuel Clausius in 1865 based on the Greek word *trope* ($\tau\rho o\pi\eta$) which means "a turn, or a change". In the 1870s, Boltzmann formulated the equation

$$S = k_B \log W,$$

for the entropy S in statistical mechanics where k_B is the Boltzmann constant and W is the number of microstates corresponding to the system's macrostate. As an equation, this is closely related to Hartley's measure of information but it is also congruent with the later definition by Shannon.

2.2.1 Entropy

- Let X be discrete random variable with pmf $p_X(x)$ for all x in its finite support $\mathcal{X}$. To lighten notation, we will typically use the shorthand $p(x)$.
- The **entropy** of X is defined as

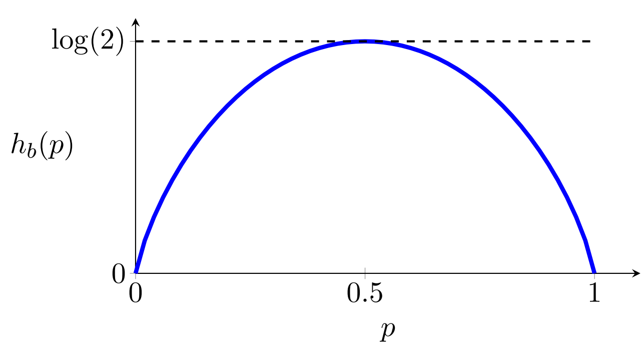
$$H(X) \coloneqq \sum_{x \in \mathcal{X}} p(x) \log\left(\frac{1}{p(x)}\right)$$

- Entropy can also be expressed as the expected value of the random variable $\log 1/p(X)$,

$$H(X) = \mathbb{E}\left[\log \frac{1}{p(X)}\right], \quad X \sim p(x)$$

- **Binary Entropy:** If X is a Bernoulli(p) random variable (i.e. $\mathbb{P}[X=1] = p$ and $\mathbb{P}[X=0] = 1-p$), then its entropy is given by the binary entropy function

$$h_b(p) \coloneqq p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} = -p \log p - (1-p) \log(1-p)$$



- ◦ The binary entropy function $h_b(p)$ is a concave. The maximum is $h_b(1/2) = \log_b(2)$ and has minimum is $h_b(0) = h_b(1) = 0$. When the context is clear, the subscript b will be dropped.

- **Example:** Two Questions
  - ◦ Will it rain Today in Durham? (say median 104 days of rain per year)
  $$h_b\left(\frac{104}{365}\right) \approx 0.862 \quad \text{bits}$$
  - ◦ Will it rain Today in Death Valley? (say median 1 day of rain per year)
  $$h_b\left(\frac{1}{365}\right) \approx 0.027 \quad \text{bits}$$

- **Fundamental Inequality:** For any base $b > 0$ and $x > 0$,
$$\left(1 - \frac{1}{x}\right) \log_b(e) \le \log_b(x) \le (x-1)\log_b(e)$$
with equalities on both sides if, and only if, $x = 1$. For the natural log, this simplifies to
$$\left(1 - \frac{1}{x}\right) \le \ln(x) \le (x-1)$$

- **Proof of upper bound:**
$$x \in (1,\infty) \qquad \Longrightarrow \qquad (x-1) - \ln(x) = \int_1^x \underbrace{\left(1 - \frac{1}{u}\right)}_{\text{strictly positive}} \mathrm{d}u > 0$$
$$x \in (0,1) \qquad \Longrightarrow \qquad (x-1) - \ln(x) = \int_x^1 \underbrace{\left(\frac{1}{u} - 1\right)}_{\text{strictly positive}} \mathrm{d}u > 0$$

- **Proof of lower bound:**
$$\ln(y) \le y - 1 \quad \Longleftrightarrow \quad 1 - y \le \ln\left(\frac{1}{y}\right) \quad \Longleftrightarrow \quad 1 - \frac{1}{x} \le \ln(x)$$

- **Theorem:** Entropy satisfies
$$0 \le H(X) \le \log|\mathcal{X}|$$

- **Proof of lower bound:** Note that $p(x) \le 1$ and so $\log 1/p(x) \ge 0$.

