Elements of Information Theory

Varun Varanasi

April 24, 2024

Contents

1 Introduction and Preview

2 Entropy, Relative Entropy, and Mutual Information

2.1 Entropy

Definition 1 (Entropy) *Entropy is understood as a quantification of uncertainity in a variable.*

$$H(X) = -\sum p(x) \log p(x)$$

Intuitively, entropy can be understood as the average length of the smallest description of a random variable. Entropy is typical calculated in base 2 as bits, but if entropy is measured with a natural logarithm we refer to the units of information as nats. Furthermore, by convention we set $0 \log 0 = 0$ to allow entropy to be invariant to additions of 0 probability events. Entropy can also be seen as a functional on the distribution of X (note that it is independent of the realized values of X).

Entropy can equivalently be understood as the expected value of $\log \frac{1}{p(X)}$.

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

This formulation of entropy is reminscent of entropy in thermodynamics.

One of the key properties of entropy is its positivity.

Lemma 2.1 $H(X) \geq 0$

Proof: $0 \leq p(x) \leq 1$ implies $\log \frac{1}{p(x)} \geq 0$

Lemma 2.2 $H_b(X) = \log_b(a) H_a(X)$

Proof: $\log_p(b) = \log_b(a) \log_a(p)$
Consider a bernoulli random variable that realizes 1 with probability p.

$$H(X) = -p \log p - (1-p) \log(1-p)$$

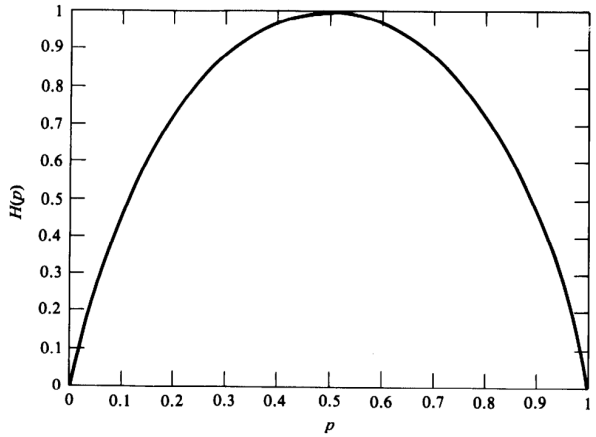


Figure 2.1. $H(p)$ versus p.

Note that the entropy curve is concave and realizes 0 when $p = 0$ or $p = 1$. Maximal entropy is achieved when $p = 1/2$.

2.2 Joint Entropy and Conditional Entropy

Definition 2 (Joint Entropy) *The joint entropy for a pair of discrete random variables is given by*

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y) = -\mathbb{E}[\log p(X,Y)]$$

Definition 3 (Conditional Entropy) *For $(X,Y) \sim p(x,y)$, the conditional entropy is given by*

$$\begin{aligned} H(Y|X) &= -\sum_x p(x) H(Y|X=x) \\ &= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) \\ &= -\sum_x \sum_y p(x,y) \log p(y|x) \\ &= -\mathbb{E}[\log p(Y|X)] \end{aligned}$$

Theorem 2.3 (Chain Rule) *Entropy of a pair of random variables can be decomposed into entropy of a single variable and the conditional entropy*

$$H(X,Y) = H(X) + H(Y|X)$$

Proof:

$$\begin{aligned} H(X,Y) &= -\sum_x \sum_y p(x,y) \log p(x,y) \\ &= -\sum_x \sum_y p(x,y) \log p(x)p(y|x) \\ &= -\sum_x \sum_y p(x,y) \log p(x) - \sum_x \sum_y p(x,y) \log p(y|x) \\ &= -\sum_x p(x,y) \log p(x) - \sum_x \sum_y p(x,y) \log p(y|x) \\ &= H(X) + H(Y|X) \end{aligned}$$

Alternatively, you can write $\log p(x,y) = \log p(x) + \log p(y|x)$ use expectations to show this property.

Corollary 2.3.1 *The chain rule can be extended to include*

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$$

2.3 Relative Entropy and Mutual Information

Relative entropy is a measure of the distance between two distributions. It can also be understood as the inefficiency metric of assuming that the distribution is q when the true distribution follows$p$.

Definition 4 (Relative Entropy) *The relative entropy or kullback Leibler distance between two probability mass functions is given by:*

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \left[\log \frac{p(x)}{q(x)} \right]$$

Relative entropy is non-negative and is zero if and only if $p = q$. Also note that relative entropy is not a true distance metric since it is not symmetric and does not abide by the triangle inquality.

Mutual information is a measure of the amount of information one random variable contains about another. It is the reduction in uncertainity in one variable due to another.

Definition 5 (Mutual Information) *The mutual information between two distributions is the relative entropy between their joint distribution and their product distribution*

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

2.4 The Relationship between Entropy and Mutual Information

$$\begin{aligned} I(X,Y) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\ &= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y) \\ &= H(X) - H(X|Y) \end{aligned}$$

By symmetry it follows that $I(X,Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$. Putting together our joint entropy decomposition with our definition of mutual entropy, we can write the expression:

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

With our mutual information definition we can also see the self-information interpretation of entropy:

$$I(X,X) = H(X) - H(X|X) = H(X)$$

These results can be summarized by the following diagram:

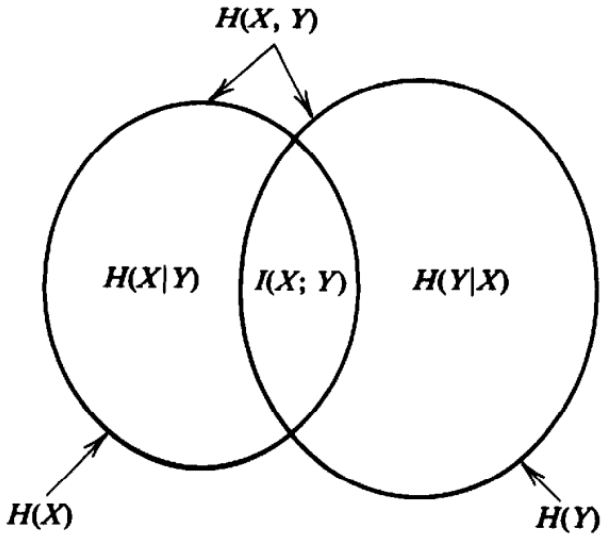


Figure 2.2. Relationship between entropy and mutual information.

2.5 Chain Rules for Entropy, Relative Entropy, and Mutual Information

Theorem 2.4 (Chain Rule for Entropy) *Let $X_1, X_2, ...X_n$ be drawn according to $p(x_1, ..., x_n)$.*

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

Proof 1:

$$\begin{aligned} H(X_1, X_2) &= H(X_1) + H(X_2|X_1) \\ H(X_1, X_2, X_3) &= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) \\ &... \\ H(X_1, X_2, ..., X_n) &= \sum_{i=1}^{n} H(X_i|X_{i-1}, ..X_1) \end{aligned}$$

Proof 2: Denote $p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i|x_{i-1}...x_1)$

$$\begin{aligned} H(X_1, X_2, ..., X_n) &= -\sum_{x_i} p(x_1, ..., x_n) \log p(x_1, ..., x_n) \\ &= -\sum_{x_i} p(x_1, ..., x_n) \log \prod_{i=1}^{n} p(x_i|x_{i-1}...x_1) \\ &= -\sum_{x_i} \sum_{i=1}^{n} p(x_1, ..., x_n) \log p(x_i|x_{i-1}...x_1) \\ &= \sum_{i=1}^{n} H(X_i|X_{i-1}, ..X_1) \end{aligned}$$

Definition 6 (Conditional Mutual Information) *is the reduction in uncertainity of a random variable due to the knowledge of Y conditional on Z.*

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = \mathbb{E}\left[\log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}\right]$$

Theorem 2.5 (Chain Rule for Mutual Information)

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

Proof:

$$\begin{aligned} I(X_1, X_2, ...X_n; Y) &= H(X_1, ...X_n) - H(X_1, ...X_n|Y) \\ &= \sum_{i=1}^{n} H(X_i|X_{i-1}, ..X_1) - \sum_{i=1}^{n} H(X_i|X_{i-1}, ..X_1, Y) \qquad = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, ...X_1) \end{aligned}$$

Definition 7 (Conditional Relative Entropy) *is the average relative entropy between conditional probability mass functions $p(y|x)$ and $q(y|x)$ averaged over $p(x)$.*

$$D\big(p(y|x)\big|\big|q(y|x)\big) = \sum_{x} p(x) \sum_{y} p(y|x) \log \frac{p(y|x)}{q(y|x)} = \mathbb{E}_{p(x)}\left[\log \frac{p(Y|X)}{q(Y|X)}\right]$$

Theorem 2.6 (Chain Rule for Conditional Relative Entropy)

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

Proof:

$$\begin{aligned} D\big(p(x,y)\big|\big|q(x,y)\big) &= \sum_{x} \sum_{y} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\ &= \sum_{x} \sum_{y} p(x,y) \log \frac{p(y|x)p(x)}{q(y|x)q(x)} \\ &= \sum_{x} \sum_{y} p(x,y) \log \frac{p(x)}{q(x)} + \sum_{x} \sum_{y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \\ &= D\big(p(x)\big|\big|q(x)\big) + D\big(p(y|x)\big|\big|q(y|x)\big) \end{aligned}$$

2.6 Jensen's Inequality and its Consequences

First, we begin by defining a convex function.

Definition 8 (Convex Function) *A convex function, $f(x)$, over an interval (a, b) is convex if for every $x_1, x_2 \in (a, b)$ and for $0 \leq \lambda \leq 1$*

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

Strictly convex refers to functions who only reach equality when $\lambda = 0, 1$. Intuitively, this means that any convex function lies below its chord.

Definition 9 (Concave Function) *A function f is concave if $-f$ is convex.*

Theorem 2.7 *If a function has second derivative that is non-negative over an interval, then the function is convex over that interval*

Proof: Consider a taylor expansion of a function around the point x:

$$f(x) = f(x_0) + f'(x)(x - x_0) + \frac{1}{2}f''(x)(x - x_0)^2$$

By hypothesis we assume $f''(x) \geq 0$. Therefore, the final term in the taylor expansion is always positive. If we let $x_0 = \lambda x_1 + (1 - \lambda)x_2$ and $x = x_1$

$$f(x_0) \geq f(x_0) + f'(x_0)(x - x_0) \tag{1}$$
$$f(x_0) \geq f(x_0) + f'(x)((1 - \lambda)(x_1 - x_2)) \tag{2}$$
$$\tag{3}$$

We can do the same and let $x = x_2$

$$f(x_0) \geq f(x_0) + f'(x)(\lambda(x_2 - x_1))$$

Now, we multiply the first equation by λ and the second by $1 - \lambda$ and sum the equations to yield the inequality in question.

