Background

- $E\left[f(X)\right] = E_{p(x)}\left[f(X)\right] = E_{p(x,y)}\left[f(X)\right]$

  Proof $E_{p(x,y)}\left[f(X)\right] = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y)\, f(x) = \sum_{x\in\mathcal{X}} f(x) \sum_{y\in\mathcal{Y}} p(x,y)$

  $= \sum_{x\in\mathcal{X}} f(x) p(x) = E_{p(x)}\left[f(X)\right]$

- $P_{X|X}(x|x) = 1$.
- $p(x,x) = p(x|x)\, p(x) = 1p(x) = p(x)$
- Convention, based on continuity arguments: $0\log 0 = 0$, $0\log\frac{0}{q} = 0$, $0\log\frac{p}{0} = \infty$.
- Let $\{p(x)\}$ and $\{q(x)\}$ be the pmf for the same alphabet set $\mathcal{X}$. We say $p = q$ if $\forall x \in \mathcal{X}$ $p(x) = q(x)$.
- **Convexity**
  - Def: A function $f(x)$ is said to be **convex** (convex $\cup$) over an interval (a,b) if $\forall x_1 \forall x_2 \in (a,b)$ and $0 \le l \le 1$, $f\left(l\, x_1 + (1-l)\, x_2\right) \le l\, f(x_1) + (1-l)\, f(x_2)$.

    A function f is said to be **strictly convex** if equality holds only if $l = 0$ or $l = 1$.
    - Ex. (strict) $x^2, |x|, e^x, x\log x$ (for $x \ge 0$)
  - Def: A function f is concave (convex $\cap$) if $-f$ is convex.
    - Ex. (strict) $\log x$, $\sqrt{x}$ for $x \ge 0$.
  - A function is convex if it always lies below any chord.

    A function is concave if it always lies above any chord.
  - If the function f has a second derivative which is non-negative (positive) everywhere, then the function is convex (strictly convex).

    Proof. Let $f''(x) > 0$ $\forall x$. By Taylor's Theorem and Lagrange Remainder Theorem,

    $f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2}(x - x_0)^2$ where x^* is between x_0 and x. So, $\frac{f''(x^*)}{2}(x - x_0)^2 \ge 0$ with equality iff $x = x_0$.

    Thus, $f(x) \ge f(x_0) + f'(x_0)(x - x_0)$ with equality iff $x = x_0$.

    $\forall x_1\ \forall x_2 \ne x_1$, let $x_0 = l\, x_1 + (1-l)\, x_2$.

    Let $x = x_1$. Then, $f(x_1) \ge f(x_0) + f'(x_0)(x_1 - x_0)$

    $= f(x_0) + f'(x_0)(1-l)(x_1 - x_2)$

    with equality iff $l = 1$.

Thus, $l\, f(x_1) \geq l\, f(x_0) + f'(x_0)\, l\,(1-l)(x_1 - x_2)$ iff $l = 1$ or $l = 0$

Similarly, let $x = x_2$, then

$(1-l)\, f(x_2) \geq (1-l)\, f(x_0) - f'(x_0)(1-l)\, l\,(x_1 - x_2)$ with equality iff $l = 0$ or $l = 1$.

So, $l\, f(x_1) + (1-l)\, f(x_2) \geq f(x_0)$ with equality iff $l = 1$ or $l = 0$.

- Linear functions $ax+b$ are both convex and concave.

- Jensen's inequality

  Let $EX = \sum_{x \in \mathcal{X}} p(x)\, x$ in discrete case and $EX = \int x f(x) dx$ in the continuous case.

  If f is a convex function and X is a random variable, then $E\left[f(X)\right] \geq f(EX)$.

  Proof by induction.

  f is convex $\cup$; so, $a_1 f(x_1) + a_2 f(x_2)$ where $a_1 + a_2 = 1$. So, $E\left[f(X)\right] \geq f(EX)$ holds for $|\mathcal{X}| = 2$.

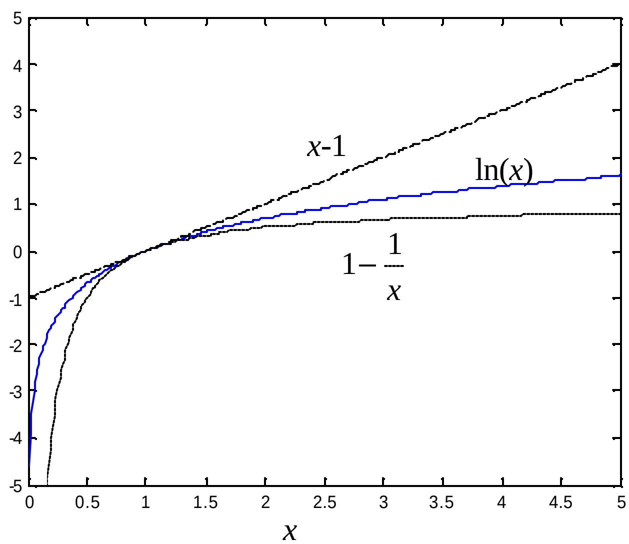
  Assume it holds for $|\mathcal{X}| = k-1$, i.e., $\sum_{i=1}^{k-1} a_i f(x_i) \geq f\left(\sum_{i=1}^{k-1} a_i x_i\right)$ where $\sum_{i=1}^{k-1} a_i = 1$.

  Then,

  If f is strictly convex, then $E\left[f(X)\right] = f(EX) \Rightarrow X = EX$ with probability 1, i.e., X is a constant.

- $\log(EX) \geq E\left[\log(X)\right]$

- **Fundamental inequality**: $\boxed{1 - \frac{1}{x} \leq \ln(x) \leq x - 1}$ with equality iff $x = 1$



Intro

- Axiomatic Derivation of Information Measure:
  Four Postulate

  A) Bayesianness: There's a function $f\left(a,b\right)$ such that $i\left(x,y\right)=f\left(a,b\right)\Big|_{\substack{a=p(x)\\ b=p(x|y)}}$.

  B) Smoothness: $f_1\left(a,b\right)=\dfrac{\partial}{\partial a}f\left(a,b\right)$ and $f_2\left(a,b\right)=\dfrac{\partial}{\partial b}f\left(a,b\right)$ exist.

  C) Successive Revelation: $f\left(a,g\right)=f\left(a,b\right)+f\left(b,g\right)$, $0\le a,b,g\le 1$.

  Justification:

  The information you get about *X* by observing $\left(W,Z\right)$ have occurred is that provided by observation that $W = w$ plus that subsequently provided by later learning that $Z = z$.

  $i\left(x,\left(w,z\right)\right)=i\left(x,w\right)+i\left(x\middle|w,z\right)$.

  $$f\left(\underbrace{p\left(x\right)}_{a},\underbrace{p\left(x\middle|w,z\right)}_{g}\right)=f\left(\underbrace{p\left(x\right)}_{a},\underbrace{p\left(x\middle|w\right)}_{b}\right)+f\left(\underbrace{p\left(x\middle|w\right)}_{b},\underbrace{p\left(x\middle|w,z\right)}_{g}\right).$$

  D) Additivity over (independent experiment): $f\left(ag,bd\right)=f\left(a,b\right)+f\left(g,d\right)$, $0\le a,b,g,d\le 1$.

  Justification:

  Consider 2 independent experiments:

  $$X\rightarrow\boxed{}\rightarrow Y$$

  $$U\rightarrow\boxed{}\rightarrow V$$

  $$p\left(x,u\right)=p\left(x\right)p\left(u\right)$$

  Then, $p\left(y,v\middle|x,u\right)=p\left(y\middle|x,u\right)p\left(v\middle|y,x,u\right)=p\left(y\middle|x\right)p\left(v\middle|u\right)$.

