Why is entropy a fundamental measure of information content?

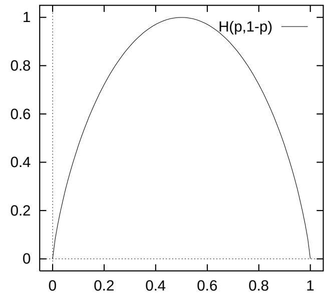

Figure 1: The binary entropy function $H(p, 1-p)$ as a function of p.

We will verify the validity of $H(X)$ as a measure of information by relating $H(X)$ to the actual number of bits needed to specify the outcome of an experiment.[1]

An ensemble 'X' is a random variable x with a set of possible *outcomes*, $\mathcal{A}_X = \{a_1, \ldots, a_I\}$, having probabilities $\mathcal{P}_X = \{p_1, \ldots, p_I\}$, with $P(x=a_i) = p_i$, $p_i \geq 0$ and $\sum_{x \in \mathcal{A}_X} P(x) = 1$.

Number of elements in a set $\mathcal{A}$ is denoted by $|\mathcal{A}|$.

The entropy of X is defined by:

$$H(X) \equiv \sum_{x \in \mathcal{A}_X} P(x) \log \frac{1}{P(x)},$$

with the convention for $P(x)=0$ that $0 \times \log 1/0 \equiv 0$, since $\lim_{\theta \to 0^+} \theta \log 1/\theta = 0$.

Note that entropy is additive for independent variables.

Perfect information content of X is:

$$H_0(X) = \log_2 |\mathcal{A}_X|$$

$H_0(X)$ is a lower bound for the number of binary questions that are guaranteed to identify the outcome. It is an additive quantity: $H_0(X, Y) = H_0(X) + H_0(Y)$.

This measure of information content does not include any probabilistic element.

Essential information. We relax the exhaustive requirement, and define:

$$H_\delta(X) = \log \min \{|T| : T \subseteq \mathcal{A}_X, \Pr(x \in T) \geq 1 - \delta\}.$$

Here the minimization seeks out the smallest possible subset T of outcomes that have the biggest possible probability. $P(x \notin T) < \delta$.

Note that $H_0(X)$ is the special case of $H_\delta(X)$ with $\delta = 0$, if $p(x) > 0$ for all $x \in X$.

[0] These and previous lecture notes are also available by `www` or `ftp` at: `ftp://131.111.48.24/pub/mackay/info-theory/course.html`

[1] I am indebted to Yaser Abu-Mostafa for the following presentation of the source coding theorem.

Example 1:
Let $\mathcal{P}_X = \{1/4, 1/4, 1/4, 3/16, 1/64, 1/64, 1/64, 1/64\}$. Then $H_0(X) = 3$ bits, $H_{1/16}(X) = 2$ bits. So if we are willing to run a risk of $\delta = 1/16$ of not having a name for x, then we can get by with half as many names as are needed if every $x \in \mathcal{A}_X$ must have a name.

Example 2: Consider $\mathbf{x} = (x_1, x_2 \ldots x_N)$ where $x_i \in \{0, 1\}$, with probabilities $p_0=0.9, p_1=0.1$. The most probable strings $\mathbf{x}$ are those with most 0's. If $r(\mathbf{x})$ is the number of 1's in $\mathbf{x}$ then

$$P(\mathbf{x}|p_0, p_1) = p_0^{N-r(\mathbf{x})} p_1^{r(\mathbf{x})}.$$

To evaluate $H_\delta(\mathbf{X})$ we must find the smallest possible subset T such that $P(\mathbf{x} \notin T) < \delta$. Clearly, this minimal subset will contain all $\mathbf{x}$ with $r(\mathbf{x}) = 0, 1, 2 \ldots$, up to some $r_{\max}(\delta)$. Figure 2 shows a graph of $H_\delta(\mathbf{X})$ against δ for the cases $N = 4$ and $N = 10$. The cusps are the points where $r_{\max}$ changes by 1.