Theorem 2.8 (Jensen's Inequality) *If f is a convex function and X is a random variable,*

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

If f is strictly convex then the equality implies that $X = \mathbb{E}[X]$.

Proof: The proof involves induction of point masses. As a base case, consider two point masses x_1 and x_2 with associated probabilities p_1 and p_2. Jensen's inequality in this form takes the shape

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$$

This statement follows directly from the definition of convexity since $p_1 + p_2 = 1$. Now, for the inductive step suppose that Jensen's inequality is true for the $k - 1$ point mass case. We introduce the quanitities $p'_i = p_i/(1 - p_k)$ for each $i = 1, ..., k - 1$ so that we can write the following expression:

$$\mathbb{E}[f(X)] = \sum_i^k p_i(f(x_i)) = p_k f(x_k) + (1 - p_k) \sum_i^{k-1} p'_i f(x_i) \tag{4}$$
$$\geq p_k f(x_k) + (1 - p_k) f\left(\sum_i^{k-1} p_i x_i\right) \tag{5}$$
$$\geq f\left(p_k f(x_k) + \sum_i^{k-1} p_i x_i\right) \tag{6}$$
$$= f(\mathbb{E}[X]) \tag{7}$$

Line 5 arises from the inductive hypothesis and line 6 is an application of the base case argument or definition of convexity. The proof can be extended to continuous distributions via continuity arguments.

Theorem 2.9 (Information Inequality) *Let $p(x)$ and $q(x)$ be two probability distributions over the same state space $\mathcal{X}$. Then,*

$$D\big(p(y|x)\big|\big|q(y|x)\big) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x.

Proof: Let A represent the support of $p(x)$.

$$\begin{aligned}
D\big(p(y|x)\big|\big|q(y|x)\big) &= -\sum_{x\in A} p(x)\log\frac{p(x)}{q(x)} \\
&= \sum_{x\in A} p(x)\log\frac{q(x)}{p(x)} \\
&\leq \log\sum_{x\in A} p(x)\frac{1(x)}{p(x)} \\
&= \log\sum_{x\in A} q(x) \\
&\leq \log\sum_{x\in \mathcal{X}} q(x) \\
&= \log 1 \\
&= 0
\end{aligned}$$

Corollary 2.9.1 (Nonnegativity of Mutual Information) *For any two random variables X and Y,*

$$I(X,Y) \geq 0$$

with equality if and only if X and Y are independent.

Proof: $I(X,Y) = D\big(p(x,y)\big|\big|p(x)p(y)\big) \geq 0$ with equality if and only if $p(x,y) = p(x)p(y)$.

Corollary 2.9.2

$$D\big(p(y|x)\big|\big|q(y|x)\big) \geq 0$$

Corollary 2.9.3

$$I(X,Y|Z) \geq 0$$

with equality if and only if X and Y are conditionally independent given Z

Theorem 2.10 (Entropy Upper Bound) *$H(X) \leq |\mathcal{X}|$ with equality if and only if X has a uniform distribution over $\mathcal{X}$*

Proof: Let $u(X) = \frac{1}{|\mathcal{X}|}$ be the uniform distribution over the state space and $p(x)$ be an arbitrary probability distribution.

$$D\big(p\big|\big|u\big) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{u(x)} = \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{u(x)} + \sum_{x\in\mathcal{X}} p(x)\log p(x) = \log|\mathcal{X}| - H(X) \geq 0$$

Therefore,

$$H(X) \leq \log|\mathcal{X}|$$

Theorem 2.11 (Conditioning reduces Entropy)

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

with equality if and only if X and Y are independent

Proof: $0 \leq I(X,Y) = H(X) - H(X|Y)$

Theorem 2.12 (Independence bound on Entropy) *Consider $X_1, X_2, ...X_n$ drawn from $p(x_1, x_2, ...x_n)$*

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

with equality if and only if the X_i are independent

Proof: Direct application of chain rule of entropies:

$$H(X_1, X_2, ...X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, ...X_1) \tag{8}$$

$$\leq \sum_{i=1}^{n} H(X_i) \tag{9}$$

2.7 The log sum inequality and its applications

Theorem 2.13 (Log sum inequality) *For non-negative numbers $a_1, ..., a_n$ and $b_1, ..., b_n$*

$$\sum_{i}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i}^{n} a_i\right) \log \frac{\sum_i^n a_i}{\sum_i^n b_i}$$

with equality if and only if $\frac{a_i}{b_i}$ is constant

Proof: First, notice that $f(t) = t \log t$ is strictly convex. We can therefore apply Jensen's inequality to show

$$\sum \alpha_i f(t_i) \geq f\left(\sum \alpha_i t_i\right)$$

when $\sum_i \alpha_i = 1$. Now, let's define $\alpha = \frac{b_i}{\sum_i b_i}$ and $t_i = \frac{a_i}{b_i}$

$$\sum \frac{a_i}{\sum_i b_i} \log(\frac{a_i}{b_i}) \geq \sum \frac{a_i}{\sum_i b_i} \log \left(\sum \frac{a_i}{\sum_i b_i}\right)$$

Theorem 2.14 (Convexity of Relative Entropy) *$D\big(p\big|\big|q\big)$ is convex in the pair (p, q)*

$$D\big(\lambda p_1 + (1-\lambda)p_2\big|\big|\lambda q_1 + (1-\lambda)q_2\big) \leq \lambda D\big(p_1\big|\big|q_1\big) + (1-\lambda) D\big(p_2\big|\big|q_2\big)$$

Proof: Direct application of the log sum inequality

$$(\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)}$$

$$\leq \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x) \log \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)}$$

REVIEW

Theorem 2.15 (Convexity of Entropy) *H(p) is concave in p*

Proof:

$$H(p) = \log |\mathcal{X}| - D\big(p\big|\big|u\big)$$

Concavity of H follows from the convexity of D

Theorem 2.16 *Let $(X, Y) \sim p(x, y) = p(y|x)p(x)$. The mutual information $I(X, Y)$ is concave of $p(x)$ for fixed $p(y|x)$ and convex in $p(y|x)$ for fixed $p(x)$.*

Proof:

$$I(X,Y) = H(Y) - H(Y|X) = H(Y) - \sum p(x)H(Y|X=x)$$

For fixed $p(y|x)$ $p(y)$ is a linear function of $p(x)$. Since $H(Y)$ is concave for $p(y)$ it follows that it must also be concave for $p(x)$ for fixed $p(y|x)$. The second term in mutual information is also linear in $p(x)$ so $I(X,Y)$ is concave for fixed $p(y|x)$.

Let's consider two conditional distributions $p_1(y|x)$ and $p_2(y|x)$. Next, consider the conditional distribution $p_\lambda(y|x) = \lambda p_1(y|x) + (1-\lambda)p_2(y|x)$. We can similarly list the joint distribution $p_\lambda(x,y) = \lambda p_1(x,y) + (1-\lambda)p_2(x,y)$ and the mixed distribution of Y $p_\lambda(y) = \lambda p_1(y) + (1-\lambda)p_2(y)$. If we define $q(x,y) = p(x)p_\lambda(y)$ we can write our mutual information as a relative entropy.

$$I(X,Y) = D\big(p_\lambda(x,y)\big|\big|q_\lambda(x,y)\big)$$

This is convex as shown above.

2.8 The Data Processing Inequality

Intuitively, the data processing inequality states that no manipulations of the data can improve inferences from the data.

Definition 10 *Random variables, X, Y, Z form a markov chain in the order $X \to Y \to Z$ if the distribution of Z only depends on Y and the distribution of Y only depends on X.*

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

X and Z are conditionally independent given Z :

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(x,z)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

Note that $Z = f(Y)$ forms a valid Markov chain.

Theorem 2.17 (Data Processing Inequality) *If $X \to Y \to Z$, then $I(X,Y) \geq I(X,Z)$*

Proof: We begin by expanding mutual information via the chain rule:

$$I(X,Y,Z) = I(X,Z) + I(X,Y|Z) = I(X,Y) + I(X,Z|Y)$$

From our definitions we know that $X \perp Z|Y$ so $I(X,Z|Y) = 0$. Since $I(X,Y|Z) \geq 0$ it follows that $I(X,Y) \geq I(X,Z)$.

Corollary 2.17.1 *If $X \to Y \to Z$, then $I(X,Y|Z) \leq I(X,Y)$*

Proof: From above we have that

$$I(X,Z) + I(X,Y|Z) = I(X,Y)$$

Since $I(X,Z) \geq 0$ it is clear that $I(X,Y|Z) \leq I(X,Y)$.

2.9 Sufficient Statistics

Definition 11 *A function $T(X)$ is said to be a sufficient statistic relative to a family of probability distributions $\{f_\theta(x)\}$ if X is independent of θ given $T(X)$ for any distribution in the family. Alternatively, $\theta \to X \to T(X)$ forms a markov chain.*

$$I(\theta, X) = I(\theta, T(X))$$

ADD EXAMPLES

Definition 12 *A statistic $T(X)$ is minimal sufficient statistic relative to a family of distributions $\{f_\theta(x)\}$ if it is a function of every other statistic U*

A minimal sufficient statistic maximally compresses information about θ in the sample.

2.10 Fano's Inequality

Fano's inequality relates the error probability in guessing a random variable X to its conditional entropy $H(Y|X)$ for some correlated variable Y.

Theorem 2.18 (Fano's Inequality) *For any estimator $\hat{X}$ that forms the markov chain $X \to Y \to \hat{X}$, with probability $P_e = Pr\{X \neq \hat{X}\}$*

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

Or, more weakly,

$$P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}$$

Proof: First, let's define an indicator random variable for error, E. We can right the join conditional entropy of this indicator as follows:

$$H(E, X|\hat{X}) = H(X|\hat{X}) + H(E|X, \hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X})$$

Note that $H(E|X, \hat{X}) = 0$. We also claim that $H(E|\hat{X}) \leq H(P_e)$ since conditioning reduces entropy and $H(P_e) = H(E)$. Finally,

$$H(X|E, \hat{X}) = Pr\{E = 0\}H(X|\hat{X}, E = 0) + Pr\{E = 1\}H(X|\hat{X}, E = 1) \leq (1 - P_e) * 0 + P_e \log |\mathcal{X}|$$

Here we 0 out the first term and upperbound the second term by the log of the state space. Combining these simplifications, we produce the first inequality shown.

We can then recover the inequality $H(X|\hat{X}) \geq H(X|Y)$ through direct application of the data processing inequality.