  $i\left(\left(y,v\right),\left(x,u\right)\right)$ should $=i\left(x,y\right)+i\left(u,v\right)$.

  $$\begin{aligned}i\left(\left(y,v\right),\left(x,u\right)\right)&=f\left(p\left(y,v\right),p\left(y,v\middle|x,u\right)\right)\\&=f\left(p\left(y\right)p\left(v\right),p\left(y\middle|x\right)p\left(v\middle|u\right)\right)\end{aligned}$$

  $$f\left(\underbrace{p\left(y\right)}_{a}\underbrace{p\left(v\right)}_{b},\underbrace{p\left(y\middle|x\right)}_{g}\underbrace{p\left(v\middle|u\right)}_{d}\right)=f\left(p\left(y\right),p\left(y\middle|x\right)\right)+f\left(p\left(v\right),p\left(v\middle|u\right)\right)$$

$A) - D) \Rightarrow\ i(x,y) = k\log\dfrac{p(x,y)}{p(x)}$

Entropy

- **Entropy of a random variable** X
  - A measure of the uncertainty of the random variable
  - A measure of the amount of information required on the average to describe the random variable.
  - Average self information of X.
  - Minimum of yes-no questions to get the value of X exactly.
  - $\boxed{\underset{\text{deterministic}}{0} \le H(X) = H(\{p(x)\}) = -E[\log p(X)] \le \underset{\text{uniform}}{\log|\mathcal{X}|}}$.

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

$$= E[i(X)]$$

$$\ge 0 \text{ with equality iff } \exists x\in \mathcal{X}\ p(x)=1$$

$$\le \log|\mathcal{X}| \text{ with equality iff } \forall x\in \mathcal{X}\ p(x)=\frac{1}{|\mathcal{X}|}$$

Proof $H(X)\ge 0$ with equality iff $\exists x\in \mathcal{X}\ p(x)=1$.

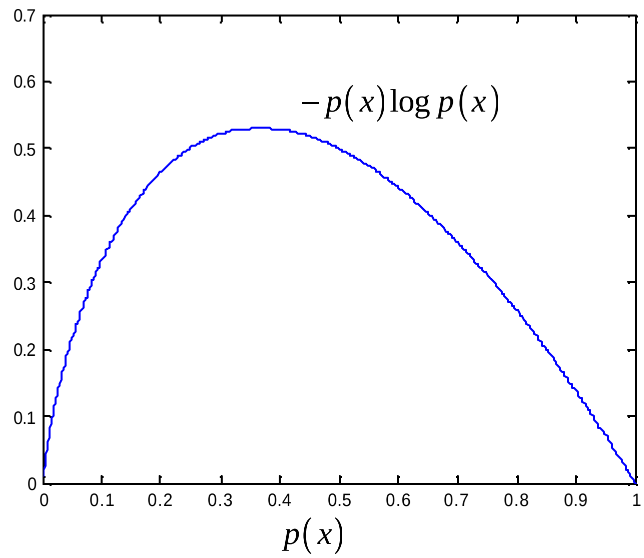
$\forall x\ p(x)$ and $-\log p(x)\ge 0$. Thus, $\forall x - p(x)\log p(x)\ge 0$.

Hence, $-\sum_{x\in\mathcal{X}} p(x)\log p(x)\ge 0$.

$H(X)=0 \Leftrightarrow \forall x - p(x)\log p(x)=0$.

But $p(x)\log p(x)=0$ if and only if $\forall x\ \ p(x)=0$ or 1.

$\forall x\ \ p(x)=0$ or 1 iff $\exists x\ p(x)=0$.



Proof $\quad H(X) \le \log|\mathcal{X}|$ with equality iff $\forall x \in \mathcal{X}\ p(x) = \frac{1}{|\mathcal{X}|}$

$$H(X) - \log|\mathcal{X}| = E[-\log p(X)] - E[|\mathcal{X}|] = E\left[\log \frac{1}{|\mathcal{X}|\, p(X)}\right]$$

$$\le E\left[\frac{1}{|\mathcal{X}|\, p(X)} - 1\right] = \sum_{x \in \mathcal{X}} p(x)\left(\frac{1}{|\mathcal{X}|\, p(X)}\right) - 1$$

$$= \sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} - 1 = \frac{|\mathcal{X}|}{|\mathcal{X}|} - 1 = 0$$

$$H(X) = \log|\mathcal{X}| \Leftrightarrow \forall x\ \frac{1}{|\mathcal{X}|\, p(x)} = 1.$$

- If the base of the logarithm is b, denote the entropy as $H_b(X)$.
- [bits] if using $\log_2(\cdot)$. [nats] if using $\log_e(\cdot)$.
- A functional of the distribution of X.
- Not depend on the actual value taken by the random variable X.
- $H(X) \ge 0$
- $H_b(X) = (\log_b a) H_a(X)$.
- Ex. entropy of a fair coin toss is $-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = -\log\frac{1}{2} = 1$.
- $H(X)$ is a function of $\{p_X(x); x \in \mathcal{X}\}$. Hence, should be written as $H(\{p_X(x)\})$.
- $H(\{p(x)\})$ is concave (convex $\cap$) in $\{p(x)\}$

$\forall l \in [0,1]$ and any two pmf $\{p_1(x), x \in \mathcal{X}\}$ and $\{p_2(x), x \in \mathcal{X}\}$,

$H(p^*) \ge l\, H(p_1) + l\, H(p_2)$ where $p^*(x) = l\, p_1(x) + (1-l)\, p_2(x)$ $\forall x \in \mathcal{X}$.

Proof

$$
\begin{aligned}
& H(p^*) - l\, H(p_1) - (1-l) H(p_2) \\
&\quad = -\sum_{x\in\mathcal{X}} p^*(x) \log p(x) \\
&\qquad + l \sum_{x\in\mathcal{X}} p_1(x) \log p(x) + (1-l) \sum_{x\in\mathcal{X}} p_2(x) \log p(x) \\
&\quad = -\sum_{x\in\mathcal{X}} \left(l\, p_1(x) + (1-l)\, p_2(x) \right) \log p^*(x) \\
&\qquad + l \sum_{x\in\mathcal{X}} p_1(x) \log p_1(x) + (1-l) \sum_{x\in\mathcal{X}} p_2(x) \log p_2(x) \\
&\quad = l \left(\sum_{x\in\mathcal{X}} p_1(x) \log \frac{p_1(x)}{p^*(x)} \right) + (1-l) \left(\sum_{x\in\mathcal{X}} p_2(x) \log \frac{p_2(x)}{p^*(x)} \right) \\
&\quad \ge l \left(\sum_{x\in\mathcal{X}} p_1(x) \left(1 - \frac{p^*(x)}{p_1(x)} \right) \right) + (1-l) \left(\sum_{x\in\mathcal{X}} p_2(x) \left(1 - \frac{p^*(x)}{p_1(x)} \right) \right) \\
&\quad = l \left(\sum_{x\in\mathcal{X}} \left(p_1(x) - p^*(x) \right) \right) + (1-l) \left(\sum_{x\in\mathcal{X}} \left(p_2(x) - p^*(x) \right) \right) \\
&\quad = l\,(1-1) + (1-l)(1-1) = 0
\end{aligned}
$$

- $\boxed{H(g(X)) \le H(X)}$ with equality iff g is one-to-one.

  Proof (1) $H(X, g(X)) = H(X) + H(g(X) \mid X)$ by chain rule.

  But $H(g(X) \mid X) = 0$; so, $H(X, g(X)) = H(X)$.

  (2) Also, by chain rule, $H(X, g(X)) = H(g(X)) + H(X \mid g(X))$.

  Because $H(X \mid g(X)) \ge 0$ with equality iff g is one-to-one, we have $H(X, g(X)) \ge H(g(X))$.

  Combining part (1) and (2), we have $H(X) \ge H(g(X))$.

- For two random variables X and Y with a joint pmf $p(x,y)$ and marginal pmf $p(x)$ and $p(y)$.
  - $\boxed{H(Y \mid X = x) = -\sum_{y\in\mathcal{Y}} p(y \mid x) \log p(y \mid x)}$.
  - **Joint entropy**: $\boxed{H(X,Y) = -E\left[\log p(X,Y)\right]} = -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(x,y)$

- **Conditional entropy** $\boxed{\underset{Y=f(X)}{0} \le H(Y|X) = -E\left[\log p(Y|X)\right] \le \underset{X,Y \text{ independent}}{H(Y)}}$

$$
\begin{aligned}
H(Y|X) &= \sum_{x\in\mathcal{X}} p(x) H(Y|X=x) \\
&= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(y|x) \\
&= -E_{p(x,y)}\left[\log p(Y|X)\right] \\
&\ge 0
\end{aligned}
$$

- Conditioning can only decrease entropy: $H(Y|X) \le H(Y)$

  Proof. $I(X;Y) = H(Y) - H(Y|X) \ge 0$.