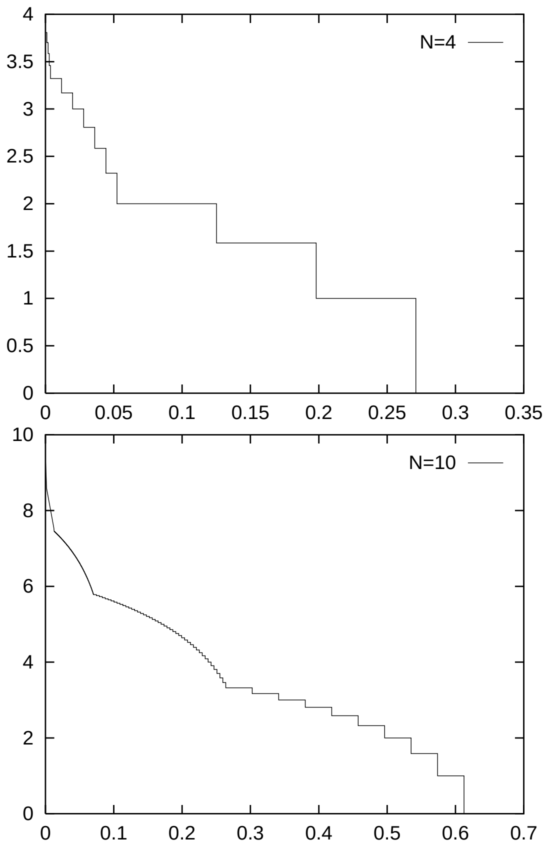

Figure 2: $H_\delta(\mathbf{X})$ (vertical axis) against δ (horizontal), for $N = 4$ and $N = 10$ binary variables with $p_1 = 0.1$.

Assymptotic Equipartition and Source Coding

We will prove the following:

Assymptotic Equipartition Principle (AEP): for an ensemble of N independent identically distributed (i.i.d.) random variables $X^N \equiv (X_1, X_2, \ldots X_N)$, with N sufficiently large, the outcome $\mathbf{x} = (x_1, x_2, \ldots x_N)$ is almost certain to belong to a subset of $\mathcal{A}_X^N$ having only $2^{NH(X)}$ members, and all having probability 'close to' $2^{-NH(X)}$.

If $H(X) < H_0(X)$ then $2^{NH(X)}$ is a *tiny* fraction of the number of possible outcomes $|\mathcal{A}_X^N| = |\mathcal{A}_X|^N = 2^{NH_0(X)}$.

The AEP is equivalent to

Shannon's source coding theorem: N i.i.d. random variables each with entropy H can be compressed into more than NH bits with negligible loss of information, as $N \to \infty$; conversely if they are compressed into fewer than NH bits there is a dramatic fall-off of information.

because we can define a compression algorithm that gives a distinct name of NH bits to each $\mathbf{x}$ in the probable subset.

We will prove the AEP by showing that for any given δ there is a sufficiently big N such that $H_\delta(X^N) \simeq NH$.

Theorem 1 *Let X be an ensemble with entropy $H(X) = H$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for $N > N_0$,*

$$\left| \frac{1}{N} H_\delta(X^N) - H \right| < \epsilon.$$

Sketch of proof. The random variable $\frac{1}{N}\log\frac{1}{p(\mathbf{x})}$ is very likely to have value very close to H. Define a subset T of $\mathcal{A}_X^N$ to be the 'typical' elements of $\mathcal{A}_X^N$ that have probability very close to 2^{-NH} (note that T does not include the most probable elements of $\mathcal{A}_X^N$, but we will show that they contribute negligible probability). Show that we can achieve $P(T) = 1 - \delta$ using only 'typical' outcomes. If $P(T) \simeq 1$, it must be that the number of elements of T is $|T| \simeq 2^{NH}$.

We can therefore capture practically all the probability of X^N using an asymptotically negligible fraction of its elements.

The law of large numbers

Mean and variance of a random variable[2] are $E[u] = \bar{u} = \sum_u P(u)u$ and $\mathrm{var}(u) = \sigma_u^2 = E[(u-\bar{u})^2] = \sum_u P(u)(u-\bar{u})^2$.

Chebyshev's inequality 1. Let t be a non-negative real random variable, and let α be a positive real number. Then $P(t \geq \alpha) \leq \bar{t}/\alpha$.

Proof: $P(t \geq \alpha) = \sum_{t\geq\alpha} P(t)$. We multiply each term by $t/\alpha \geq 1$ and obtain: $P(t \geq \alpha) \leq \sum_{t\geq\alpha} P(t)t/\alpha$. We add the (non-negative) missing terms and obtain: $P(t \geq \alpha) \leq \sum_t P(t)t/\alpha = \bar{t}/\alpha$.

Chebyshev's inequality 2. Let x be a random variable, and let α be a positive real number. Then $P\left((x-\bar{x})^2 \geq \alpha\right) \leq \sigma_x^2/\alpha$.

Proof: Take $t = (x-\bar{x})^2$ and apply the previous proposition.