If we let $\hat{X} = Y$ we see that for any two random variables

$$H(P) + P \log |\mathcal{X}| \geq H(X|Y)$$

Theorem 2.19 *If X and X' are iid with entropy $H(X)$*

$$Pr\{X = X'\} \geq 2^{-H(X)}$$

with equality if and only if X has a uniform distribution

Proof: If $X \sim p(x)$ we can apply Jensen's inequality as follows:

$$2^{\mathbb{E}[\log p(X)]} \leq \mathbb{E}[2^{\log p(X)}]$$

Expanding each side,

$$2^{-H(X)} = 2^{\sum p(x) \log p(x)} \leq \sum p(x) 2^{\log p(x)} = \sum p^2(x) = Pr\{X = X'\}$$

Corollary 2.19.1 *If X and X' are independent with distributions $p(X)$ and $r(X)$ respectively,*

$$Pr\{X = X'\} \geq 2^{-H(p) - D\left(p\middle|\middle|r\right)} \qquad Pr\{X = X'\} \geq 2^{-H(r) - D\left(r\middle|\middle|p\right)}$$

Proof:

$$\begin{aligned}
2^{-H(p) - D\left(p\middle|\middle|r\right)} &= 2^{\sum p(x) \log p(x) + \sum p(x) \log \frac{r(x)}{p(x)}} \\
&= 2^{\sum p(x) \log r(x)} \\
&\leq \sum p(x) 2^{\log r(x)} \\
&= \sum p(x) r(x) \\
&= Pr\{X = X'\}
\end{aligned}$$

3 Asymptotic Equipartition Property

The asymptotic equipartition property is the information theory analog to the law of large numbers. In fact, it follows directly from the weak law of large numbers. Recall that the law of large numbers states that for n draws of iid random variables, $\frac{1}{n}\sum X_i$ is the best estimate for $\mathbb{E}[X]$ in the limit $n \to \infty$. The AEP on similarly states $H \approx \frac{1}{n}\log\frac{1}{p(X_1,...,X_n)}$ where $X_1, ..., X_n$ are iid random variables with probability of being observed in that sequence $p(X_1, ..., X_n)$. It follows that $p(X_1, ..., X_n) \approx 2^{-nH}$.

With this definition we can define a typical set of random variables with entropy close to true entropy. Properties proved for this set then occur with high probability for large samples.

Definition 13 (Convergence of Random Variables) *A sequence of random variables $X_1, X_2...$ is said to converge to X*

1. *In probability: if for every $\epsilon > 0$ $Pr\{|X_n - X| > \epsilon\} \to 0$*
2. *In mean square: if $\mathbb{E}[(X_n - X)^2] \to 0$*
3. *With probability 1 (almost surely): if $Pr\{\lim_{n\to\infty} X_n = X\} = 1$*

3.1 Asymptotic Equipartition Property Theorem

Theorem 3.1 (Asymptotic Equipartition Property) *If $X_1, X_2...$ are iid $\sim p(x)$ then,*

$$\frac{1}{n}\log\frac{1}{p(X_1, X_2, ...X_n)} \to H(X) \quad \textit{in probability}$$

Proof: The key insight for this proof is to notice that functions of random variables are random variables themselves. Therefore, we can consider $\log p(X_i)$ to be an r.v.

$$\begin{aligned} -\frac{1}{n}\log\frac{1}{p(X_1, X_2, ...X_n)} &= -\frac{1}{n}\sum\log\frac{1}{p(X_i)} && (10)\\ &\approx \mathbb{E}[-\log p(X)] \qquad \text{Weak law of large numbers} && (11)\\ &= H(X) && (12)\end{aligned}$$

Definition 14 (Typical Set) *A typical set, $A_\epsilon^{(n)}$ with respsect to a probability distribution $p(x)$ is the set of sequences that obey*

$$2^{-n(H(X)+\epsilon)} \leq p(X_1, X_2, ...X_n) \leq 2^{-n(H(X)-\epsilon)}$$

Theorem 3.2 *The following properties arise from the definition of a typical set*

1. *If $(x_1, ..., x_n) \in A_\epsilon^{(n)}$, then $H(X) - \epsilon \leq -\frac{1}{n}\log\frac{1}{p(X_1,X_2,...X_n)} \leq H(X) + \epsilon$*
2. *$Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$ for n sufficiently large*
3. *$|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$ where $|A|$ is the number of elements in the set*
4. *$|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$ for n sufficiently large*

Conceptually, we understand these statements as the probability of the typical set being nearly 1, all elements in the typical set are equally probable, and the number of items in the typical set is approximately $2^{nH(X)}$

Proof:

1) is immediate from the definition of a typical set.

2) We approach this proof my envoking the asymptotic equipartition theorem. First, notice that using 1) we can write the condition of existing in $A_\epsilon^{(n)}$ as $|-\frac{1}{n}\log p(x_1, ..., x_n) - H(x)| < \epsilon$ Next, by the definition of convergence in probability we know that for every $\delta > 0$, there exists n_0 such that for $n \geq n_0$

$$Pr\{|-\frac{1}{n}\log p(x_1,...,x_n) - H(x)| < \epsilon\} < 1-\delta$$

If we set $\delta = \epsilon$ we recover the second property.

3)

$$\begin{aligned}
1 &= \sum_{x\ in\mathcal{X}} p(x) \\
&\geq \sum_{x\in A_\epsilon^{(n)}} p(x) \\
&\geq \sum_{x\in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \qquad \text{Definition of Typical Set} \\
&= 2^{-n(H(X)+\epsilon)}|A_\epsilon^{(n)}|
\end{aligned}$$

From here we can trivially see that $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$.

4) Beginning with 2, we have

$$\begin{aligned}
1-\epsilon &< Pr\{A_\epsilon^{(n)}\} \\
&\leq \sum_{x\in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \qquad \text{Definition of Typical Set} \\
&= 2^{-n(H(X)+\epsilon)}|A_\epsilon^{(n)}|
\end{aligned}$$

From here it follows $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$.

3.2 Consequences of the AEP: Data Compression

Theorem 3.3 *Let $X^n \sim p(x)$ by iid. Let $\epsilon > 0$. Then there exists a code that maps sequences x^n of length n into binary strings such that the mapping is 1-to-1 and*

$$\mathbb{E}[\frac{1}{n}l(X^n)] \leq H(X) + \epsilon$$

for sufficiently large n

This implies that we can represents sequeneces X^n using $nH(X)$ bits on average.

Proof: First, we order the sequence $A_\epsilon^{(n)}$ by giving each entry an index. Since here are $\leq 2^{n(H+\epsilon)}$ sequences in this set, we can represent them by $\leq n(H+\epsilon)+1$ bits. Note that the additional bit accounts for non-integer values of $H+\epsilon$. To specify that these sequences are in the typical set we introduce an additional pre-fix bit for a total maximum of $\leq n(H+\epsilon)+1$ bits. The same reasoning can be applied to sequences no in the typical sets with for a maximumal bit representation of $n\log|\mathcal{X}|+2$.

Next, consider the notation $l(x^n)$ as the length of the codeword of the sequence corresponding to x^n. If n is sufficiently large, we can apply the condition $Pr\{A_\epsilon^{(n)}\} > 1-\epsilon$. Our expected code length is then:

$$\mathbb{E}[l(x^n)] = \sum_{x^n \in \mathcal{X}} p(x^n) l(x^n) \tag{13}$$

$$= \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) l(x^n) + \sum_{x^n \in A_\epsilon^{(n)c}} p(x^n) l(x^n) \tag{14}$$

$$\leq \sum_{x^n \in A_\epsilon^{(n)}} p(x^n)(n(H+\epsilon)+2) + \sum_{x^n \in A_\epsilon^{(n)c}} p(x^n)(n \log |\mathcal{X}| + 2) \tag{15}$$

$$= Pr\{A_\epsilon^{(n)}\}(n(H+\epsilon)+2) + Pr\{A_\epsilon^{(n)^C}\}(n \log |\mathcal{X}| + 2) \tag{16}$$

$$\leq n(H+\epsilon) + \epsilon n \log |\mathcal{X}| + 2 \tag{17}$$

$$= n(H+\epsilon') \tag{18}$$

3.3 High-Probability Sets and the Typical Set

Theorem 3.4 *Let $X_1, ...X_n$ be iid $\sim p(x)$. For $\delta < 1/2$ and any $\delta' > 0$, if $PR\{B_\delta^n\} > 1 - \delta$, then*

$$\frac{1}{n} \log |B_\delta^{(n)}| > H - \delta'$$

for n sufficiently large.

$B_\delta^{(n)}$ must have at least 2^{nH} elements which is approximately the same as $A_\epsilon^{(n)}$.

4 Entropy Rates of a Stochastic Process

4.1 Markov Chains

A general stochastic process is a series of indexed random variables.

Definition 15 (Stationary) *A stochastic process is said to be stationary if the joint distribution of any subset of the varibales is invariant to shifts in the index.*

$$Pr\{X_1, X_2, ...X_n\} = Pr\{X_{1+l}, X_{2+l}, ...X_{n+l}\}$$

for any n and any shift l.

A markov process is a special subset of stochastic processes where the current random variable is conditionally independent of all previous random variables given the most recent random variable. In other words, a markov process is a series of random variables that are only dependent on the previous random variable. Realizations of the Markov random variable are referred to as states.

Definition 16 (Markov Chain) *A discrete stochastic process is said to be Markov if for $n = 1, 2, ...$*

$$Pr\{X_{n+1}|X_n, ...X_1\} = Pr\{X_{n+1}|X_n\}$$

This definition implies that we can break down the joint probability distribution as follows:

$$p(x_1, x_2, ...x_n) = p(x_1)p(x_2|x_1)...p(x_{n+1}|x_n)$$

Next, we introduce the concept of time-invariance to denote markov chains that are independent of n and only depend on the state itself. Alternatively, time-invariance can be understood as the conditional probability of each realization of the random variable staying constant over time.

Definition 17 (Time-Invariance) *A Markov chain is said to be time-invariant if the conditional probability $p(x_{n+1}|x_n)$ is independent of n.*

$$Pr\{X_{n+1} = a|X_n = b\} = Pr\{X_2 = a|X_1 = b\}$$

A time-invariance Markov chain can then be fully captured by an initial state x_0 and a probability transitition matrix $P_{i,j}Pr\{X_n = i|X_n = j\}$. Markov chains can also be classified as irreducible if it is possible to jump from any state i to any other state j in a finite number of steps. If the greatest common divisor of every path from a state to itself is 1, then the state is said to be aperiodic; otherwise, it is simply periodic. A stationary distribution is one that has the same distribution of states at time n and $n+1$. We can calculate the evolution of the distribution of states using the transition matrix. In fact, matrix multiplication allows us to write $p(x_{n_1}) = \sum_{x_n} p(x_n)P(x_n, x_{n+1})$ as

$$\pi_{n+1} = \pi_n P$$

If π was stationary we would have $\pi_{n+1} = \pi_n$. An important theorem for Markov chains is that if a finite-state Markov chain is aperiodic and irreducible and has a unique stationary distribution, then the distrubtion X_n approaches the stationary distribution as $n \to \infty$. For a more formal discussion of Markov chains and their properties refer to a text on stochastic processes.