- $H(X|X) = 0$

  Proof $p(X=y|X=x) = \begin{cases} 1 & ,y=x \\ 0 & ,y\ne x \end{cases}$

$$
p(X=y, X=x) = \begin{cases} p(x) & ,y=x \\ 0 & ,y\ne x \end{cases}
$$

$$
\begin{aligned}
H(X|X) &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{X}} p(x,y) \log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} \left(p(x,x) \log p(x|x) + \sum_{\substack{y\in\mathcal{X} \\ y\ne x}} p(x,y) \log p(y|x) \right) \\
&= -\sum_{x\in\mathcal{X}} \left(p(x) \log 1 + \cancelto{0}{\sum_{\substack{y\in\mathcal{X} \\ y\ne x}} 0 \log 0} \right) = 0
\end{aligned}
$$

- $H(g(X)|X) = 0$

  Proof $p(g(X)=y|X=x) = \begin{cases} 1 & ,y=g(x) \\ 0 & ,y\ne g(x) \end{cases}$

$$
p(g(X)=y, X=x) = \begin{cases} p(x) & ,y=g(x) \\ 0 & ,y\ne g(x) \end{cases}
$$

$$H\left(g(X)\middle|X\right) = -\sum_{x\in\mathfrak{X}}\sum_{y\in g(\mathfrak{X})} p_{X,g(X)}(x,y)\log p_{g(X)|X}\left(y\middle|x\right)$$

$$= -\sum_{x\in\mathfrak{X}}\left(\sum_{\substack{y\in g(\mathfrak{X})\\ y=g(x)}} p_{X,g(X)}(x,y)\log p\left(y\middle|x\right) + \sum_{\substack{y\in g(\mathfrak{X})\\ y\neq g(x)}} p(x,y)\log p_{g(X)|X}\left(y\middle|x\right)\right)$$

$$= -\sum_{x\in\mathfrak{X}}\left(\sum_{\substack{y\in g(\mathfrak{X})\\ y=g(x)}} p(x)\log 1 + \sum_{\substack{y\in g(\mathfrak{X})\\ y\neq g(x)}} 0\log 0\right)$$

$$= 0$$

- $H\left(g\left(\vec{X}\right)\middle|\vec{X}\right) = 0$

  Proof $H\left(g\left(\vec{X}\right)\middle|\vec{X}=x\right) = -\sum_{\vec{y}} p_{g(\vec{X})|\vec{X}}\left(\vec{y}\middle|\vec{x}\right)\log p_{g(\vec{X})|\vec{X}}\left(\vec{y}\middle|\vec{x}\right)$

  $$= -\sum_{\vec{y}=g(\vec{x})} p_{g(\vec{X})|\vec{X}}\left(\vec{y}\middle|\vec{x}\right)\log p_{g(\vec{X})|\vec{X}}\left(\vec{y}\middle|\vec{x}\right) - \sum_{\vec{y}\neq g(\vec{x})} p_{g(\vec{X})|\vec{X}}\left(\vec{y}\middle|\vec{x}\right)\log p_{g(\vec{X})|\vec{X}}\left(\vec{y}\middle|\vec{x}\right)$$

  $$= -p_{g(\vec{X})|\vec{X}}\left(g(\vec{x})\middle|\vec{x}\right)\log p_{g(\vec{X})|\vec{X}}\left(g(\vec{x})\middle|\vec{x}\right) - \sum_{\vec{y}\neq g(\vec{x})} 0\log 0$$

  $$= -1\log 1 + 0 = 0$$

  $$H\left(g\left(\vec{X}\right)\middle|\vec{X}\right) = \sum_{\vec{x}} p_{\vec{X}}(\vec{x})\, H\left(g\left(\vec{X}\right)\middle|\vec{X}=x\right) = \sum_{\vec{x}} p_{\vec{X}}(\vec{x})\,0 = 0.$$

- Chain rule: $H(X,Y) = H(X) + H\left(Y\middle|X\right)$.

  Proof $p(x,y) = p(x)p\left(y\middle|x\right)$

  - $H\left(X,Y\middle|Z\right) = H\left(X\middle|Z\right) + H\left(Y\middle|X,Z\right)$.

    Proof $p\left(x,y\middle|z\right) = p\left(x\middle|z\right)p\left(y\middle|x,z\right)$.

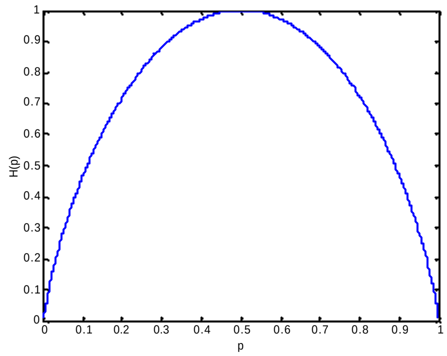
- In general $H\left(Y\middle|X\right) \neq H\left(X\middle|Y\right)$

- $H(Y) - H\left(Y\middle|X\right) = H(X) - H\left(X\middle|Y\right)$

  Proof $p(x,y) = p(x)p\left(y\middle|x\right) = p(y)p\left(x\middle|y\right)$

- $H\left(\{p(x)\,p(y)\}\right) = H\left(\{p(y)\}\right) + H\left(\{p(x)\}\right)$.

- Def: $H(p) = -p\log p - (1-p)\log(1-p)$



- Entropy of a collection of random variables.
  - Let X_1^n represents $X_1, X_2, \ldots, X_n$.
  - Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$.
  - **Joint entropy**: $H(X_1^n) = -E\left[\log p(X_1^n)\right]$

    $$H(X_1, X_2, \ldots, X_n) = -\sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} \cdots \sum_{x_n \in \mathcal{X}_n} p(x_1, x_2, \ldots, x_n) \log p(x_1, x_2, \ldots, x_n)$$
    $$= -E\left[\log p(X_1, X_2, \ldots, X_n)\right]$$
  - $H(X_1^n, Y) = H(X_1^n) + H(Y | X_1^n)$

    Proof $p(x_1^n, y) = p(x_1^n) p(y | x_1^n)$.
  - **Chain rule for entropy**: $H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1)$.

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

    Proof $p(x_1^n) = \prod_{i=1}^{n} p(x_i | x_1^{i-1})$.
  - $$\boxed{H(X_1^n) = -E\left[\log p(X_1^n)\right] = \sum_{i=1}^{n} H(X_i | X_1^{i-1}) \le \sum_{i=1}^{n} H(X_i) \underset{X_i\text{'s are independent}}{}}$$
  - $H(X, Y) = H(X) + H(Y | X) = H(Y) + H(X | Y)$
  - $H(X_1^n | Y) = \sum_{i=1}^{n} H(X_i | X_1^{i-1}, Y)$
  - $H(X, Y | Z) = H(X | Z) + H(Y | X, Z) = H(Y | Z) + H(X | Y, Z)$
- $H(X) \le \log|\mathcal{X}|$, where $|\mathcal{X}|$ denotes the number of elements in the range of X with equality if and only if X has a uniform distribution over $\mathcal{X}$.

- Conditioning reduce entropy: $H(X|Y) \le H(X)$ with equality iff X and Y are independent.

  Proof $I(X;Y) = H(X) - H(X|Y) \ge 0$ with equality iff X and Y are independent.

  - Knowing another random variable Y can only reduce the uncertainty in X.
  - No general comparison between $H(X|Y=y)$ and $H(X)$.

- Independence bound on entropy: $H(X_1^n) \le \sum_{i=1}^{n} H(X_i)$ with equality if and only if the X_i are independent.

Relative Entropy

- **Relative entropy** / Kullback Leibler "distance" between two probability mass functions $p(x)$ and $q(x)$
  - $\boxed{\underset{p=q}{0} \le D(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}} = E\left[\log\frac{p(X)}{q(X)}\right]$ [bits]
  - A measure of the inefficiency of assuming that the distribution is q when the true distribution is p.