- **Proof of upper bound:**
$$\begin{aligned}
\sum_x p(x) \log \frac{1}{p(x)} &= \sum_x p(x) \log\left(\frac{|\mathcal{X}|}{p(x)|\mathcal{X}|}\right) \\
&= \log(|\mathcal{X}|) + \sum_x p(x) \log\left(\frac{1}{p(x)|\mathcal{X}|}\right) \\
&\le \log(|\mathcal{X}|) + \sum_x p(x) \log(e)\left(\frac{1}{p(x)|\mathcal{X}|} - 1\right) \qquad \text{Fundamental Inq.} \\
&= \log(|\mathcal{X}|) + \log(e) - \log(e) \\
&= \log(|\mathcal{X}|)
\end{aligned}$$

- The **joint entropy** of two random variables, X and Y, with joint distribution $p_{XY}(x,y)$ (written as $p(x,y)$ in shorthand) is simply the entropy of the vector (X,Y)

$$H(X,Y) \coloneqq \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{1}{p(x,y)}\right)$$

  If X and Y are independent, then $p(x,y)=p(x)p(y)$ and $\log(p(x)p(y)) = \log p(x) + \log p(y)$ together imply that $H(X,Y) = H(X) + H(Y)$.

- The entropy of an n-dimensional random vector $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$ with pmf $p(\boldsymbol{x})$ is defined as

$$H(\boldsymbol{X}) = H(X_1, X_2, \cdots, X_n) = \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})\log\left(\frac{1}{p(\boldsymbol{x})}\right)$$

  If $\boldsymbol{X}$ is a vector of iid random variables, then $H(\boldsymbol{X}) = nH(X_1)$ by induction.

- **Conditional Entropy**: The entropy of a random variable Y conditioned on the event $\{X = x\}$ is a function of the conditional distribution $p_{Y|X}(y \mid x)$ (written as $p(y \mid x)$ in shorthand):

$$H(Y \mid X = x) \coloneqq \sum_{y\in\mathcal{Y}} p(y\mid x)\log\left(\frac{1}{p(y\mid x)}\right)$$

  Averaging over x gives the conditional entropy of Y given X as a function of $p(x,y)$:

$$H(Y\mid X) = \sum_{x\in\mathcal{X}} p(x)H(Y\mid X=x) = \sum_{x,y} p(x,y)\log\left(\frac{1}{p(y\mid x)}\right)$$

- **Warning:** Note that $H(Y \mid X)$ is *not* a random variable! This is differs from the usual convention for conditioning where, for example, $\mathbb{E}[Y \mid X]$ and $\mathsf{Var}(X \mid Y)$ are random variables.

- **Chain Rule:** The joint entropy of X and Y can be decomposed as

$$H(X,Y) = H(X) + H(Y\mid X)$$

  and, more generally, for any random vector $\boldsymbol{X} = (X_1,\dots,X_n)$, we have

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

- **Proof of chain rule:**

$$\begin{aligned}
H(X,Y) &= \sum_{x,y} p(x,y)\log\left(\frac{1}{p(x,y)}\right)\\
&= \sum_{x,y} p(x,y)\log\left(\frac{1}{p(x)}\frac{1}{p(y\mid x)}\right)\\
&= \sum_{x,y} p(x,y)\left[\log\left(\frac{1}{p(x)}\right) + \log\left(\frac{1}{p(y\mid x)}\right)\right]\\
&= H(X) + H(Y\mid X)
\end{aligned}$$

  and the general result follows by induction.

2.2.2 Mutual Information

- **Mutual information** is a measure of the amount of information that one random variable contains about another random variable

$$I(X;Y) := \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) = \mathbb{E}\left[\log\left(\frac{p(X,Y)}{p(X)p(Y)}\right)\right]$$

- Note: $I(X;Y) \geq 0$ follows from the fundamental lower bound $\log_b(x) \geq \left(1-\frac{1}{x}\right)\log_b e$ and this implies that equality is achieved if and only if $p(x,y) = p(x)p(y)$.