4.2 Entropy Rate

If a markov chain is in its stationary distribution, the entropy of the random variable is simply given by the entropy of that distribution. However, if are not in the stationary distribution, what can we say about the entropy of our Markov chain? In particular, we ask the question, how does the entropy grow with n?

Definition 18 (Entropy of a Stochastic Process) *The entropy of a stochastic process is given by*

$$\mathcal{X} = \lim_{n\to\infty} \frac{1}{n} H(X_1, ..., X_n)$$

We can apply this definition to a couple of common stochastic processes:

- **Uniform Distribution**: Consider the entropy limit of a stochastic process where each random variable is drawn from a uniform distribution with probability m. for a sequence of n variables $H(X_1, ..., X_n) = \log m^n$. It follows that the entropy rate is:
$$\mathcal{H}(X_1, ..., X_n) = \log m$$
- **i.i.d Random Variables**:
$$\mathcal{H}(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, ..., X_n) = \lim_{n\to\infty} \frac{nH(X_1)}{n} = H(X_1)$$
- **Independent Random Variables**:
$$\mathcal{H}(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, ..., X_n) = \lim_{n\to\infty} \frac{\sum_i H(X_i)}{n}$$
We can cleverly select random variables $X_1, ...X_n$ such that this limit does not exist.

We can alternatively define entropy rate via the quantity:

$$\mathcal{H}'(X) = \lim_{n\to\infty} H(X_n|X_{n-1}, ...X_1)$$

Intuitively,this entropy rate refers to the entropy rate of the last random variable given all the previous random variables.

Theorem 4.1 *For a stationary stochastic process, the limits of the entropy rate definitions exist and are equal.*

$$\mathcal{H}(X) = \mathcal{H}'(X)$$

Theorem 4.2 *For a stationary stochastic process, $H(X_n|X_{n-1}, ...X_1)$ is nonincreasing and hasthe limit $\mathcal{H}'(X)$.*

Proof:

$$H(X_{n+1}|X_n, ...X_1) \leq H(X_{n+1}|X_n..., X_2) = H(X_n|X_{n-1}, ...X_1)$$

The first inequality follows from the fact that conditioning does not increase entropy and the second equliaty is an application of stationarity. Since $H(X_n|X_{n-1}, ...X_1)$ is a decreasing seqeunce of nonnegative numbers the limit $\mathcal{H}'(X)$ exists.

Why decreasing and not nonincreasing?

Theorem 4.3 (Cesaro Mean) *If $a_n \to a$ and $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$ then $b_n \to a$*

Proof Sketch: Informally, we can say that since most terms in $\{a_i\}$ are eventually close to a, then b_n, the average of a_i is also close to a.

Now, we have the necessary background to prove Theorem 4.1.
Proof: We can apply the chain rule decomposition to $\mathcal{H}(X)$

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

We can now apply the Cesaro Mean theorem since our conditional entropy terms approach $\mathcal{H}'(X)$, we can claim that their average must as well. Therefore, $\frac{H(X_1,...,X_n)}{n} \to \mathcal{H}'(X)$ or equivalently $\mathcal{H}(X) = \mathcal{H}'(X)$.

Entropy rate is particularly useful as an average description length for a stationary ergodic process. However, the proof and validation of this interpretation comes later in the text.

Markov Chains

Consider a stationary Markov chain where the entropy rate is given by

$$\mathcal{H}(X) = \mathcal{H}'(X) = \lim H(X_n|X_{n-1}, ...X_1) = \lim H(X_n|X_{n-1}) = H(X_2|X_1)$$

The last two equalities arise from the Markov property and the time-invariance proerties of the Markov Chain. Note that the conditional entropy is calculated using the stationary distribution. Why?

Theorem 4.4 *Let $\{X_i\}$ be a stationary Markov chain with stationary distribution μ and transition matrix P. The entropy rate is*

$$\mathcal{H}(X) = -\sum_{i,j} \mu_i P_{i,j} \log P_{i,j}$$

This follows from the stationary distribution property $\mu_j = \sum_{i,j} \mu_i P_{i,j}$. Note that if a Markov chain is irreducible, aperiodic, and has a unique stationary distribution, then any initial distribution tends towards the stationary distribution. Since the Entropy Rate is a comment on long run behavior, it follows that the entropy rate of the chain is independent of the initial distribution. Conseequently, we can solve for the entropy rate for starting from the stationary distribution and apply that to all other initial distributions.

4.3 Example: Entropy Rate of Random Walk on a Weighted Graph

Consider a random walk on a weighted graph. Given that you are on node i at time t, at the next time step you will jump to one of the neighboring nodes with probability proportional to edge weights between the nodes. $P_{ij} = W_{i,j}/\sum_k W_{ik}$. We propose the stationary distribution

$$u_i = \frac{\sum_j W_{ij}}{\sum_{i,j} W_{ij}}$$

We can easily verify this via detailed balance:

$$u_i P_{ij} = u_j P_{ji} \tag{19}$$

$$\frac{\sum_j W_{ij}}{\sum_{i,j} W_{ij}} \cdot \frac{W_{i,j}}{\sum_k W_{ik}} = \frac{\sum_k W_{jk}}{\sum_{i,j} W_{ij}} \cdot \frac{W_{j,i}}{\sum_k W_{jk}} \tag{20}$$

$$W_{ij} = W_{ji} \tag{21}$$

Now we can calculate the entropy rate:

$$H(\mathcal{X}) = H(X-2|X_1) \tag{22}$$
$$= -\sum_i \mu_i \sum_j P_{i,j} \log P_{i,j} \tag{23}$$
$$= -\sum_i \frac{\sum_j W_{ij}}{\sum_{i,j} W_{ij}} \sum_j \frac{W_{i,j}}{\sum_k W_{ik}} \log \frac{W_{i,j}}{\sum_k W_{ik}} \tag{24}$$
$$= -\sum_i \sum_j \frac{W_{ij}}{\sum_{i,j} W_{ij}} \log \frac{W_{i,j}}{\sum_k W_{ik}} \tag{25}$$
$$= -\sum_i \sum_j \frac{W_{ij}}{\sum_{i,j} W_{ij}} \log \frac{W_{i,j}}{\sum_k W_{ik}} + \sum_i \sum_j \frac{W_{ij}}{\sum_{i,j} W_{ij}} \log \frac{\sum_k W_{ik}}{\sum_{i,j} W_{ij}} \tag{26}$$
$$= H(...\frac{W_{ij}}{\sum_{i,j} W_{ij}}...) - H(...\frac{\sum_k W_{ik}}{\sum_{i,j} W_{ij}}...) \tag{27}$$

4.4 Second Law of Thermodynamics

The Second Law of Thermodynamics states that the entropy of a isolated system is non-decreasing. In statistical mechanics, particularly under the microcanonical ensemble, the probability of the system being in a specified state is assumed to be uniformly distributed across all states with equivalent energy. If we model the system as a Markov chain over these states, we find a handful of interpretations of the second law.

1. **Relative Entropy Decreases with n**: Consider two probability distributions over the Markov chain state space μ and μ'. At time n and $n+1$ the distributions are given by μ_n, μ'_n and μ_{n+1}, μ'_{n+1} respectively. Let p and q denote the joint mass distributions for each function such that $p(x_n, x_n+1) = p(x_n)r(x_{n+1}|x_n)$ where r denotes the transition probability. Similarly, for q, $q(x_n, x_n+1) = q(x_n)r(x_{n+1}|x_n)$. With these definitions, we can consider the chain rule for relative entropy:

   $$D\big(p(x_n, x_{n+1})\big|\big|q(x_n, x_{n+1})\big) = D\big(p(x_n)\big|\big|q(x_n)\big) + D\big(p(x_{n+1}|x_n)\big|\big|q(x_{n+1}|x_n)\big)$$

   We can also decompose the joint distribution as

   $$D\big(p(x_n, x_{n+1})\big|\big|q(x_n, x_{n+1})\big) = D\big(p(x_{n+1})\big|\big|q(x_{n+1})\big) + D\big(p(x_n|x_{n+1})\big|\big|q(x_n|x_{n+1})\big)$$

   Now we argue that $p(x_{n+1}|x_n) = q(x_{n+1}|x_n)$ since both probability distributions are dependent on the transition proability function. Therefore, $D\big(p(x_{n+1}|x_n)\big|\big|q(x_{n+1}|x_n)\big) = 0$. Since relative entropy is non-negative, we know that $D\big(p(x_n|x_{n+1})\big|\big|q(x_n|x_{n+1})\big) \geq 0$. It follows that,

   $$D\big(p(x_n)\big|\big|q(x_n)\big) \geq D\big(p(x_{n+1})\big|\big|q(x_{n+1})\big)$$

   equivalently,

   $$D\big(\mu_n\big|\big|\mu'_n\big) \geq D\big(\mu_{n+1}\big|\big|\mu'_{n+1}\big)$$

   We can then state that the distance between probability mass functions decreases with time for any Markov Chains. What if the chain is transient? Does the same limit property hold – it feels counter intuitive but the math feels like it holds

2. **Relative Entropy between a distribution over states and the stationary distribution decreases with n**: This can be seen by setting μ' in the previous derivation to the stationary distribution. It folllows that $\pi_n = \pi_{n+1}$ and so the equality holds as well.

   $$D\big(\mu_n\big|\big|\pi\big) \geq D\big(\mu_{n+1}\big|\big|\pi\big)$$

   The state distribution gets closer to the stationary distribution as time passes. The limit of this sequence is 0 if the stationary distribution is unique. Doesn't this require specific Markov chain conditions?

3. **Entropy increases if the stationary distribution is uniform**: If the stationary distribution is uniform we can write relative entropy as

$$D(\mu_n||\pi) = \log|\mathcal{X}| - H(\mu_n) = \log|\mathcal{X}| - H(X_n)$$

The monotonic decrease in relative entropy implies a monotonic increase in entropy. This example is the case that is referred to in the microcanonical ensemble.

Definition 19 (Doubly Stochastic) *A probability transition matrix is said to be doubly stochastic if*

$$\sum_i P_{i,j} = 1 \qquad \sum_j P_{i,j} = 1$$

The uniform distribution is the stationary distrubtion of P if and only if the transition matrix is doubly stochastic.