    If we knew the true distribution $\{p(x)\}$ of the random variable, then we could construct a code with average description length $H(p)$. If, instead, we used the code for a distribution q, we would need $H(p) + D(p\|q)$ bits on the average to describe the random variable.

    Proof. $(-p(x)\log q(x)) - (-p(x)\log p(x)) = p(x)\log\frac{p(x)}{q(x)}$. So,

    $E[-\log q(X)] = E[-\log p(X)] + E\left[\log\frac{p(X)}{q(X)}\right]$.

  - ≥ 0, $= 0$ iff $p = q$.

    Proof. $D(p\|q) = \sum_x p(x)\log\frac{p(x)}{q(x)} \ge \sum_x p(x)\left(1 - \frac{q(x)}{p(x)}\right)$

    $\ge \sum_x (p(x) - q(x)) = 1 - 1 = 0$

    - Note that this just means if we have two vectors $\vec{u}, \vec{v}$ with the same lengths, each have elements which summed to 1. Then, $\sum_i u_i \log\frac{u_i}{v_i} \ge 0$.

- In general, $D(p\|q) \ne D(q\|p)$.

- Log sum inequality: for non-negative numbers, $a_1,\dots,a_n$ and $b_1,\dots,b_n$,

$$\boxed{\sum_{i=1}^{n} a_i \log\frac{a_i}{b_i} \ge \left(\sum_{i=1}^{n} a_i\right)\log\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}}$$ with equality iff $\frac{a_i}{b_i}$ = constant $\forall i$.

  Proof. Define $a_i' = \frac{a_i}{\sum_{i=1}^{n} a_i} = \frac{a_i}{A}$ and $b_i' = \frac{b_i}{\sum_{i=1}^{n} b_i} = \frac{b_i}{B}$. Then, from

  $$D(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \ge 0,$$ we have

  $$0 \le \sum_{i=1}^{n} a_i'\log\frac{a_i'}{b_i'} = \sum_{i=1}^{n}\frac{a_i}{A}\log\frac{\frac{a_i}{A}}{\frac{b_i}{B}} = \sum_{i=1}^{n}\frac{a_i}{A}\log\frac{a_i}{b_i}\frac{B}{A} = \sum_{i=1}^{n}\frac{a_i}{A}\log\frac{a_i}{b_i} - \sum_{i=1}^{n}\frac{a_i}{A}\log\frac{A}{B}$$

  $$= \frac{1}{A}\sum_{i=1}^{n} a_i\log\frac{a_i}{b_i} - \log\frac{A}{B} = \frac{1}{A}\left(\sum_{i=1}^{n} a_i\log\frac{a_i}{b_i} - A\log\frac{A}{B}\right)$$

  Thus, $\sum_{i=1}^{n} a_i\log\frac{a_i}{b_i} \ge (A)\log\frac{A}{B}$.

  Equality iff $a_i' = b_i'\ \forall i \Leftrightarrow \frac{a_i}{b_i} = \frac{A}{B}\ \forall i$.

  - $a\log\frac{a}{0} = \infty$ if $a > 0$, and $0\log\frac{0}{0} = 0$.

- Not a true distance since symmetry and triangle inequality fail. Nonetheless, it is often useful to think of it as a distance between distributions.
- $D(p\|q)$ is convex $\cup$ in the pair (p,q).

  If (p_1,q_1) and (p_2,q_2) are two pairs of probability mass functions, then

  $$D\left(l\,p_1 + (1-l)\,p_2 \,\middle\|\, l\,q_1 + (1-l)\,q_2\right) \le l\,D(p_1\|q_1) + (1-l)\,D(p_2\|q_2) \quad \forall \quad 0 \le l \le 1.$$

- For fixed p, $D(q\|p)$ is a convex $\cup$ function of q.

  $$D\left(l\,q_1 + (1-l)\,q_2 \,\middle\|\, p\right) \le l\,D(q_1\|p) + (1-l)\,D(q_2\|p).$$

  Proof.

$$
\begin{aligned}
p_0(x)\log\frac{p_0(x)}{q_0(x)} &= \left(l\, p_1(x)+(1-l)\,p_2(x)\right)\log\frac{l\,p_1(x)+(1-l)\,p_2(x)}{l\,q_1(x)+(1-l)\,q_2(x)} \\
&= A\log\frac{A}{B} \\
&\le \sum_{i=1}^{n} a_i \log\frac{a_i}{b_i} \\
&= l\,p_1(x)\log\frac{l\,p_1(x)}{l\,q_1(x)} + (1-l)\,p_2(x)\log\frac{(1-l)\,p_2(x)}{(1-l)\,q_2(x)} \\
&= l\,p_1(x)\log\frac{p_1(x)}{q_1(x)} + (1-l)\,p_2(x)\log\frac{p_2(x)}{q_2(x)}
\end{aligned}
$$

- **Conditional relative entropy** $D\left(p(y|x)\|q(y|x)\right)$

  $$D\left(p(y|x)\|q(y|x)\right) = E\left[\log\frac{p(Y|X)}{q(Y|X)}\right] = \sum_x p(x)\sum_y p(y|x)\log\frac{p(y|x)}{q(y|x)}.$$

- $D\left(p(\vec{x})\|q(\vec{x})\right) \ge 0$

  Proof. Map $\mathcal{X}^n \xrightarrow[onto]{1-1} \left\{i : i = 1,\ldots,|\mathcal{X}|^n\right\}$, $p(\vec{x}) \to u_i$, $q(\vec{x}) \to v_i$. Then, $\sum_i u_i = 1$, and $\sum_i v_i = 1$. Use $\sum_i u_i \log\frac{u_i}{v_i} \ge 0$.

- $D\left(p(x|z)\|q(x|z)\right) \ge 0$

  Proof.. For any given z, $\sum_x p(x|z) = 1$, and $\sum_x q(x|z) = 1$; thus, $\sum_x p(x|z)\log\frac{p(x|z)}{q(x|z)} \ge 0$. $D\left(p(x|z)\|q(x|z)\right) = \sum_z p(z)\underbrace{\sum_x p(x|z)\log\frac{p(x|z)}{q(x|z)}}_{\ge 0}$.

- $D\left(p(\vec{x}|\vec{z})\|q(\vec{x}|\vec{z})\right) \ge 0$

- Chain rule for relative entropy:

  $$D\left(p(x,y)\|q(x,y)\right) = D\left(p(x)\|q(x)\right) + D\left(p(y|x)\|p(y|x)\right)$$

  Proof $\dfrac{p(x,y)}{q(x,y)} = \dfrac{p(x)p(y|x)}{q(x)q(y|x)} = \dfrac{p(x)}{q(x)}\dfrac{p(y|x)}{q(y|x)}$.

Mutual Information

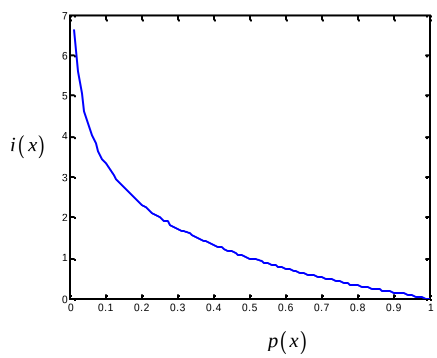
- $i(x,y) = \log \frac{p(x,y)}{p(x)\,p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}$; can be negative.

  $i(x,y) = i(y,x)$; more precisely $i(X=x, Y=y) = i(Y=y, X=x)$.