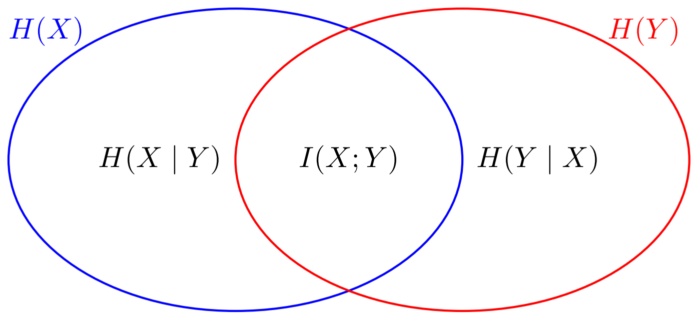
- Mutual information can also be expressed as the amount by which knowledge of X reduces the entropy of Y:

$$I(X;Y) = H(Y) - H(Y \mid X)$$
$$I(X;Y) = H(X) - H(X \mid Y)$$

- **Proof:**

$$\begin{aligned}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\left[\log\left(\frac{1}{p(y)}\right) - \log\left(\frac{1}{p(y\mid x)}\right)\right]\\ &= \underbrace{\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{1}{p(y)}\right)}_{H(Y)} - \underbrace{\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{1}{p(y\mid x)}\right)}_{H(Y|X)}\end{aligned}$$

- Venn diagram of entropy, conditional entropy, and mutual information



- The conditional mutual information between X and Y given Z is

$$I(X;Y \mid Z) := \sum_{x,y,z} p(x,y,z)\log\left(\frac{p(x,y\mid z)}{p(x\mid z)p(y\mid z)}\right)$$

- Using this, we get the chain rule for mutual information:

$$I(X;Y_1,Y_2) = I(X;Y_1) + I(X;Y_2 \mid Y_1)$$

and more generally

$$I(X;Y_1,Y_2,\cdots,Y_n) = \sum_{i=1}^{n} I(X;Y_i \mid Y_1,Y_2,\cdots,Y_{i-1})$$

2.2.3 Example: Testing for a disease

There is a 1% chance I have a certain disease. There exists a test for this disease which is 90% accurate (i.e. $\mathbb{P}[\text{test is pos} \mid \text{I have disease}] = \mathbb{P}[\text{test is neg} \mid \text{I don't have disease}] = 0.9$). Let

$$X = \begin{cases} 1, & \text{I have disease} \\ 0, & \text{I don't have disease} \end{cases} \quad \text{and} \quad Y_i = \begin{cases} 1, & i\text{th test is positive} \\ 0, & i\text{th test is negative} \end{cases}$$

Assume the the test outcomes $\boldsymbol{Y} = (Y_1, Y_2)$ are conditionally independent given X.

- The probability mass functions can be computed as

$p(x, \boldsymbol{y})$	$\boldsymbol{y} = (0,0)$	$\boldsymbol{y} = (0,1)$	$\boldsymbol{y} = (1,0)$	$\boldsymbol{y} = (1,1)$
$x = 0$	0.8019	0.0891	0.0891	0.0099
$x = 1$	0.0001	0.0009	0.0009	0.0081

  and

	$p(x)$
$x = 0$	0.99
$x = 1$	0.01

	$p(\boldsymbol{y})$
$\boldsymbol{y} = (0,0)$	0.8020
$\boldsymbol{y} = (0,1)$	0.0900
$\boldsymbol{y} = (1,0)$	0.0900
$\boldsymbol{y} = (1,1)$	0.0180

	$p(y_1)$
$y_1 = 0$	0.8920
$y_1 = 1$	0.1080

- The individual entropies are

$$H(X) = H_b(0.01) \approx 0.0808$$
$$H(Y_1) = H(Y_2) = H_b(0.1080) \approx 0.4939$$