4. **Conditional Entropy increases with n for a stationary Markov process**:

$$\begin{aligned} H(X_n|X_1) &\geq H(X_n|X_1, X_2) && \text{Conditioning} && (28)\\ &= H(X_n|X_2) && \text{Markovity} && (29)\\ &= H(X_{n-1}|X_1) && \text{stationarity} && (30)\end{aligned}$$

We can also show this through the data processing inequality:

$$I(X_1; X_{n-1}) \geq I(X_1; X_n)$$

Expanding each side,

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

By stationarity we know that $H(X_{n-1}) = H(X_n)$, so

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

5. **Shuffles Increase Entropy**: For a random shuffle T independent of the original state X:

$$H(TX) \geq H(X)$$

4.5 Functions of Markov Chains

Consider a stationary Markov Chain $X_1, ...X_n$ with $Y_i = \psi(X_i)$. Unfortunately, we can't claim that $\{Y_i\}$ is a markov chain, but we do know that it must be stationary. Therefore, we can meaningfully speak about the entropy rate $H(\mathcal{Y})$. However, since we don't know the rate at which $H(Y_n|Y_{n-1}, ...Y_1)$converges, we can't make conclusions about the limit. Why is this true?

Instead, we can try to place bounds on $H(\mathcal{Y})$. As shown in Theorem 4.2, we know that in a stationary stochastic process $H(Y_n|Y_{n-1}, ...Y_1)$ converges to $H(\mathcal{Y})$ from above. Therefore, we can focus on defining a lower bound by $H(Y_n|Y_{n-1}, ...Y_2, X_1)$.

Lemma 4.5 $H(Y_n|Y_{n-1}, ...Y_2, X_1) \leq H(\mathcal{Y})$

Proof: Consider $k = 1, 2, ...$

$$\begin{aligned} H(Y_n|Y_{n-1}, ...Y_2, X_1) &= H(Y_n|Y_{n-1}, ...Y_2, Y_1, X_1) && Y_1 \text{ is a function of } X_1 && (31)\\ &= H(Y_n|Y_{n-1}, ...Y_2, Y_1X_1, X_0, ...X_{-k}) && \text{Applying Markov Property} && (32)\\ &= H(Y_n|Y_{n-1}, ...Y_2, Y_1X_1, X_0, ...X_{-k}, Y_0, ...Y_{-k}) && Y_1 \text{ is a function of } X_1 && (33)\\ &\leq H(Y_n|Y_{n-1}, ...Y_2, Y_1, Y_0, ...Y_{-k}) && \text{Conditioning Reduces Entropy} && (34)\\ &= H(Y_{n+k+1}|Y_{n+k}, ...Y_2, Y_1) && \text{Stationarity} && (35)\end{aligned}$$

This inequality holds for all k so it holds in the limit

$$H(Y_n|Y_{n-1}, ...Y_2, X_1) \leq \lim_k H(Y_{n+k+1}|Y_{n+k}, ...Y_2, Y_1) = H(\mathcal{Y})$$

Lemma 4.6 $H(Y_n|Y_{n-1},...Y_1) - H(Y_n|Y_{n-1},...Y_2,X_1) \to 0$

Proof: Notice that we can write this interval as a mutual information term.

$$H(Y_n|Y_{n-1},...Y_1) - H(Y_n|Y_{n-1},...Y_2,X_1) = I(Y_n,X_1|Y_{n-1},...Y_1)$$

We can place a bound on mutual information

$$I(X_1;Y_n,Y_{n-1},...Y_1) \leq H(X_1)$$

such that

$$H(X_1) \geq \lim_{n\to\infty} I(X_1;Y_n,Y_{n-1},...Y_1) \tag{36}$$

$$= \lim_{n\to\infty} \sum_{i=1}^{n} I(Y_n,X_1|Y_{n-1},...Y_1) \tag{37}$$

$$= \sum_{i=1}^{\infty} I(Y_n,X_1|Y_{n-1},...Y_1) \tag{38}$$

Since the sum is infinite, bounded, and each term is non-negative we can conclude the terms must approach 0.

Theorem 4.7 *If $X_1,X_2,...X_n$ form a stationary markov chain and $Y_i = \psi(X_i)$, then*

$$H(Y_n|Y_{n-1},...Y_1,X_0) \leq H(\mathcal{Y}) \leq H(Y_n|Y_{n-1},...Y_1,)$$

and

$$\lim H(Y_n|Y_{n-1},...Y_1,X_0) = H(\mathcal{Y}) = \lim H(Y_n|Y_{n-1},...Y_1,)$$

We can apply even tighter bounds if Y_i is another stochastic process (i.e. Hidden Markov Model) Work these out Do we have any constraints on ϕ?

5 Data Compression

Data compression involves assigning small code words to frequent occurences in the data. By doing so, the goal is to optimally reduce the length required to transmit/store given information. Morse code is one of such examples where the most frequent symbol is represented by a dot.

5.1 Examples of Codes

Definition 20 (Source Code) *A source code C for a random variable X is a mapping from the state space $\mathcal{X}$ to $\mathcal{D}^*$ the set of finite length strings of symbols from a D-ary alphabet.*

For a given x, $C(x)$ is its associated code word and $l(x)$ is its length. Note that we can also assume a D-ary alphabet is $\mathcal{D} = \{0,...,D-1\}$

Definition 21 (Expected Length) *The expected length $L(C)$ of a source code C for a random variable with probability distribution $p(x)$*

$$L(C) = \sum_{x\in\mathcal{X}} p(x)l(x)$$

Definition 22 (Nonsingular) *A code is nonsingular if every element of $\mathcal{X}$ maps to a unique value in $\mathcal{D}^*$*

Definition 23 (Extension) *The extension C^* of a code C is a mapping from finite strings of $\mathcal{X}$ to strings of $\mathcal{D}$*

$$C(x_1,...x_n) = C(x_1)C(x_2)...C(x_n) \qquad \textit{Concatenation of code words}$$

Definition 24 (Uniquely Decodable) *A uniquely decodable code is one whose extension is non-singular*

Simply, a uniquely decodable string is one that only has one possible source string. Note that this does not imply that you can interpret the code by reading it in order.

Definition 25 (Prefix Code) *A prefix or instantaneous code is one in which no code is a prefix of another.*

Instantaneous codes can be interpretted immediately since the end of a codeword is recognizable. In this way, instantaneous codes are self-punctuating. These definitions provide a natural hierchical order of codes:

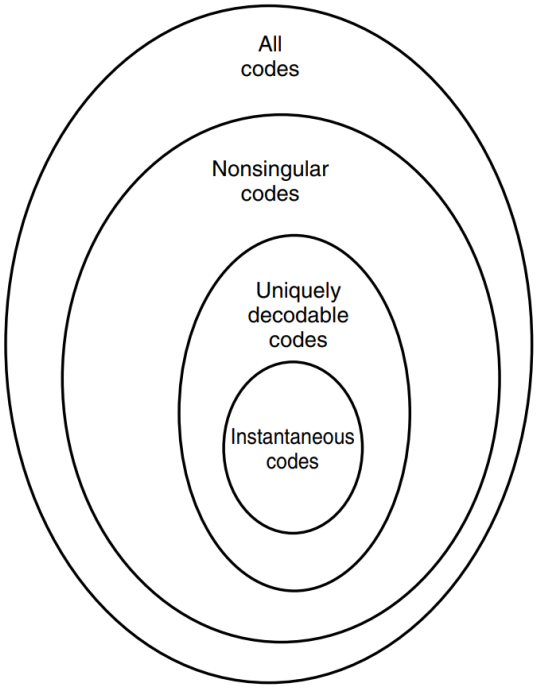


FIGURE 5.1. Classes of codes.

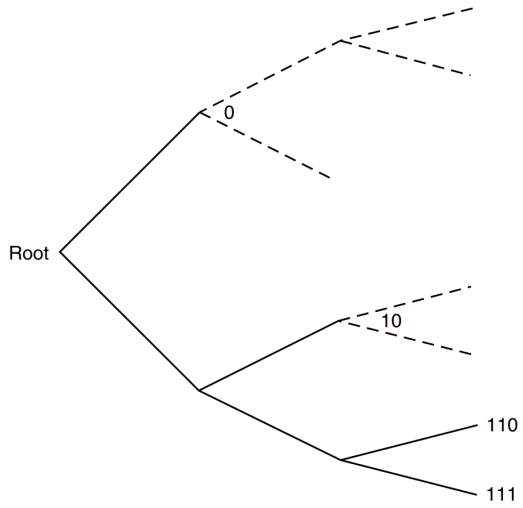
5.2 Kraft Inequality

Theorem 5.1 (Kraft Inequality) *For any instantaneous code over an alphabet of size D, the codeword lengths must satisfy*

$$\sum_i D^{-l_i} \leq 1$$

Conversely, if the lengths of code words satisfy the Kraft inequality, then there exists an instantaneous code for those code lengths.

Proof: We can conceptualize the prefix free requirement by considering a tree with each node having D children. Each branch corresponds to a symbol in the codeword with the final codewords being represented by leafs on the tree. For example, one could read off the path from a leaf to the root to reconstruct the codeword.



Since no leaf can be an ancestor for another, we ensure that no codeword can be a prefix for another. Now, consider l_{max} the maximum length coded. If we consider a tree depth of l_{max} we will some leaves of length l_{max} and some empty leaves that do not correspond to codewords. The important thing to note is that no codewords exist at a greater depth or that no descendant of codewords at this level exist in the set. For an arbitrary codeword length l_i, we will have at most $D^{l_{max}-l_i}$ descendants in the l_{max} layer. It follows that each of these descendants are disjoint and so we can sum over all codewords and still be bounded by the total number of possible leaves in the max layer:

$$\sum D^{l_{max}-l_i} \leq D^{l_{max}}$$

The Kraft Inequality is equivalently

$$\sum D^{-l_i} \leq 1$$

We can argue the converse by constructing a prefix code using the same tree system.

Theorem 5.2 (Extended Kraft Theorem) *For a countably infinite set of codewords that form a prefix code, the codewords satisfy*

$$\sum_{i}^{\infty} D^{l_i} \leq 1$$

THe converse also holds true that we can construct a prefix code with code lengths l_1 ,... given that they satisfy the extended Kraft inequality.