Weak law of large numbers. Take x to be the average of N independent random variables $h_1, \ldots, h_N$, having common mean $\bar{h}$ and common variance σ_h^2: $x = \frac{1}{N}\sum_{n=1}^N h_n$. Then $P((x-\bar{h})^2 \geq \alpha) \leq \sigma_h^2/\alpha N$.

Proof: obtained by showing that $\bar{x} = \bar{h}$ and that $\sigma_x^2 = \sigma_h^2/N$.

We are interested in being very close to the mean (α very small). No matter how large σ_h^2 is, and no matter how small the required α is, and no matter how small the probability of $(x-\bar{h})^2 \geq \alpha$ is desired to be, we can always achieve it by taking N large enough.

Proof of the theorem

We apply the law of large numbers to the random variable $\frac{1}{N}\log\frac{1}{P(\mathbf{x})}$ defined for $\mathbf{x}$ drawn from the ensemble X^N. This random variable can be written as the average of N terms $\log(1/P(x_n))$, each of which is a random variable with mean $H = H(X)$ and variance $\sigma^2 \equiv \mathrm{var}[\log(1/P(x_n))]$. We define a 'typical' subset with parameters N and β thus:

$$T_{N\beta} = \left\{ \mathbf{x} \in \mathcal{A}_X^N : \left[\frac{1}{N}\log\frac{1}{P(\mathbf{x})} - H\right]^2 < \beta^2 \right\}.$$

For all $\mathbf{x} \in T_{N\beta}$, $2^{-N(H+\beta)} < P(\mathbf{x}) < 2^{-N(H-\beta)}$. And by the law of large numbers,

$$P(\mathbf{x} \in T_{N\beta}) \geq 1 - \frac{\sigma^2}{\beta^2 N}.$$

Part 1 of theorem. $\frac{1}{N}H_\delta(X^N) < H + \epsilon$.

We show how *small* $H_\delta(X^N)$ must be by calculating the largest cardinality that $T_{N\beta}$ could have. Since the smallest possible probability that a member of $T_{N\beta}$ can have is $2^{-N(H+\beta)}$, and the largest total probability that $T_{N\beta}$ could contain is 1, we can bound

$$|T_{N\beta}| < 2^{N(H+\beta)}.$$

Setting $\beta = \epsilon$ and N_0 such that $\frac{\sigma^2}{\epsilon^2 N} \leq \delta$, so that $P(T_{N\beta}) \geq 1 - \delta$, $T_{N\beta}$ becomes a witness to the fact that $H_\delta(X^N) < N(H+\epsilon)$.

Part 2 of theorem. $\frac{1}{N}H_\delta(X^N) > H - \epsilon$.

We prove that an alternative *smaller* subset T' having $|T'| \leq 2^{N(H-2\beta)}$ and achieving $P(\mathbf{x} \in T') \geq 1 - \delta$ cannot exist (for N greater than an N_0 that we will specify). The probability of the subset T' is $P(\mathbf{x} \in T' \cap T_{N\beta}) + P(\mathbf{x} \in T' \cap \overline{T_{N\beta}})$, where $\overline{T_{N\beta}}$ denotes the complement $\{\mathbf{x} \notin T_{N\beta}\}$.

[2] Technical note: strictly I am assuming here that u is a function $u(x)$ of a sample x from a finite discrete ensemble X. Then the summations $\sum_u P(u)f(u)$ should be written $\sum_x P(x)f(u(x))$. This means that $P(u)$ is a finite sum of delta functions. This restriction guarantees that the mean and variance of u do exist, which is not the case for general $P(u)$.

The maximum value of the first term is found if $T' \cap T_{N\beta}$ contains $2^{N(H-2\beta)}$ outcomes all with the maximum probability, $2^{-N(H-\beta)}$. The maximum value of the second term is $P(\mathbf{x} \notin T_{N\beta})$:

$$P(\mathbf{x} \in T') \leq 2^{-N(H-\beta)} 2^{N(H-2\beta)} + \frac{\sigma^2}{\beta^2 N} = 2^{-N\beta} + \frac{\sigma^2}{\beta^2 N}.$$

We set $\beta = \epsilon/2$ and N_0 such that $P(\mathbf{x} \in T') < 1 - \delta$. This establishes that *any* subset T' such that $|T'| \leq 2^{N(H-\epsilon)}$ has probability less than $1 - \delta$, so by the definition of H_δ, $H_\delta(X^N) > N(H - \epsilon)$.