- The conditional entropy of X given Y_1 is computed as follows:

$$H(X|Y_1 = 1) = H_b(0.9167) \approx 0.4137$$
$$H(X|Y_1 = 0) = H_b(0.0011) \approx 0.0126$$

  and so

$$H(X|Y) = \mathbb{P}[Y_1 = 1]H(X|Y_1 = 1) + \mathbb{P}[Y = 0]H(H|Y_1 = 0) \approx 0.0559$$

- The mutual information is

$$I(X; Y_1) = H(X) - H(X|Y_1) \approx 0.0249$$
$$I(X; Y_1, Y_2) = H(X) - H(X|Y_1, Y_2) \approx 0.0469$$

- The conditional mutual information is

$$I(X; Y_2|Y_1) = H(X|Y_1) - H(X|Y_1, Y_2) \approx 0.0220$$

2.2.4 Relative Entropy

- The relative entropy between a distributions p and q is defined by

$$D(p \,\|\, q) \coloneqq \sum_{x \in \mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

  This is also known as the Kullback-Leibler divergence. It can be expressed as the expectation of the expectation of the log likelihood ratio

$$D(p \,\|\, q) = \mathbb{E}[\Lambda(X)], \qquad X \sim p, \qquad \Lambda(x) = \log\left(\frac{p(x)}{q(x)}\right)$$

- Note that if there exists x such that $p(x) > 0$ and $q(x) = 0$, then $D(p \,\|\, q) = \infty$.

- **Warning:** $D(p \,\|\, q)$ is not a metric since it is not symmetric and it does not satisfy the triangle inequality.

- The mutual information between X and Y is equal to the relative entropy between $p_{X,Y}(x, y)$ and $p_X(x)p_Y(y)$,

$$I(X; Y) = D(p_{X,Y}(x, y) \,\|\, p_X(x)p_Y(y))$$

- **Theorem:** Relative entropy is nonnegative, i.e $D(p \,\|\, q) \geq 0$. It is equal to zero if and only if $p = q$.

- **Proof:**

$$\begin{aligned}
D(p \,\|\, q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\
&\geq \sum_x p(x) \log(e) \left(1 - \frac{q(x)}{p(x)}\right) && \text{Fundamental Inq.} \\
&= \log(e) \sum_x p(x) - \log(e) \sum_x q(x) \\
&= 0,
\end{aligned}$$

  with equality if and only if $p(x) = q(x)$ for all $x \in \mathcal{X}$.

- Important consequences of the non-negativity of relative entropy:

  - Mutual information is nonnegative, $I(X; Y) \geq 0$, with equality if an only if X and Y are independent.
  - This means that $H(X) - H(X|Y) \geq 0$, and thus **conditioning cannot increase entropy**,

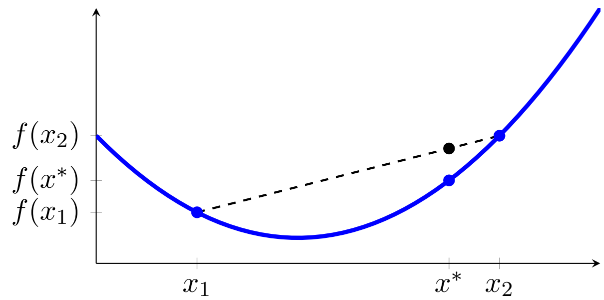
$$H(X|Y) \leq H(X)$$

  - **Warning**: Although conditioning cannot increase entropy (in expectation), it is possible that the entropy of X conditioned on an specific event, say $\{Y = y\}$, is greater than $H(X)$, i.e. $H(X|Y = y) > H(X)$. In fact, such an event must exist unless $H(X|Y = y) = H(X|Y)$ for all $y \in \mathcal{Y}$.

2.3 Convexity & Concavity

- A function $f(x)$ is convex over an interval $(a, b) \subseteq \mathbb{R}$ if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,
$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2).$$
The function is strictly convex if equality holds only if $\lambda = 0$ or $\lambda = 1$.