Skipped Proof

5.3 Optimal Codes

The Kraft inequality provides us a test to see if a codeword set can satisfy the prefix condition, but does not speak on the shortest possible prefix code. The question becomes, how can we minimize the average code length $L = \sum p_i l_i$ over all integers satisfying $\sum D^{-l_i} \leq 1$ Let's solve this as a simple minimization problem. First, we simplify the problem by removing the integer constraint and setting the contraint to an equality statement. Now, we can write this problem as a Lagrange Multiplier minimization problem.

$$J = \sum p_i l_i + \lambda \left(\sum D^{-l_i} \right)$$

Differentiating w.r.t l_i

$$\frac{\partial J}{\partial l_i} = p_i - \lambda D^{-l_i} \log_e D$$

Now, we set this derivative to 0 and solve for D^{-l_i}

$$D^{-l_i} = \frac{p_i}{\lambda \log_e D}$$

We can plug this expression into our equality constraint and find $\lambda = \frac{1}{\log_e D}$ Therefore,

$$D^{-l_i} = p_i$$

So the optimal code lengths are given by $l_i^* = \log_D p_i$. The average optimal length is then

$$L^* = \sum p_i l_i^* = -\sum p_i \log_D p_i = H_D(X)$$

Note that we assumed that l_i can realize non-integer values. In practice we select integer code lengths as close as possible to optimal conditions.

Theorem 5.3 *The expected length of any instantaneous D-ary code for a random variable X is lower bounded by the entropy* $H_D(X)$

$$L \geq H_D(X)$$

with equality if and only if $D^{-l_i} = p_i$

Proof:

$$\begin{aligned} L - H_D(X) &= \sum p_i l_i - \sum p_i \log_D \frac{1}{p_i} \\ &= \sum p_i \log_D D^{-l_i} + \sum p_i \log_D p_i \end{aligned}$$

If we define $r_i = \frac{D^{-l_i}}{\sum_j D^{-l_j}}$ and $c = \sum D^{-l_i}$

$$\begin{aligned} L - H_D(X) &= \sum p_i \log \frac{p_i}{r_i} - \log_D c \\ &= D(p||r) + \log_D \frac{1}{c} \\ &\geq 0 \end{aligned}$$

The Kraft inequality ensures that $c \leq 1$ so the last summation is necessarily non-negative by the non-negativity of relative entropy.

Definition 26 (D-adic) *A probability distribution is said to be D-iadic if every probability is equal to* D^-n *for some n.*

With this definition in hand, we can claim that equality is reached if and only if the probability distribution is D-adic.

5.4 Bounds on the Optimal Code Length

We construct a near optimal code by using codeword lenghts rounded up to the nearest integer of the optimal solution.

$$l_i = \left\lceil \log_D \frac{1}{p_i} \right\rceil$$

We see that this code must satisfy the Kraft Inequality

$$\sum D^{-l_i} = \sum D^{-\lceil \log_D \frac{1}{p_i} \rceil} \leq \sum D^{\log_D \frac{1}{p_i}} = 1$$

since each $l_i \geq l_i^*$. Next, we write the bounds and multiply by p_i and sum over all i to find

$$\log_D \frac{1}{p_i} \leq l_i \leq \log_D \frac{1}{p_i} + 1$$
$$H_D(X) \leq L < H_D(X) + 1$$

Optimal codes must be better than this code.

Theorem 5.4 *Let L_i^* denote the optimal code lengths for some probability distribution p over a D-ary alphabet. Let L^* be the expected code length for the optimal encoding*

$$H_D(X) \leq L^* < H_D(X) + 1$$

Proof: We already know that $L^* \geq H_D(X)$. Now, using the encoding suggested above, we argue that $L^* \leq L < H_D(X) + 1$.

One way to go about reducing the additional overhead bit is to spread it across the many symbols. Consider a system where we send information in large chunks of $x_1, ...x_n$. Let L_n denote the expected codeword length per symbol.

$$L_n = \frac{1}{n} \sum p(x_1, ...x_n) l(x_1, ...x_n) = \frac{1}{n} \mathbb{E}[l(x_1, ...x_n)]$$

For an optimal code, we can apply the entropy bounds above:

$$H(X_1, ..., X_n) \leq \mathbb{E}[l(X_1, ...X_n)] \leq H(X_1, ..., X_n) + 1$$

If we assume that the draws X_i are iid, we can simplify the expression

$$H(X_1, ..., X_n) \leq \mathbb{E}[l(X_1, ...X_n)] \leq H(X_1, ..., X_n) + 1$$
$$nH(X) \leq \mathbb{E}[l(X_1, ...X_n)] \leq nH(X) + 1$$
$$H(X) \leq L_n \leq H(X) + \frac{1}{n}$$

Therefore, if we select large block lengths we can achieve expected code length arbitrarily close to entropy.

Alternatively, we can also consider a stochastic process where the X_i are not iid.

$$\frac{H(X_1, ..., X_n)}{n} \leq L_n \leq \frac{H(X_1, ..., X_n)}{n} + \frac{1}{n}$$

If the process is stationary we see that in the limit the expected code length approaches the entropy rate. Entropy rate can therefore be interpretted as the expected number of symbols to describe a stationary stochastic process.

Theorem 5.5 (Wrong Code) *The expected length under $p(x)$ of the code assignment $l(x) = \left\lceil \log \frac{1}{q(x)} \right\rceil$ (Shannon Code Assignment) satisfies*

$$H(p) + D(p||q) \leq \mathbb{E}_p l(X) \leq H(p) + D(p||q) + 1$$

Believing the wrong distribution incurs a penalty of $D(p||q)$ for the average description length. Relative entropy is the increase in descriptive complexity due to incorrect information.

Proof:

$$\begin{aligned}
\mathbb{E}[l(x)] &= \sum p(x)\Big[\log x \frac{1}{q(x)}\Big] \\
&< \sum p(x)\left(\log \frac{1}{q(x)} + 1\right) \\
&= \sum_x p(x) \log \frac{p(x)}{q(x)} \frac{1}{q(x)} + 1 \\
&= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \log \frac{1}{p(x)} + 1 \\
&= D(p||q) + H(p) + 1
\end{aligned}$$

The derivation for the lower bound is analagous.

5.5 Kraft Inequality For Uniquely Decodable Codes

Theorem 5.6 (McMillan) *The codeword lengths of any uniquely decodable D-ary code must satisfy the Kraft inequality*

$$\sum D^{-l_i} \leq 1$$

Conversely, any codeword set that obey this inequality can be used to create a uniquely decodable code.

Proof: For the extension code, we can write the length of the extension as $l(x_1, ..., x_k) = \sum_{i=1}^{k} l(x_i)$. Consider the quantity

$$\left(\sum_{x\in\mathcal{X}} D^{-l(x)}\right)^k = \sum_{x_1} ... \sum_{x_k} D^{-l(x_1)} D^{-l(x_2)} ... D^{-l(x_k)} = \sum_{x_1,...,x_k} D^{-l(x_1,...x_k)}$$

Let $a(m)$ denote the count of count with length m. We can write this summation as

$$\sum_{x_1,...,x_k} D^{-l(x_1,...x_k)} = \sum_{m=1}^{kl_{max}} a(m) D^{-m}$$

We can bound $a(m) \leq D^m$ since each code of length m has D^m possiblities. If the code is uniquely decodable we know that we cannot have multiple strings map to the same code. It follows that we have at most one string map to each of the at most D^m possibilites.

$$\begin{aligned}
\left(\sum_{x\in\mathcal{X}} D^{-l(x)}\right)^k &= \sum_{m=1}^{kl_{max}} a(m) D^{-m} \\
&\leq \sum_{m=1}^{kl_{max}} D^m D^{-m} \\
&= kl_{max}
\end{aligned}$$

Then, we get $\sum_{x\in\mathcal{X}} D^{-l(x)} \leq (kl_{max})^{1/k}$. Since this holds for all k, in the limit $k \to \infty$ we recover

$$\sum_{x\in\mathcal{X}} D^{-l(x)} \leq 1$$

Corollary 5.6.1 *A uniquely decodable code for an infinite source alphabet X also satisfies the Kraft inequality*

We can augment the above proof by arguing that any subset of an uniquely decodable code is also uniquely decodable. Since any finite subset satisfies the Kraft inequality, we can argue

$$\sum_{i}^{\infty} D^{-l_i} = \lim_{N\to\infty} \sum_{i}^{N} D^{-l_i} \leq 1$$

This theorem implies that uniquely decodable codes do not offer any further set choices than prefix codes. The set of achievable code lengths is the same for both types. Therefore, the bounds calcualted above still hold for uniquely decodable codes.

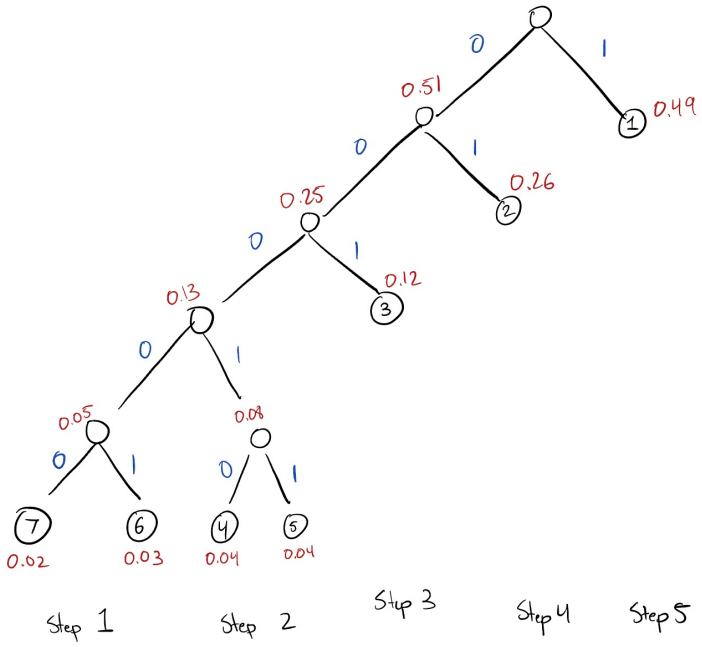
5.6 Huffman Codes

Huffman codes are prefix codes that provide the shortest expected length for a given distribution. In a D-ary alphabet, the general strategy to construct a Huffman code is to denote the D lowest probability events by the longest string and a differing final symbol. The process to construct a Huffman code involves sequentially combining the lowest D probaiblity events into singular events. Once we have a singular event left, we denote this with the shortest codeword and work our way down the constructed tree.

Example
Consider the following probability distribution.