COMMENT

The theorem has two parts, $\frac{1}{N} H_\delta(X^N) - H < \epsilon$, and $H - \frac{1}{N} H_\delta(X^N) < \epsilon$. Both results are interesting.

The first part tells us that even if δ is extremely small, the number of bits per outcome needed to specify $\mathbf{x}$ with vanishingly small error probability, $\frac{1}{N} H_\delta(X^N)$, does not have to exceed $H + \epsilon$ bits. We only need to have a tiny tolerance to error, and the number of bits required drops significantly from $NH_0(X)$ to $N(H+\epsilon)$.

What happens if we are yet more tolerant to compression errors? Part 2 tells us that even if δ is very close to 1, so that an error is made most of the time, the number of bits per outcome needed to specify $\mathbf{x}$ still must be at least $H - \epsilon$ bits. These two extremes tell us that regardless of our specific allowance for error, the number of bits per outcome needed to specify X boils down to H bits; no more and no less.

Thus for large enough N, $\frac{1}{N} H_\delta(X^N)$ is essentially a constant function of δ. Figure 3 illustrates this assymptotic tendency for the example discussed earlier with N binary variables with $p_1 = 0.1$. As N increases, $\frac{1}{N} H_\delta(X^N)$ becomes an increasingly flat function, except for tails at $\delta = 0$ and 1. The limiting value of the plateau is $H(X) = 0.47$.

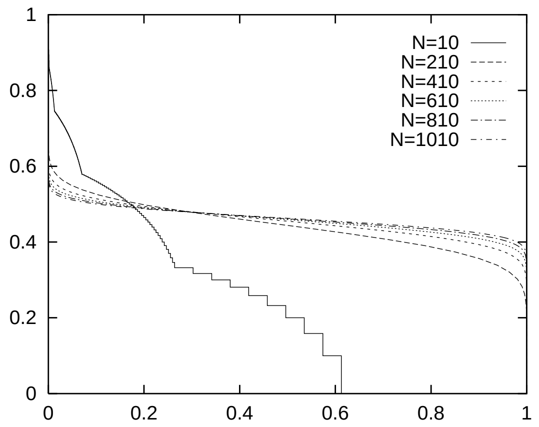

Figure 3: $\frac{1}{N} H_\delta(\mathbf{X})$ (vertical axis) against δ (horizontal), for $N = 10, \ldots, 1010$ binary variables with $p_1 = 0.1$.

In the next lecture we will discuss more practical data compression schemes that are practical for small block sizes and are guaranteed to function without error.

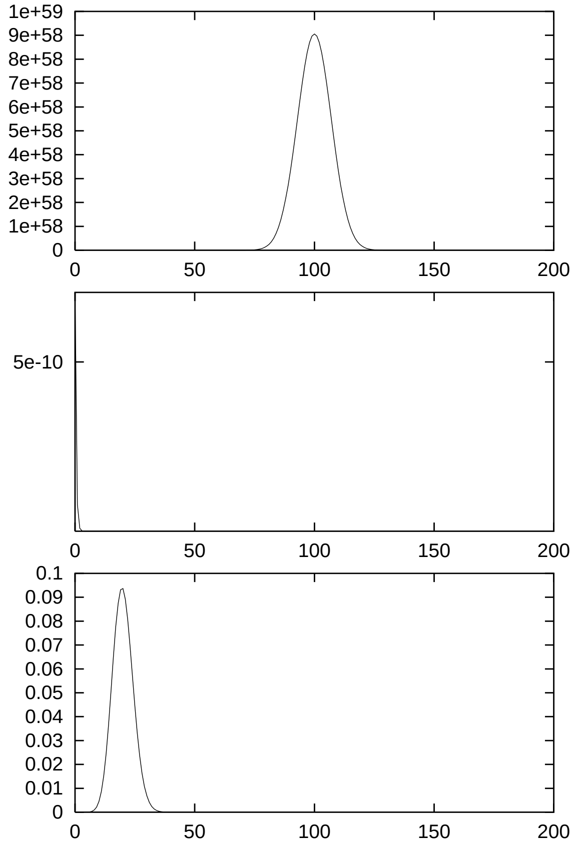

Figure 4: Anatomy of the typical set T

For $N = 200$ and $p_1 = 0.1$, these graphs show the number of strings containing r 1s (top), the probability of a single string that contains r 1s (middle), and the total probability of all strings that contain r 1s (bottom). The bottom graph is the product of the upper two. The number r is on the horizontal axis. The typical set used in the proof is all strings that contain about 20 1s. Note that this set does not include the most probable strings, which have fewer 1s. We do not bother including them because they have negligible total probability.