- Illustration of convexity. Let $x^* = \lambda x_1 + (1-\lambda)x_2$



- **Theorem:** $H(X)$ is a concave function $p(x)$, i.e.
$$H(\underbrace{\lambda p_1 + (1-\lambda)p_2}_{p^*}) \geq \lambda H(p_1) + (1-\lambda) H(p_2)$$
  - This can be proved using the fundamental inequality (try it yourself)
  - Here is an alternative proof which uses the fact that conditioning cannot increase entropy. Let Z be Bernoulli(λ)and let
$$X \sim \begin{cases} p_1, & Z = 1 \\ p_2, & Z = 0 \end{cases} \quad \iff \quad p(x,z) = \begin{cases} p_1(x), & z = 1 \\ p_2(x), & z = 0 \end{cases}$$
Then,
$$H(X) = H(\lambda p_1 + (1-\lambda)p_2)$$
Since conditioning cannot increase entropy,
$$H(X) \geq H(X|Z) = \lambda H(X|Z=1) + (1-\lambda) H(X|Z=0).$$
Combining the displays completes the proof.

- **Jensen's Inequality:** If f is a convex function over an interval $\mathcal{I}$ and X is a random variable with support $\mathcal{X} \subset \mathcal{I}$ then
$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$
Moreover, if f is strictly convex, equality occurs if and only if $X = \mathbb{E}[X]$ is a constant.

- **Example:** For any set of positive numbers $\{x_i\}_{i=1}^n$, the geometric mean is no greater than the arithmetic mean:
$$\left(\prod_{i=1}^n x_i\right)^{1/n} \leq \frac{1}{n}\sum_i^n x_i$$
Proof: Let Z be uniformly distributed on $\{x_i\}$ so that $\mathbb{P}[Z = x_i] = 1/n$. By Jensen's inequality,
$$\log\left(\prod_{i=1}^n x_i\right)^{1/n} = \frac{1}{n}\sum_{i=1}^n \log x_i = \mathbb{E}[\log(Z)] \leq \log(\mathbb{E}[Z]) = \log\left(\frac{1}{n}\sum_i^n x_i\right)$$

2.4 Data Processing Inequality

- **Markov Chain:** Random variables X, Y, Z form a Markov chain, denoted

$$X \to Y \to Z$$

  if X and Z are independent conditioned on Y.

$$p(x, z \mid y) = p(x \mid y)p(z \mid y) \quad \Longleftrightarrow \quad p(x, y, x) = p(x, y)p(z \mid y)$$

  - alternatively

$$\begin{aligned} p(x, y, z) &= p(x)p(y, z \mid x) && \text{always true} \\ &= p(x)p(y \mid x)p(z \mid x, y) && \text{always true} \\ &= p(x)p(y \mid x)p(z \mid y) && \text{if Markov chain} \\ &= p(x, y)p(z \mid y) \end{aligned}$$

  - Note $X \to Y \to Z$ implies $Z \to Y \to X$ by the a similar argument

$$p(x, z \mid y) = p(x \mid y)p(z \mid y) \iff p(x, y, z) = p(x \mid y)p(y, z) \iff p(x, y, z) = p(z)p(y \mid z)p(x \mid y).$$

  - If $Z = f(Y)$ then $X \to Y \to Z$.

- **Theorem:** (Data Processing Inequality) If $X \to Y \to Z$, then

$$I(X; Y) \geq I(X; Z)$$

- In particular, for any function g defined on $\mathcal{Y}$, we have $X \to Y \to g(Y)$ and so

$$I(X; Y) \geq I(X; g(Y)).$$

  No clever manipulation of Y can increase the mutual information!

- **Proof:** By chain rule, we can expand mutual information two different ways:

$$\begin{aligned} I(X; Y, Z) &= I(X; Z) + I(X; Y \mid Z) \\ &= I(X; Y) + I(X; Z \mid Y) \end{aligned}$$

  Since X and Z are conditionally independent given Y, we have $I(X; Z \mid Y) = 0$. Since $I(X; Y \mid Z) \geq 0$, we have

$$I(X; Y) \geq I(X; Z)$$

2.5 Fano's Inequality

- Suppose we want to estimate a random variable X from an observation Y.