Length	Code	X	Probability
1	1	X_1	0.49
2	01	X_2	0.26
3	001	X_3	0.12
5	00010	X_4	0.04
5	00011	X_5	0.04
5	00001	X_6	0.03
5	00000	X_7	0.02

The corresponding tree fromwhich we construct the codes:



5.7 Some Comments on Huffman Codes

1. Huffman Encoding is conceptually equivalent to an ideal game of 20 questions
2. Huffman Encoding can be applied to weights ($p_i \geq 1$). In this case, the algorithm will minimize the sum code lengths
3. Something about Slice Codes
4. Code words for infrequent bits are much longer in Shannon Codes than in Huffman Codes. Either code can be shorter for a given symbol, but the Huffman code is always shorter on average (differ by at most 1 bit)

5. Fano Codes are non-optimal and achieve $L(C) \leq H(X) + 2$

5.8 Optimality of Huffman Codes

Lemma 5.7 *For any distribution, there exists an optimal instantaneous code that satisfies the following properties:*

1. *Lengths are inversely ordered with probabilities*
2. *The two longest codewords have the same length*
3. *Two of the longest codewords only differ by the last bit and correspond to the two least likely symbols*

Proof:

- Lengths are inversely ordered with probabilities

  Consider an optimal code C and a code C' that swaps two code lengths i and j such that $l_i \leq l_j$ and $p_i > p_j$.

  $$L(C) - L(C') = \sum_k p_k(l_k - l'_k) = p_i(l_i - l_j) + p_j(l_j - l_i) = (p_j - p_i)(l_j - l_i) \leq 0$$

  Since the swap increases code length we know it to be non-ideal.

- Two longest Codewords have the same length

  If the two longest codewords did not have the same length you can simply trim the larger code and reduce average code length. We can ensure that this would not coincide with an existing code by citing the prefix property.

- Two longest codewords only differ by the last bit and correspond ot the two least likely symbols

  We can replicate the argument above to state that the longest codeword must have a sibiling. From here we can construct a code in which the two lowest probability events are sibilings. This ensures that the two longest codewords differ by the final bit andalso do not change the average length.

Optimal codes that satisfy the lemma are known as canonical codes.

For an ordered probability distribution over an alphabet of size m we define the Huffman reduction as $p' = (p_1, ...p_{m-2}, p_{m-1} + p_m)$ defined over an alphabet of size $m - 1$. Optimality will be proved by showing that we can extend the optimal code for the Huffman reduction.

In the optimal code C^*_{m-1} consider the code corresponding to $p_{m-1} + p_m$. Extend it by adding a 0 to correspond to p_{m-1} or a 1 for p_m. The average length following this addition is

$$L(p) = L^*(p') + p_{m-1} + p_m$$

We can do the same with the optimal code C^*_m by joining the codes corresponding to p_{m-1} and p_m. It follows that

$$L(p') = L^*(p) - p_{m-1} - p_m$$

We can combine these expressions to write

$$(L(p') - L^*(p')) + (L(p) - L^*(p)') = 0$$

Each term in this expression is bound by ≥ 0 since the L^* corresponds to the optimal length. The expression can only evaluate to 0 if both terms are equal to 0 implying that the respective extensions are optimal. We can then use induction starting with a base case of 2 symbols to show that this is optimal for all code lengths.

Theorem 5.8 *Huffman Encoding, C^*, is Optimal*

$$L(C^*) - L(C') \leq 0$$

Note that Huffman encoding is a greedy algorithm but global optimality is shown above.

5.9 Shannon-Fano-Elias Coding

In this section we will discuss a coding procedure involving the cumulative distribution function for some probability distribution. Let our state space be $\mathcal{X} = \{1, 2, ..., m\}$. The CDF is given by

$$F(x) = \sum_{a \leq x} p(a)$$

Now, we introduce a modified CDF

$$\bar{F}(x) = \sum_{a < x} p(a) + \frac{1}{2} p(x)$$

Intuitively, this modified CDF corresponds to the midpoint of the steps in the discrete CDF. $F(x)$ provides a unique code for any x; however, to represent $F(X)$ we require a theoretical infinite number of bits to represent the real numbers it can realize. Instead, we can approximate $F(x)$ via $[F(x)]_{l(x)}$ where we truncate $F(x)$ to $l(x)$ bits. This rounding places an inherent bound

$$F(x) - [F(x)]_{l(x)} < \frac{1}{2^{l(x)}}$$

If we select $l(x) = [\log \frac{1}{p(x)}] + 1$,

$$\frac{1}{2^{l(x)}} < \frac{p(x)}{2} = \bar{F}(x) - \bar{F}(x-1)$$

We see that if we cleverly select $l(x)$ the truncation lies between each $F(x)$ step. Therefore, we do not need any additional bits to capture the state.

Now, let's verify the prefix free property of this code. We argue this by saying that each code $z_1 z_2 ... z_l$ is the interval $[0.z_1 z_2 ... z_l + \frac{1}{2^l})$. The code is prefix free if these intervals are disjoint. Since each interval is less than half of the step size, we can safely conclude that the sets are disjoint. Probably make this more rigorous.

The average code length for this code is

$$L = \sum p_i [\log \frac{1}{p_i} + 1] < H(X) + 2$$

5.10 Competitive Optimality of Shannon Code

Competitive Optimality considers the conditions underwhich a code performs better than others. You can think of it as a game between two codes who each collect points for producing shorter codes for randomly generated source codes. This is difficult to talk about for Huffman codes since there is no explicit form for lengths of codes, but we can meaningfully speak about Shannon codes.

Theorem 5.9 *Let $l(x)$ be the codeword lengths associated with the Shannon Code and $l'(x)$ the codeword lengths associated with another uniquely decodable code.*

$$Pr\{l(X) \geq l'(X) + c\} \leq \frac{1}{2^{c-1}}$$

Proof: internalize!

$$Pr\{l(X) \geq l'(X) + c\} = Pr\{\left\lceil \log \frac{1}{p(X)} \right\rceil \geq l'(X) + c\} \tag{39}$$

$$\leq Pr\{\log \frac{1}{p(X)} \geq l'(X) + c - 1\} \tag{40}$$

$$= Pr\{p(X) \leq 2^{-l'(X)-c+1}\} \tag{41}$$

$$= \sum_{x:p(x)\leq 2^{-l'(X)-c+1}} p(x) \tag{42}$$

$$\leq \sum_{x:p(x)\leq 2^{-l'(X)-c+1}} 2^{-l'(X)-c+1} \tag{43}$$

$$\leq \sum_{x} 2^{-l'(X)-c+1} \tag{44}$$

$$\leq 2^{-c+1} \tag{45}$$

FINISH LATER

5.11 Generation of Discrete Distribution from Fair Coins

Given a sequence of fair coin tosses $Z_1, Z_2...$ how can we generate a discrete random variable with probability mass function $(p_1, ...p_m)$? The idea is to map binary strings of coin flips to possible outcomes of the random variable in the form of a tree. Trees generated by this algorithm must 1) Be complete: every node is a leaf or has 2 descendants 2) Probability of a leaf at depth k is 2^-k and 3) The expected number of flips required to generate X is equal to the expected depth of the tree.

Lemma 5.10 *For any complete tree, consider a probability distribution on the leaves such that all leaves at depth k have probability 2^{-k}. The expected depth of this tree is the entropy of this distribution.*

Proof: The expected depth is given by

$$\mathbb{E}[T] = \sum_{y\in\mathcal{Y}} k(y)2^{-k(y)}$$

Entropy is given by

$$H(Y) = -\sum_{y\in\mathcal{Y}} \frac{1}{2^{k(y)}} \log \frac{1}{2^{k(y)}} = \sum_{y\in\mathcal{Y}} k(y)2^{-k(y)}$$

Therefore, $\mathbb{E}[T] = H(Y)$

Theorem 5.11 *For any algorithm generating X, the expected number of fair bits is greater than entropy*

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

Proof: Any algorithm generating X from fair bits can be represented by a tree. If we label our leaves $\mathcal{Y}$ we can define a random variable Y on this state space with probability 2^{-k} where k is the depth of y. From the lemma we know that $\mathbb{E}[T] = H(Y)$; however, we know that Y is a function of X so $H(X) \leq H(Y)$. It follows that $\mathbb{E}[T] \geq H(Xs)$

Theorem 5.12 *If the distribution is diadic then the expected number of fair bits to reproduce X is given by* $\mathbb{E}[T] = H(X)$

Proof: We can construct a Huffman tree which we know to saturate the entropy bound.

For non-dyadic distributions we break probabilities into sums of reciprocal powers of 2.

$$p(i) = \sum_{j\geq 1} p_i^{(j)} \qquad p_i^{(j)} = 2^{-j} \text{ or } 1$$

We then allot binary strings with the corresponding leaves to the event in question.

Theorem 5.13 *The expected number of fair bits required to generate X by the optimal algorithm is bounded by*

$$H(X) \leq \mathbb{E}[T] \leq H(X) + 2$$

Add Proof

6 Gambling and Data Compression

6.1 The Horse Race

Consider a gambler betting on horses. There are m horses that have a probability p_i of winning and an associated payoff of o_i. Let b_i denote the fraction of their wealth that the gambler places on the ith horse. Since the gambler receives nothing if the horse loses their total wealth will increase by $b_i o_i$ if the ith horse wins. To better understand how to maximize the the value of this random variable, we can consider repeating this process over n races. Let S_n be the gambler's total wealth:

$$S_n = \prod_{i=1}^{n} S(X_i)$$

Definition 27 (Wealth Relative) *, $S(X)$, is the factor by which the gambler's wealth increases if horse X wins*

Definition 28 (Doubling Rate) *is defined as*

$$W(b,p) = \mathbb{E}[\log S(X)] = \sum_{k=1}^{m} p_k \log b_k o_k$$

Theorem 6.1 *Suppose the outcome of each race is drawn iid $p(x)$. Then, the wealth of the gambler grows exponentially according to $W(b,p)$*

$$S_n = 2^{nW(b,p)}$$

Proof: Since functions of independent random variables are also independent, we can consider $\log(S(X_1))...\log(S(X_n))$ as a sequence of iid random variables as well. Now,invoking the weak law of large numbers,

$$\frac{1}{n}\log S_n = \frac{1}{n}\sum_{i=1}^{n}\log(S(X_i)) \to \mathbb{E}[S(X)]$$

Therefore, it is clear that we desire to optimize $W(b,p)$ over the possible wealth allocations b. The optimal doubling rate is given by proportional gambling. Consider the following optimization problem:

$$\max_b W(b,p) = \sum p_k \log b_k o_k \quad \text{wrt} \quad \sum b_i = 1$$

We can rewrite this as a lagrange multiplier optimization problem and consider

$$J(b) = \sum p_i \ln b_i o_i + \lambda \sum b_i$$

Differentiating across each b_i and setting the result to 0 yields

$$b_i = -\frac{p_i}{\lambda}$$

The constraint only holds true if $\sum b_i = \sum -\frac{p_i}{\lambda} = 1$ or alternatively $\lambda = -1$. This implies that $b_i = p_i$ or that the optimal betting strategy is proportional to the probabilities of occurence. This gambling strategy is known as Kelly gambling. Note that we ignored second order optimality considerations and will verify that this is indeed the optimum in the following section:

Theorem 6.2 (Proportional Gambling is log-optimal) *Optimal doubling rate is*

$$W^*(b,p) = \sum p_i \log o_i - H(p)$$

and this value is optimized when $b^* = p$

Proof:

$$W(b,p) = \sum p_i \log b_i o_i \tag{46}$$
$$= \sum p_i \log \frac{b_i}{p_i} o_i p_i \tag{47}$$
$$= \sum p_i \log o_i - H(p) - D(p||b) \tag{48}$$
$$\leq \sum p_i \log o_i - H(p) \tag{49}$$

Equality is only reached if $b = p$

Consider an alternate intterpretation of the doubling rate. We take a race track with fair odds (no track take). Let $r_i = 1/o_i$ be the bookie's estimate of win probabilities.

$$W(b,p) = \sum p_i \log b_i o_i = \sum p_i \log \left(\frac{b_i}{p_i} \frac{p_i}{r_i} \right) = D(p||r) - D(p||b)$$

The doubling rate can be interpretted as a measure of the distance between the true distrubtion and the bookie's estimate and your estimate and the true distribution. You only have positive doubling rate if your estimate is closer than the bookies. If the odds are m-for-1 or fair with respect to a uniform distribution we can further simplify this to say

$$W^*(p) = D(p||\frac{1}{m}) = \log m - H(p)$$

Duality to data compression?