  If $p(x|y) = 1$, the mutual info is equivalent to the self-information of symbol x.

  $$i(x) = i(x,y)\Big|_{p(x|y)=1} = \log \frac{p(x|y)}{p(x)}\Bigg|_{p(x|y)=1} = \log \frac{1}{p(x)} = -\log p(x).$$

  $$i(x) = i(x,x) = \log \frac{p(x|x)}{p(x)} = \log \frac{1}{p(x)} = -\log p(x).$$



- Average Mutual information
  - A measure of the amount of information that one random variable contains about another random variable. $\left(H(X|Y) = H(X) - I(X;Y)\right)$.
  - The reduction in the uncertainty of one random variable due to the knowledge of the other.
  - A special case relative entropy.
    - Need on average $H\left(\{p(x,y)\}\right)$ info bits to describe (x,y). If instead, assume that X and Y are independent, then would need on average $H\left(\{p(x)\,p(y)\}\right) + D\left(p(x,y)\,\|\,p(x)\,p(y)\right)$ info bits to describe (x,y).
  - **Average mutual information**

$$\boxed{\underset{\text{iff independent}}{0} \le I(X;Y) = E\left[\log \frac{P(X,Y)}{p(X)\,q(Y)}\right] = E\left[\log \frac{P(X|Y)}{p(X)}\right] = E\left[\log \frac{Q(Y|X)}{q(Y)}\right]}.$$

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)} \\ &= E_{p(x,y)}\left[\log\frac{p(X,Y)}{p(X)p(Y)}\right] = E\left[i(X;Y)\right] \\ &= D\left(p(x,y)\,\|\,p(x)\,p(y)\right) \\ &\geq 0 \text{ with equality iff } X \text{ and } Y \text{ are independent} \end{aligned}$$

Proof

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)} \\ &= \frac{1}{\ln(2)}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\ln\frac{p(x,y)}{p(x)\,p(y)} \\ &\geq \frac{1}{\ln(2)}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\left[1-\left(\frac{p(x,y)}{p(x)p(y)}\right)^{-1}\right] \\ &= \frac{1}{\ln(2)}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\left[p(x,y)-p(x)\,p(y)\right] = \frac{1}{\ln(2)}(1-1) = 0 \end{aligned}$$

- $$\boxed{\begin{aligned} I(X;Y) &= H(X)+H(Y)-H(X,Y) \\ &= H(Y)-H(Y|X) = H(X)-H(X|Y) \end{aligned}}$$

  Proof. $\dfrac{p(x,y)}{p(x)\,p(y)} = \dfrac{p(y|x)}{p(y)} = \dfrac{p(x|y)}{p(x)}$.

- $I(X;X) = H(X) \Rightarrow$ entropy = self-information.

  Proof. $I(X;X) = H(X) - H(X|X) = H(X)$.

- $I(X;Y) = I(Y;X)$
  - The X says, on average, as much about Y as Y says, on average, about X.

- **<u>Conditional mutual information</u>** of random variables X and Y given Z,

  $$\boxed{\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= E_{p(x,y,z)}\log\frac{p(X,Y|Z)}{P(X|Z)p(Y|Z)} \\ &\geq 0 \text{ with equality iff } X \text{ and } Y \text{ are conditionally independent given } Z \end{aligned}}$$

  Proof. $\dfrac{p(x,y|z)}{P(x|z)p(y|z)} = \dfrac{1}{p(x|z)}\dfrac{p(y|z)p(x|y,z)}{P(y|z)} = \dfrac{1}{p(x|z)}p(x|y,z)$.

Proof. $I(X;Y|Z) = D\left(p(x,y|z) \| q(x,y|z)\right) \geq 0$ where $q(x,y|z) = p(x|z)p(y|z)$.

- $I(X_1^n;Y) = E\left[\log \dfrac{p(X_1^n,Y)}{p(X_1^n)p(Y)}\right] = \sum_{x_1\in\mathcal{X}_1}\sum_{x_2\in\mathcal{X}_2}\cdots\sum_{x_n\in\mathcal{X}_n}\sum_{y\in\mathcal{Y}} p(x_1^n,y)\log\dfrac{p(x_1^n,y)}{p(x_1^n)p(y)}$

$$= H(X_1^n) + H(Y) - H(X_1^n,Y)$$

$$= H(X_1^n) - H(X_1^n|Y) = H(Y) - H(Y|X_1^n)$$

Proof $E\left[\log \dfrac{p(X_1^n,Y)}{p(X_1^n)p(Y)}\right] = E\left[\log p(X_1^n,Y)\right] - E\left[\log p(X_1^n)\right] - E\left[\log p(Y)\right]$

$$= -H(X_1^n,Y) + H(X_1^n) + H(Y)$$

- Chain rule for information:

$$I(X_1,X_2,\ldots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_{i-1},X_{i-1},\ldots,X_1)$$

$$\boxed{I(X_1^n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_1^{i-1})}$$

Proof $I(X_1^n;Y) = H(X_1^n) - H(X_1^n|Y) = \sum_{i=1}^{n} H(X_i|X_1^{i-1}) - \sum_{i=1}^{n} H(X_i|X_1^{i-1},Y)$

$$= \sum_{i=1}^{n}\left(H(X_i|X_1^{i-1}) - H(X_i|X_1^{i-1},Y)\right) = \sum_{i=1}^{n} I(X_i;Y|X_1^{i-1})$$

$$H(X_i|X_1^{i-1}) - H(X_i|X_1^{i-1},Y) = H(X|Z) - H(X|Z,Y) = I(X;Y|Z)$$

$$= I(X_i;Y|X_1^{i-1})$$

$$I(X_1,X_2;Y) = I(X_1;Y) + I(X_2;Y|X_1)$$

Stationary Information Sources

- Consider stationary source $\{U(k)\}$. Common alphabet $\mathcal{U}$.
  - $H(U_1^n) = H(U_k^{k+n-1})$
- Per letter entropy of an L-block:
  - $\boxed{H_L = \dfrac{H(U_1^L)}{L}} = \dfrac{H(U_k^{k+L-1})}{L}$.
  - $H_{Volumetric} = \lim_{L\to\infty} H_L$
- Incremental entropy change

- $\boxed{h_L = H\left(U_L \middle| U_1^{L-1}\right)} = H\left(U_1^L\right) - H\left(U_1^{L-1}\right)$
- $H_{Incremental} = \lim_{L\to\infty} h_L$.
- For stationary Markov chain (the initial state of the Markov chain is drawn according to a stationary distribution.):

  $h_{L,markov} = H\left(U_L \middle| U_1^{L-1}\right) = H\left(U_L \middle| U_{L-1}\right) = H\left(U_2 \middle| U_1\right) \quad \forall L \ge 2$.

  $H = \lim_{L\to\infty} h_L = H\left(U_2 \middle| U_1\right) = -\sum_{u_1,u_2} p\left(u_1\right) p\left(u_2 \middle| u_1\right) \log p\left(u_2 \middle| u_1\right)$.

- $h_1 = H_1 = H\left(U_1\right) = H\left(U_k\right)$
- Both h_L and H_L are non-increasing ↓ function of L, converging to same limit H.

$\boxed{\lim_{L\to\infty} H_L = \lim_{L\to\infty} h_L = H}$. Also, $\boxed{h_L \le H_L}$ $\left(H\left(X_n \middle| X_1^{n-1}\right) \le \frac{H\left(X_1^n\right)}{n} \right)$.

Proof. $h_L \le h_{L-1}$.

$$h_L = H\left(U_L \middle| U_1^{L-1}\right) \le H\left(U_L \middle| U_2^{L-1}\right) \underset{\text{staionary}}{=} H\left(U_{L-1} \middle| U_1^{L-2}\right) = h_{L-1}.$$

Proof $\boxed{H\left(U_1^L\right) = \sum_{k=1}^{L} h_k}$

$$H\left(U_1^L\right) = \sum_{k=1}^{L} H\left(U_k \middle| U_1^{k-1}\right) = \sum_{k=1}^{L} h_k$$

Proof $h_L \le H_L$

$$H\left(U_1^L\right) = \sum_{k=1}^{L} h_k \ge \sum_{k=1}^{L} h_L = L h_L. \text{ So, } h_L \le \frac{H\left(U_1^L\right)}{L} = H_L.$$

Proof $H_L \le H_{L-1}$

$$\begin{aligned} H_L &= \frac{1}{L} H\left(U_1^{L-1}\right) + \frac{h_L}{L} = \frac{L-1}{L} \frac{H\left(U_1^{L-1}\right)}{L-1} + \frac{h_L}{L} = \frac{L-1}{L} H_{L-1} + \frac{h_L}{L} \\ &\le \frac{L-1}{L} H_{L-1} + \frac{H_L}{L} \end{aligned}$$

$$\frac{L-1}{L} H_L \le \frac{L-1}{L} H_{L-1}$$

$$H_L \le H_{L-1}$$

Proof $\lim_{L\to\infty} H_L = \lim_{L\to\infty} h_L$

From $h_L \le H_L$, $\lim_{L\to\infty} H_L \ge \lim_{L\to\infty} h_L$.

$$H_{L+M} = \frac{H\left(U_1^{L+M}\right)}{L+M} = \frac{\sum_{k=L}^{L+M} H\left(U_k \middle| U_1^{k-1}\right) + H\left(U_1^{L-1}\right)}{L+M} = \frac{\sum_{k=L}^{L+M} h_k + H\left(U_1^{L-1}\right)}{L+M}$$
$$\le \frac{\sum_{k=L}^{L+M} h_L + H\left(U_1^{L-1}\right)}{L+M} = \frac{(M+1)h_L + H\left(U_1^{L-1}\right)}{L+M}$$

Take $M \to \infty$. $\lim_{M\to\infty} H_{L+M} \le h_L$

Take $L \to \infty$. $\lim_{L\to\infty}\lim_{M\to\infty} H_{L+M} = \lim_{L\to\infty} H_L \le \lim_{L\to\infty} h_L$.