- The probability of error for an estimator $\hat{X} = \phi(Y)$ is

$$P_e = \mathbb{P}\left[\hat{X} \neq X\right]$$

- **Theorem:** (Fano's Inequality) For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$,

$$h_b(P_e) + P_e \log(|\mathcal{X}|) \geq H(X \mid Y)$$

and thus

$$P_e \geq \frac{H(X \mid Y) - \log 2}{\log(|\mathcal{X}|)}$$

- **Remark:** Fano's Inequality provides a lower bound on P_e for any possible function of Y!
- **Proof of Fano's inequality:**
  - Let E be a random variable that indicates whether an error has occurred:

$$E = \begin{cases} 1, & \hat{X} = X \\ 0, & \hat{X} \neq X \end{cases}$$

  - By the chain rule, the entropy of (E, X) given $\hat{X}$ can be expanded two different ways

$$\begin{aligned} H(E, X|\hat{X}) &= H(X \mid \hat{X}) + \underbrace{H(E \mid X, \hat{X})}_{=0} \\ &= \underbrace{H(E \mid \hat{X})}_{\leq h_b(P_e)} + \underbrace{H(X \mid E, \hat{X})}_{\leq P_e \log |\mathcal{X}|} \end{aligned}$$

  - $H(X \mid \hat{X}) \geq H(X \mid Y)$ by the data processing inequality,
  - $H(E \mid X, \hat{X}) = 0$ because E is a deterministic function of X and $\hat{X}$.
  - $H(E \mid \hat{X}) \leq H(E) = H_b(P_e)$ since conditioning cannot increase entropy
  - Furthermore,

$$H(X|E, \hat{X}) = \mathbb{P}[E = 1] \underbrace{H(X|\hat{X}, E = 1)}_{=0} + \mathbb{P}[E = 0] \underbrace{H(X|\hat{X}, E = 0)}_{\leq \log |\mathcal{X}|}$$

  - Putting everything together proves the desired result.
  - For the second result, we use $h_b(P_e) \leq \log 2$.

As an example, let $\mathcal{C} \subseteq \{0, 1\}^n$ be a rate-R code (i.e., $|\mathcal{C}| = 2^{nR}$) and consider a random codeword $X^n \in \mathcal{C}$ observed through a BSC as Y^n. If $H(X^n|Y^n) \geq \epsilon n$, then Fano's inequality shows that

$$P_e \geq \frac{\epsilon n - \log 2}{nR} \geq \epsilon / R.$$

2.6 Summary of Basic Inequalities

- **Jensen's inequality:**
  - If f is a convex function then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

  - if f is a concave function then

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$$

- **Data Processing Inequality:** If $X \to Y \to Z$ form a Markov chain, then
$$I(X;Y) \geq I(X;Z)$$

- **Fano's Inequality:** If $X \to Y \to \hat{X}$ forms a Markov chain, then
$$\mathbb{P}\Big[X \neq \hat{X}\Big] \geq \frac{H(X|Y) - \log 2}{\log(|\mathcal{X}|)}$$

2.7 Axiomatic Derivation of Mutual Information [Optional]

This section is based on lecture notes from Toby Berger.

- Let X, Y denote discrete random variables with respective alphabets $\mathcal{X}$ and $\mathcal{Y}$. (Assume $|\mathcal{X}| < \infty$ and $|\mathcal{Y}| < \infty$.)
- Let $i(x,y)$ be the amount of information about event $\{X = x\}$ conveyed by learning $\{Y = y\}$
- Let $i(x,y|z)$ be the amount of information about event $\{X = x\}$ conveyed by learning $\{Y = y\}$ conditioned on the event $\{Z = z\}$
- Consider the four postulates:

  (A) **Bayesianness:** $i(x,y)$ depends only on $p(x,y)$, i.e.
  $$i(x,y) = f(\alpha,\beta)\Big|_{\substack{\alpha=p(x)\\ \beta=p(x|y)}}$$
  for some function $f : [0,1]^2 \to \mathbb{R}$.