Theorem 6.3 (Conservation Theorem) *For uniform fair odds the sum of the doubling rate and entropy rate is constant*

$$W(p) + H(p) = \log m$$

Every bit of entropy decrease increases the gamblers doubling rate. Therefore, low entropy races are the most profitable. If we remove the condition that the gambler must invest their full wealth in each race, we can study the system as 3 subcases. First, we can rewrite the total welath equation as

$$S(X) = b(0) + b(X)o(X)$$

where $b(0)$ denotes the proportion of wealth uninvested.

Fair Odds

For fair odds the analysis does not change. Proportional betting remains optimal. We can argue this by betting $b_i = \frac{1}{o_i}$ for each outcome i. Under this scheme $S(X) = 1$ regardless of whether the proprotions include the withheld or non-withheld cash. In other words, the withheld cash can be proportionally distributed across existing bets and that would not change the outcome. Therefore, the analysis remains the same and we can conclude that proportional betting is optimal.

Need to internatlize this argument more

Superfair

Under a superfair betting schemea,$\sum \frac{1}{o_i} < 1$. In this situation, it is ideal to place all cash into the race. Once again, proportional betting is optimal. We can also follow the Dutch betting schemea which allocates wealth according to $b_i = \frac{c}{o_i}$ where c is given by $\frac{1}{\sum \frac{1}{c_i}}$ This yields $o_i b_i = c$ regardless of which horse wins. There is no risk under this allotment since $S(X) = \frac{1}{\sum \frac{1}{o_i}} > 1$. The Dutch Book does not optimize doubling rate.

I don't follow Dutch betting. What is c_i? Show that the dutch book doesn't optimize

Subfair

Subfair betting odds refers to $\sum \frac{1}{o_i} > 1$. This system is representative of real life where the race organizers take a portion of the winnings. Proportional gambling is not optimal under this schemea; instead, it is ideal to leave some cash aside.

6.2 Gambling and Side Information

This section addresses the value of additional information when betting. In terms of the horse racing analogy we have been constructing, we can think of side information as the outcomes of horses in past races. Let's consider a horse $X = \{1, 2, .., m\}$ win the race with probability $p(x)$ and payout $o(x)$. Furthermore, let Y represent the side information. (X, Y) have a joint distribution $p(x, y)$. Consider the conditional betting allocation $b(x|y) \geq 0$ and $\sum_x b(x|y) = 1$. In comparison, let $b(x)$ denote the unconditional betting allocaiton. We have two associated doubling rates:

$$W^*(X) = \max_{b(x)} \sum_x p(x) \log b(x) o(x) \qquad W^*(X|Y) = \max_{b(x|y)} \sum_{x,y} p(x, y) \log b(x|y) o(x)$$

Let $\Delta W = W^*(X|Y) - W^*(X)$

Theorem 6.4 *The increase in doubling rate assocaited with side information is given by the mutual information between the side information and the betting variable*

$$\Delta W = I(X; Y)$$

Proof: First, we argue that the optimal conditional betting allocation $b(x|y) = p(x|y)$.

$$\begin{aligned} W(X|Y) = \max_{b(x|y)} \mathbb{E}[\log S(X)] &= \max_{b(x|y)} \sum p(x|y) \log b(x|y) o(x) \\ &= \sum p(x|y) \log p(x|y) o(x) \\ = \sum p(x) o(x) - H(X|Y) & \end{aligned}$$

Review why the optimal argument for fucntions of the form of entropy is the probability

From previous work we know that the optimal non-conditional betting strategy is also proportional: $W(X) = \sum p(x) o(x) - H(X)$. Putting it together, the increase in doubling rate assocaited with side information is

$$\Delta W = \sum p(x) o(x) - H(X|Y) - \sum p(x) o(x) - H(X) = H(X) - H(X|Y) = I(X; Y)$$

6.3 Dependent Horse Races and Entropy Rate

Consider a series of horse races. It is likely that the outcomes of future races are informed by the outcomes of previous runs. We can model this dependence via a stochastic process. Our betting strategy will then be dependent on this stochastic process. The doubling rate will be given by:

$$W(X_n|X_{n-1}, ..., X_0) = \mathbb{E}\Big[\max \mathbb{E}[\log S(X_n|X_{n-1}, ..., X_0)]\Big] = \log m - \log H(X_n|X_{n-1}, ..., X_0)$$

Optimality is reached when $b(x_n|o_{n-1}, ...x_0) = p(x_n|o_{n-1}, ...x_0)$. The gambler's wealth is still given by

$$S_n = \prod S(X_i)$$

Using the same reasoning as before,

$$\begin{aligned} \frac{1}{n}\mathbb{E}[\log S_n] &= \frac{1}{n} \sum \mathbb{E}[\log S(X_i)] \\ &= \frac{1}{n} \sum (\log m - H(X_i|X_{i-1}, ...X_0)) \\ &= \log m - \frac{H(X_1, ..., X_n)}{n} \end{aligned}$$

If our race outcomes follow a stationary process the limit of the second term is the entropy rate. It follows that the sum of the doubling rate and the entropy rate is constant.

Review betting on cards example

6.4 Entropy of English

6.5 Data Compression and Gambling

A good gambler makes accurate predictions on the distribution of data. This data can be in turn used as a data comprehension scheme so a good gambler is actually a good data compressor. Add example

6.6 Gambling Estimate of the Entropy of English

Review previous section + see if there is any insight

7 Information Theory and Portfolio Theory

7.1 The Stock Market: Some Definitions

We can formalize the notion of the stock market by considering it as a bundle of stocks represented by a vector $X = (X_1, ..., X_m)$. Each entry in the vector X corresponds to the daily change in the *ith* stock (price relative). Under this representation we have m total stocks. Suppose $X \sim F(X)$ which is the joint distribution of the price relatives. We define a portfolio to be a wealth allocation across the stocks. Portfolios are represented by another vector $b = (b_1, ..., b_m)$ where each entry represents the proportion of wealth invested in the *ith* stock. Under this definition $\sum b_i = 1$. At the end of each day, for price relative X and portfolio b the wealth relative is given by $S = b^T X = \sum b_i X_i$. The goal is to roughly optimize S but since it is a random variable we have multiple choices for the best distribution S.

Traditional financial investment simplifies this problem to focusing on the mean and variance of the distribution. This formulation born from the Sharpe-Markowitz theory of investment is dependent solely on the first and second moments of the distribution. Under this theory,for a given mean and variance we can understand the optimal allocation as a set of portfolios lying on the efficient frontier. We can further simplify portfolio optimization by introducing a risk-free asset which allows you to work with a 0-variance and fixed interest rate asset. The introduction of this financial entity allows you to expand your optimal portfolio space.This theory also implies the Capital Asset Pricing Model (CAPM) or that there is a fixed true price for a stock. This evaluation can be used to determine if an asset is under or overvalued.

Definition 29 (Growth Rate) *The growth rate of a portfolio b with respect to a stock distribution $F(x)$ is given by*

$$W(b, F) = \int \log b^t x dF(x) = \mathbb{E}[\log b^t X]$$

The growth rate is a doubling rate if the logarithm is base 2. Intuitively, we understand the expectation of a logarithm as a consequence of daily reinvestment in the stock market. Traditionally, whwen we want to consider the long run behavior of a random variable we consider the sum off i.i.d realizations of the variable; however, in the stock market we tned to reinvest our money at the end of each day. This means that our wealth grows as the product of performances during each day. Therefore, an expected logarithm is more apt at capturing the long run behavior.

Definition 30 *The Optimal Growth Rate is given as the maximum portfolio allocation for a given stock distribution*

$$W^*(F) = \max_b W(b, F)$$

The optimal portfolio is referred to as the log-optimal or growth optimal portfolio.

Theorem 7.1 *Let $X_1, ..X_n$ be i.i.d. according to the stock distribution $F(x)$. Suppose S_n denotes the wealth accrued using a daily rebalanced portfolio b^*.*

$$S_n = \prod_{i=1}^{n} b^{*t} X_i$$

Then

$$\frac{1}{n}\log S_n^* \to W^*$$

with probability 1

Proof: This follows from the strong law of large numbers:

$$\frac{1}{n}\log S_n^* \to \frac{1}{n}\sum \log b^{*t} X_i \to W^*$$

Lemma 7.2 *$W(b,F)$ is concave in b and linear in F. $W^*(F)$ is concave in F.*

Proof: Condsider the growth rate formula

$$W(b,F) = \int \log b^t x dF(x)$$

The growth rate is linear in F since the integral is. We can show the concavity with respect to b by showing the concavity of the logarithm. Finish Proof Later

Lemma 7.3 *The set of log-optimal portfolios with respect to a given distribution is convex*

Proof: This follows directly from the concavity of $W(b,F)$ with respect to b.

7.2 Kuhn-Tucker Characterization of the Log-Optimal Portfolio

Theorem 7.4 *The log-optimal portfolio b^* for a stock distribution obeys the following neccesary and sufficient conditions*

$$\begin{aligned} \mathbb{E}\left(\frac{X_i}{b^T X_i}\right) &= 1 \qquad \text{if } b_i^* > 0 \\ &\leq 1 \qquad \text{if } b_i^* = 0 \end{aligned}$$

Proof:

8 Channel Capacity

9 Differential Entropy