- **<u>Entropy rate</u>** of stationary source $\{U_k\}$

  $\boxed{H\left(\{U_\ell\}\right) = H_U = \lim_{L\to\infty} \frac{H\left(U_1^L\right)}{L} = \lim_{L\to\infty} H\left(U_L \middle| U_1^{L-1}\right)}$.

  - So, for stationary source, the entropy $H\left(U_1^L\right)$ grows (asymptotically) linearly with L at a rate H_U.
  - For stationary Markov chain of order r, $H_U = H\left(U_{r+1} \middle| U_1^r\right) = h_{r+1}$.
  - For stationary Markov chain of order 1, $H_U = H\left(U_2 \middle| U_1\right) = h_2 < H(U_1) = H(U_2)$.
    - Let $\{X_i\}$ be a stationary Markov chain with stationary distribution $\vec{u}$ and transition matrix P. Then, the entropy rate is $H = -\sum_{ij} u_i P_{ij} \log P_{ij}$.

      $P_{ij} = \Pr\left[\text{Next state is } j \middle| \text{Current state is } i\right] = \Pr\left[X_2 = j \middle| X_1 = i\right]$.
      $u_i = \Pr\left[X_1 = i\right]$.

    - More than one communicating class: $H_U = \sum_i \Pr[\text{class}_i] H\left(U_2 \middle| U_1, \text{class}_i\right)$.
  - The best achievable data compression.

Variable-length (VL) lossless source codes

- Stationary discrete memoryless?? source $\{U_k\}$, finite alphabet $\mathcal{U}$.
- A variable-length D-ary source codes is a mapping $f : \mathcal{U} \to \{0,\ldots,D-1\}^*$
  - binary $D = 2$
  - D = coding alphabet cardinality.
- Def: f is **<u>uniquely decipherable</u>** if $\forall M \forall N$ and any $\underline{U} = (U_1,\ldots,U_M)$, $\underline{U}' = (U_1',\ldots,U_N')$, $f(\underline{U}) = f\left(\underline{U}'\right) \Rightarrow \underline{U} = \underline{U}'$.

  (No two distinct source strings get mapped into same code string.

- $\ell(u)$ is length of D-ary string $f(u)$.
- Def: $\bar{\ell} =$ the mean code word length $= E\left[\ell(u)\right] = \sum_{u \in \mathcal{U}} p(u)\ell(u)$.
- Optimum = min $\bar{\ell}$, uniquely decipherable.
- **<u>Morse's principle</u>**: To minimize $\bar{\ell}$, if $p(u) = \Pr[U_k = u]$ is small, make $\ell(u)$ large, and conversely.
- **<u>Prefix code</u>**: no short code word is prefix of a longer one. $\Rightarrow$ uniquely decipherable
- **<u>Kraft Inequality</u>** $\boxed{\text{(KI): } \sum_{u \in \mathcal{U}} D^{-\ell(u)} \le 1.}$
- Property of a length set $\{\ell(u), u \in \mathcal{U}\}$

  (1) If $\{\ell(u), u \in \mathcal{U}\}$ satisfying KI, then there exists a prefix code (hence, UD) with these lengths.

  (2) Every D-ary UD code has word lengths $\{\ell(u), u \in \mathcal{U}\}$ that satisfy KI.

  $\boxed{\text{KI} \Rightarrow \exists \text{ prefix (UD)}}$, $\boxed{\text{UD (including prefix)} \Rightarrow \text{KI}}$

  We are looking for a UD code with min $\bar{\ell}$. Suppose we find one. Because it is UD, from (2), it's length set satisfies KI. Then, (1) tells us that there exists a prefix code with exactly the same length set and thus also minimize $\bar{\ell}$. So, (1) and (2) let us restrict search for optimal code to prefix codes.

  Proof (2)

  Consider L-vector $\underline{u} = (u_1, u_2, \ldots, u_L)$.

  $f(\underline{u}) = f(u_1) f(u_2) \cdots f(u_L)$. $\ell(\underline{u}) = \ell(u_1) + \ell(u_2) + \cdots + \ell(u_L)$.

$$\begin{aligned}
\sum_{\underline{u} \in \mathcal{U}^L} D^{-\ell(\underline{u})} &= \sum_{\underline{u} \in \mathcal{U}^L} D^{-(\ell(u_1) + \ell(u_2) + \cdots + \ell(u_L))} = \sum_{\underline{u} \in \mathcal{U}^L} D^{-\ell(u_1)} D^{-\ell(u_2)} \cdots D^{-\ell(u_L)} \\
&= \sum_{u_1 \in \mathcal{U}} \sum_{u_2 \in \mathcal{U}} \cdots \sum_{u_L \in \mathcal{U}} D^{-\ell(u_1)} D^{-\ell(u_2)} \cdots D^{-\ell(u_L)} \\
&= \left(\sum_{u_1 \in \mathcal{U}} D^{-\ell(u_1)} \right) \left(\sum_{u_2 \in \mathcal{U}} D^{-\ell(u_2)} \right) \cdots \left(\sum_{u_L \in \mathcal{U}} D^{-\ell(u_L)} \right) = \left(\sum_{u \in \mathcal{U}} D^{-\ell(u)} \right)^L
\end{aligned}$$

  So, we have $\sum_{\underline{u} \in \mathcal{U}^L} D^{-\ell(\underline{u})} = \left(\sum_{u \in \mathcal{U}} D^{-\ell(u)} \right)^L$. (*)

  Let $\ell_{\min} = \min_u \ell(u)$, $\ell_{\max} = \max_u \ell(u)$.

  A_n = the number of $\underline{u} \in \mathcal{U}^L$ such that $\ell(\underline{u}) = n$.

Note that $L\ell_{\min} \le n \le L\ell_{\max}$. The maximum is attained when every u_k in $\underline{u}$ corresponds to $\ell_{\max}$. The minimum is attained when every u_k in $\underline{u}$ corresponds to $\ell_{\min}$. Also, $\sum_{\underline{u}\in \mathcal{U}^L} D^{-\ell(\underline{u})} = \sum_{n=L\ell_{\min}}^{L\ell_{\max}} A_n D^{-n}$.

UD implies that $A_n \le D^n$. (There are only D^n different code sequences of length n. If $A_n > D^n$, then there are at least two $\underline{u}$ which map to the same code sequence.)

$$\sum_{\underline{u}\in \mathcal{U}^L} D^{-\ell(\underline{u})} = \sum_{n=L\ell_{\min}}^{L\ell_{\max}} A_n D^{-n} \le \sum_{n=L\ell_{\min}}^{L\ell_{\max}} D^n D^{-n} = \sum_{n=L\ell_{\min}}^{L\ell_{\max}} 1 = L\ell_{\max} - L\ell_{\min} + 1$$
$$\le L\ell_{\max}$$

So, UD requires $\sum_{\underline{u}\in \mathcal{U}^L} D^{-\ell(\underline{u})} \le L\ell_{\max}$. (**).

Combining (*) and (**), we have $\left(\sum_{u\in\mathcal{U}} D^{-\ell(u)}\right)^L \le L\ell_{\max}$, or equivalently,

$\sum_{u\in\mathcal{U}} D^{-\ell(u)} \le \left(L\ell_{\max}\right)^{\frac{1}{L}}$. This has to be true for all L.