  (B) **Smoothness:** partial derivatives of $f(\cdot,\cdot)$ exist.
  $$f_1(\alpha,\beta) = \frac{\partial f(\alpha,\beta)}{\partial \alpha}, \quad f_2(\alpha,\beta) = \frac{\partial f(\alpha,\beta)}{\partial \beta}$$

  (C) **successive revelation:** Let $y = (w,z)$. Then
  $$i(x,y) = i(x,w) + i(x,z|w)$$
  where $i(x,w) = f(p(x), p(x|w))$ and $i(x,z|w) = f(p(x|w), p(x|z,w))$ and so the function $f(\cdot,\cdot)$ must obey
  $$f(\alpha,\gamma) = f(\alpha,\beta) + f(\beta,\gamma), \quad 0 \leq \alpha,\beta,\gamma \leq 1$$

  (D) **Additivity:** If (X,Y) and (U,V) are independent, i.e. $p(x,y,u,v) = p(x,y)p(u,v)$, then
  $$i((x,u),(y,v)) = i(x,y) + i(u,v)$$
  where $i(x,u) = f(p(x,u), p(x,u|y,v)) = f(p(x)p(u), p(x|y)p(u|v))$ and so the function $f(\cdot,\cdot)$ must obey
  $$f(\alpha\gamma, \beta\delta) = f(\alpha,\beta) + f(\gamma,\delta) \quad 0 \leq \alpha,\beta,\gamma,\delta \leq 1$$

- **Theorem:** The function
$$i(x,y) = \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$
is the is the only function which satisfies our four postulates above.

2.7.1 Proof of uniqueness of $i(x, y)$

- Because of B, we can apply $\frac{\partial}{\partial \beta}$ to left and right sides of C

$$0 = f_2(\alpha, \beta) + f_1(\beta, \gamma) \quad \Longrightarrow \quad f_2(\alpha, \beta) = -f_1(\beta, \gamma)$$

Thus $f_2(\alpha, \beta)$ must be a function only of β, say $g'(\beta)$. Integrating w.r.t. β gives

$$\int f_2(\alpha, \beta) d\beta = f(\alpha, \beta) + c(\alpha)$$

i.e.

$$\int g'(\beta) d\beta = g(\beta) = f(\alpha, \beta) + c(\alpha)$$

and so

$$f(\alpha, \beta) = g(\beta) - c(\alpha)$$

- Put this back into C

$$\begin{aligned} f(\alpha, \gamma) = g(\gamma) - c(\alpha) &= g(\beta) - c(\alpha) + g(\gamma) - c(\beta) \\ &\Rightarrow c(\beta) = g(\beta) \\ &\Rightarrow f(\alpha, \beta) = g(\beta) - g(\alpha) \end{aligned}$$

- Next, write D in terms of $g(\cdot)$

$$g(\beta\delta) - g(\alpha\gamma) = g(\beta) - g(\alpha) + g(\delta) - g(\gamma)$$

Take derivative w.r.t δ of both sides to get

$$\beta g'(\beta\delta) = g'(\delta)$$

Set $\delta = 1/2$ (could be $\delta = 1$ but scared to try)

$$\beta g'(\beta/2) = g'(1/2) = K, \quad \text{a constant}$$

and so

$$g'(\beta/2) = K/\beta$$

Take the integral of both sides with respect to β to get

$$g(\beta/2) = K \ln(\beta) + C$$

So

$$g(x) = K \ln(2x) + C$$

or

$$g(x) = K \ln(x) + \tilde{C}$$

Thus

$$f(\alpha, \beta) = g(\beta) - g(\alpha) = K \ln(\beta) - K \ln(\alpha) = K \ln(\beta/\alpha)$$

- By A,

$$i(x, y) = K \ln\left(\frac{p(x|y)}{p(x)}\right)$$

Choosing K is equivalent to choosing the log base:

  - $K = 1$ corresponds to measuring information in nats
  - $K = \log_2(e)$ corresponds to measuring information in bits