Note that $\left(L\ell_{\max}\right)^{\frac{1}{L}}$ is strictly decreasing as L increase. $\lim_{L\to\infty}\left(L\ell_{\max}\right)^{\frac{1}{L}} = 1$. Thus, $\left(L\ell_{\max}\right)^{\frac{1}{L}}$ can get arbitrary close to 1 from above. If $\sum_{u\in\mathcal{U}} D^{-\ell(u)} > 1$, there will exist L_0 such that $\sum_{u\in\mathcal{U}} D^{-\ell(u)} > \left(L\ell_{\max}\right)^{\frac{1}{L}}$ for all $L > L_0$. So, to have $\sum_{u\in\mathcal{U}} D^{-\ell(u)} \le \left(L\ell_{\max}\right)^{\frac{1}{L}}$, need $\sum_{u\in\mathcal{U}} D^{-\ell(u)} \le 1$.

Proof (1) by induction

We will show that we can embed these KI satisfying word lengths as the terminal nodes in a D-ary branching tree.

Note that putting a terminal node on level ℓ prunes away $D^{L-\ell}$ nodes from level $L \ge \ell$.

Suppose each u such that $\ell(u) \le \ell - 1$ has been assigned a terminal node on level $\ell(u)$. Now, we want to assign terminal nodes on level ℓ to all u such that $\ell(u) = \ell$. We then need there to be at least $\left|\{u : \ell(u) = \ell\}\right|$ nodes on level ℓ not yet pruned away.

Originally, there were D^{ℓ} nodes on level ℓ. We have pruned away $\sum\limits_{\substack{u \\ \ell(u)\le \ell-1}} D^{\ell-\ell(u)}$

of them. So, we need $D^{\ell} - \sum\limits_{\substack{u \\ \ell(u)\le \ell-1}} D^{\ell-\ell(u)} \ge \left|\left\{u : \ell(u) = \ell\right\}\right|$.

Trick: $\left|\left\{u : \ell(u) = \ell\right\}\right| = \sum\limits_{\substack{u \\ \ell(u)=\ell-1}} 1 = \sum\limits_{\substack{u \\ \ell(u)=\ell-1}} D^{\ell-\ell(u)}$.

So, need $1 \ge \sum\limits_{\substack{u \\ \ell(u)\le \ell}} D^{-\ell(u)}$.

If $\left\{\ell(u), u \in \mathcal{U}\right\}$ satisfying KI, then $\sum\limits_{u\in\mathcal{U}} D^{-\ell(u)} \le 1$, and therefore,

$\sum\limits_{\substack{u \\ \ell(u)\le \ell}} D^{-\ell(u)} \le \sum\limits_{u\in\mathcal{U}} D^{-\ell(u)} \le 1$.

- For any UD D-ary code, and any distribution $\left\{p(u), u \in \mathcal{U}\right\}$,
  $\overline{\ell} \ge H_D\left(\left\{p(u)\right\}\right) = -\sum\limits_{u\in\mathcal{U}} p(u)\log_D p(u)$.

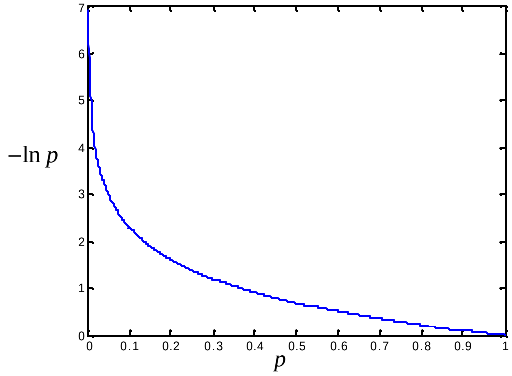
  Proof. $\overline{\ell} - H_D\left(\left\{p(u)\right\}\right)$

$$
\begin{aligned}
&= \sum_{u\in\mathcal{U}} p(u)\ell(u) + \sum_{u\in\mathcal{U}} p(u)\log_D p(u) = \sum_{u\in\mathcal{U}} p(u)\left(\ell(u) + \log_D p(u)\right) \\
&= \sum_{u\in\mathcal{U}} p(u)\left(\log_D D^{\ell(u)} + \log_D p(u)\right) = \sum_{u\in\mathcal{U}} p(u)\left(\log_D p(u) D^{\ell(u)}\right) \\
&= \frac{1}{\ln D}\sum_{u\in\mathcal{U}} p(u)\ln p(u) D^{\ell(u)} \\
&\ge \frac{1}{\ln D}\sum_{u\in\mathcal{U}} p(u)\left(1 - \frac{1}{p(u) D^{\ell(u)}}\right) = \frac{1}{\underbrace{\ln D}_{>0}}\left(\sum_{u\in\mathcal{U}} p(u) - \sum_{u\in\mathcal{U}} D^{-\ell(u)}\right) \\
&\overset{(a)}{\ge} \frac{1}{\ln D}(1-1) = 0
\end{aligned}
$$

(a) Code is UD; thus length set satisfies KI.

- $\boxed{KI \Rightarrow \overline{\ell} \ge \underset{\ell_i = \log_D p_i}{H_D\left(\left\{p_i\right\}\right)}}$

- Shannon-Fano codes: $\left\{\ell(u) = \left\lceil -\log_D p(u)\right\rceil = \left\lceil i(u)\right\rceil, u \in \mathcal{U}\right\}$.



- This length assignment is possible because it satisfies KI.

  Proof. Because $\lceil -\log_D p(u) \rceil \ge -\log_D p(u) \ge 0$, $-\lceil -\log_D p(u) \rceil \le \log_D p(u)$, and $D^{-\lceil -\log_D p(u) \rceil} \le D^{\log_D p(u)}$.

  Thus, $\sum_{u \in \mathcal{U}} D^{-\ell(u)} = \sum_{u \in \mathcal{U}} D^{-\lceil -\log_D p(u) \rceil} \le \sum_{u \in \mathcal{U}} D^{\log_D p(u)} = \sum_{u \in \mathcal{U}} p(u) = 1$.

- $H_D(\{p(u)\}) \le \overline{\ell}_{SF} < 1 + H_D(\{p(u)\})$

  Proof. $1 + -\log_D p(u) \ge \lceil -\log_D p(u) \rceil \ge -\log_D p(u)$

  So, $1 + E[-\log_D p(u)] > E\lceil -\log_D p(u) \rceil \ge E[-\log_D p(u)]$.

  Hence, $1 + H_D(\{p(u)\}) > \overline{\ell}_{SF} \ge H_D(\{p(u)\})$.

- If $-\log_D p(u)$ is an integer (***D*-adic**) for all $u \in \mathcal{U}$, then $\overline{\ell}_{SF} = H_D(\{p(u)\})$.

  - Ex. for $D = 2$, $\{p(u)\} = \left\{ \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots, \frac{1}{2^n}, \frac{1}{2^n} \right\}$.

  - Ex. for general D, $\{p(u)\} = \left\{ \underbrace{\frac{1}{D}, \ldots, \frac{1}{D}}_{D-1 \text{ times}}, \underbrace{\frac{1}{D^2}, \ldots, \frac{1}{D^2}}_{D-1 \text{ times}}, \ldots, \underbrace{\frac{1}{D^n}, \ldots, \frac{1}{D^n}}_{D-1 \text{ times}}, \frac{1}{D^n} \right\}$.

    Proof $\quad (D-1)\dfrac{\frac{1}{D} - \frac{1}{D^{n+1}}}{1 - \frac{1}{D}} + \frac{1}{D^n} = \left(1 - \frac{1}{D^n}\right) + \frac{1}{D^n} = 1$.

- $\boxed{H_D(\{p(u)\}) \le \overline{\ell}_{opt} \le \overline{\ell}_{SF} < 1 + H_D(\{p(u)\})}$.
- Block-to-variable length codes
  - Instead of $f : \mathcal{U} \to \{0, \ldots, D-1\}^*$, use $f : \mathcal{U}^L \to \{0, \ldots, D-1\}^*$ with corresponding $\{p(\underline{u}), \underline{u} \in \mathcal{U}^L\}$.
  - Super-letters.

- Super letter source is still stationary.
- Entropy rate per letter is the same as that of original source (H).

  Proof For original source, entropy rate per letter $H = \lim_{L\to\infty} H_L = \lim_{L\to\infty} \frac{H\left(U_1^L\right)}{L}$.

  For the new one, entropy rate per super letter =

  $H_{\sup er} = \lim_{n\to\infty} \frac{H\left(\underline{U}_1^n\right)}{n} = \lim_{n\to\infty} \frac{H\left(U_1^{Ln}\right)}{n}$. Thus, entropy rate per letter

  $= \lim_{n\to\infty} \frac{H\left(U_1^{Ln}\right)}{Ln}$. Note that sequence $\frac{H\left(U_1^{Ln}\right)}{Ln}$ is a subsequence of $\frac{H\left(U_1^n\right)}{n}$.

  Because the sequence $\frac{H\left(U_1^n\right)}{n}$ converges, the subsequence converge to the same limit.
- 2-L delay and extra complexity.
- As $L \to \infty$, $\bar{\ell} \to H$, the source's entropy rate.

  Proof
  $$\begin{aligned}\bar{\ell} &= \frac{E\left[\left\lceil -\log p\left(\underline{U}\right)\right\rceil\right]}{L} \\ &< \frac{E\left[1-\log p\left(\underline{U}\right)\right]}{L} = \frac{1+E\left[-\log p\left(\underline{U}\right)\right]}{L} \\ &= \frac{1+H\left(U_1^L\right)}{L} = \frac{1}{L} + H_L\end{aligned}$$

  Note: $E\left[-\log p\left(\underline{U}\right)\right] = H\left(U_1^L\right)$ because the source is stationary.

  $\lim_{L\to\infty} \bar{\ell} \le H_L = H$. And we already know that treating a super letter as normal letter, $H\left(U_1^L\right) \le \bar{\ell}_{\sup er} < H\left(U_1^L\right) + 1$. So, $\frac{H\left(U_1^L\right)}{L} \le \frac{\bar{\ell}_{\sup er}}{L} < \frac{H\left(U_1^L\right)}{L} + \frac{1}{L}$, and thus $\lim_{L\to\infty} \frac{\bar{\ell}_{\sup er}}{L} = \lim_{L\to\infty} \frac{H\left(U_1^L\right)}{L} = H$.

- Huffman code, $D = 2$.

  Given $\left\{p_j, 0 \le j \le M-1\right\}$. Want to build a minimum $\bar{\ell}$ binary prefix code by assigning lengths $\left\{\ell_j, 0 \le j \le M-1\right\}$ that minimize $\bar{\ell} = \sum_{j=0}^{M-1} \ell_j p_j$.

  $M = \left|\mathcal{U}\right|$. Assume $p_0 \ge p_1 \ge \cdots \ge p_{M-1}$.

  Assume that we have optimal length described by $\left\{\ell_j, 0 \le j \le M-1\right\}$.

  Let $\ell_{\max}$ be the longest of the optimum ℓ_j's.

Least likely source symbol has $\ell = \ell_{\max}$. (Morse. If not, switch its assignment with the one that has $\ell_{\max}$ will give lower $\bar{\ell}$.)

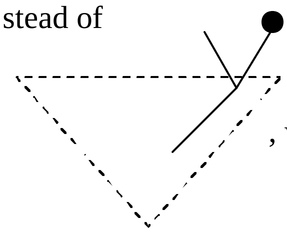
Next-to-least likely source symbol should have $\ell = \ell_{\max}$ also.

Suppose not. Assume the next-to-least likely source symbol has $\ell < \ell_{\max}$.

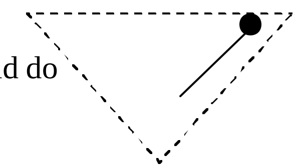
Then, note that no other symbols can have $\ell = \ell_{\max}$. It has larger probability than the next-to-least likely symbol; so, it should not be assigned larger ℓ.

This means the least likely source symbol is the only one on the level ℓ. This is not optimal because

in stead of



, we should do



Without loss of generality, let's have code strings for these two letters identical through level $\ell_{\max}$-1. (Then, one of them ends with 0, the other with 1.) This means they are assigned a common ancestor on level $\ell_{\max} - 1$.

Then, define new alphabet set with $|\mathcal{U}'| = M - 1$.

$p'_i = p_i$ for $0 \le i \le M-3$. $p'_{M-2} = p_{M-2} + p_{M-1}$.

$$\bar{\ell} = \sum_{k=0}^{M-1} \ell_k p_k = \sum_{k=0}^{M-3} \ell_k p_k + \ell_{\max}\left(p_{M-2} + p_{M-1}\right)$$

$$= \sum_{k=0}^{M-3} \ell'_k p'_k + \ell_{\max}\left(p'_{M-2}\right) = \sum_{k=0}^{M-3} \ell'_k p'_k + \left(\ell'_{M-2} + 1\right)\left(p'_{M-2}\right)$$

$$= \sum_{k=0}^{M-2} \ell'_k p'_k + p'_{M-2}$$

Because p'_{M-2} is constant, we then want to minimize $\bar{\ell}' = \sum_{k=0}^{M-2} \ell'_k p'_k$.

This can be accomplished by recursively applying the above argument.

- D-ary Huffman code, $D \ge 2$.
  - Full D-ary tree: one with D branches out of every internal node.
  - Full tree has $D + k(D-1)$ terminal nodes for some non-negative integer k. So, if $|\mathcal{U}|$ is not in that form, then can't have full tree.

    First full-fan has D terminal nodes. Growing from this fan, adding one full-fan takes one terminal node off; so, net increase = $D - 1$.
  - Optimal code should have full tree except one of the top.

    If any one at lower level is not full, then can move one from the top down and reduce $\bar{\ell}$.

- If there exists k such that $|\mathcal{U}| = D + k(D-1)$, then, hang D least likely off common ancestor and proceed iteratively as in binary.

  If no such k exists, there is a non-full fan of the least likely letters' terminals on $\ell_{\max}$. Let N = size of this fan. Then, $2 \le N \le D$.

  $|\mathcal{U}| = D + k(D-1) + N - 1$ because we add one fan of N terminal nodes to a full tree. This adds N terminal nodes but takes out 1 terminal node.

  $$\begin{aligned}|\mathcal{U}| &= D + k(D-1) + N - 1 = (k+1)(D-1) + 1 + N - 1 \\ &= (k+1)(D-1) + N \\ &= (k+1)(D-1) + 2 + (N-2)\end{aligned}$$

  Note that $0 \le N-2 \le D-2$. Therefore, $(N-2) \bmod (D-1) = N-2$.

  So, $(|\mathcal{U}|-2) \bmod (D-1) = N-2$. So, $N = 2 + (|\mathcal{U}|-2) \bmod (D-1)$.

  Grab $N = 2 + (|\mathcal{U}|-2) \bmod (D-1)$ least likely at first, then always D at a time.

- For $D = 3$, $N = \begin{cases} 2, & \text{even } |\mathcal{U}| \\ 3, & \text{odd } |\mathcal{U}| \end{cases}$

* Universal lossless coding
  - Borisfitingof

    Discrete memoryless stationary source with alphabet $\mathcal{U}$ but unknown distribution $\{p(u), u \in \mathcal{U}\}$.

    Encoding: Gather n-block, $\underline{u} = (u_1, \ldots, u_n)$. Compute empirical distribution

    $\tilde{p}(u) = \frac{n(u)}{n}$. $n(u) = |\{k : u_k = u\}|$.

    Also need to send $\{\tilde{p}(u)\}$. Note that $0 \le n(u) \le n$ for every $u \in \mathcal{U}$. So, $\{\tilde{p}(u)\}$ cannot assume more than $(n+1)^{|\mathcal{U}|}$ values. So, take no more than $\left\lceil \log_2 (n+1)^{|\mathcal{U}|} \right\rceil$ binary digits to specify $\{\tilde{p}(u), u \in \mathcal{U}\}$.

    Thus, per source letter, use fewer than $\frac{\left\lceil \log_2 (n+1)^{|\mathcal{U}|} \right\rceil}{n}$. As $n \to \infty$, this $\to 0$.

  - Lempel-Ziv (LZ codes)
  - Arithmetic (Pasco, Rissanen, Langdon)