Week 1 - Entropy (Speaker: Roman Vershynin)

1 Entropy

Definition 1. Let X be a discrete (for now) random variable with $\mathbb{P}[X_i = x_i] = p_i$. We define the **entropy** of X to be

$$H(X) = -\sum_i p_i \log p_i$$

$$= \mathbb{E}\left[\log \frac{1}{p_X(x)}\right],$$

where the logarithm is to the base 2 and p_X is the probability mass function of X. We say that X has $H(X)$ **bits** of entropy.

Roughly speaking, entropy quantifies how much "information" is in a random variable.

Example 1. Say there are n possible outcomes for X, each occurring with probability $p_i = \frac{1}{n}$. Then

$$H(X) = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n}$$

$$= \log n.$$

Example 2. Suppose X is identically zero. Using the convention that $0 \cdot \log 0 = 0$, we have that $H(X) = 0$.

Example 3. Suppose X is a Bernoulli random variable with success probability p. Then the entropy of X is given by

$$H(X) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$$

$$=: h(p).$$

We call $h(p)$ the binary entropy function. Some basic calculus shows that this function is maximized when $p = \frac{1}{2}$. Intuitively speaking, a Bernoulli trial that has success probability is maximally "unpredictable", so its outcome carries more information.

Example 4. Suppose X is a geometric random variable with probability p. Then $\mathbb{P}[X = i] = (1-p)^i p$. The entropy of X is then

$$H(X) = \sum_{i=0}^{\infty} (1-p)^i p \log \frac{1}{p(1-p)^i}$$

$$= \frac{h(p)}{p}.$$

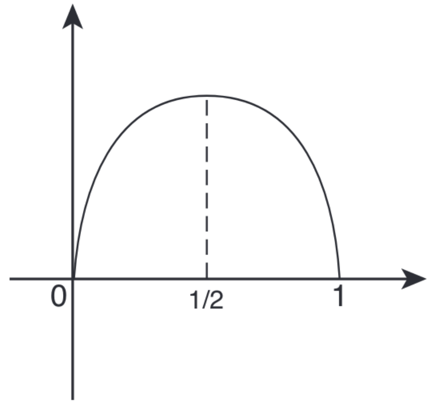

Figure 1: The graph of the binary entropy function $h(p)$.

Now we state a theorem about some of the nice properties entropy has. One can show that any function satisfying these properties is exactly our definition of entropy up to multiplication by a constant factor.

Theorem 1 (Properties of Entropy).

1. *$H(X) \geq 0$, with equality if and only if H is constant.*
2. *$H(X) \leq \log n$ if X has n possible outcomes, with equality if and only if X is uniformly distributed. (The most "unpredictable" variable is a uniform one.)*
3. *$H(X) = H(f(x))$ for any bijective f. (The labels don't matter, only the probabilities.)*
4. *$H(X|Y) \leq H(X)$, with equality if and only if X and Y are independent. (More information, i.e. conditioning, lowers uncertainty.)*
5. *$H(X, Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$. (A "chain rule".)*
6. *$H(X) \geq H(f(X))$, with equality if and only if f is injective.*
7. *$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{j<i})$, with equality if and only if the X_i are mutually independent. (A bigger "chain rule".)*

Proof.

1. Since $0 \leq p_i \leq 1$, $-\log p_i \geq 0$, so the sum defining entropy has only nonnegative terms.
2. The logarithm is concave, so we can apply Jensen's inequality:

$$\begin{aligned} H(X) &= \mathbb{E}[\log 1/p_X(x)] \\ &\leq \log \mathbb{E}[1/p_X(x)] \\ &= \log \mathbb{E}\left[\sum_{i=1}^{n} 1\right] \\ &= \log n. \end{aligned}$$

We didn't prove it during the seminar, but a slick proof for the "only if uniform" part I found online uses the weighted AM-GM inequality:

$$
\begin{aligned}
2^{H(X)} &= \prod_{i=1}^{n} p_i^{-p_i} \\
&\leq \sum_{i=1}^{n} p_i \cdot \frac{1}{p_i} \\
&= n,
\end{aligned}
$$

with equality if and only if the p_i are all equal.

3. $H(X)$ depends only on the probabilities associated to the outcomes of X, not the outcomes themselves.

4. We skipped the proof for this. It's allegedly coming later.

5. Follows from the definition of the joint distribution.

6. $x \mapsto (x, f(x))$ is injective, so by properties 3 and 5 we have

$$
\begin{aligned}
H(X) &= H(X, f(X)) \\
&= H(f(X)) + H(X|f(X)) \\
&\geq H(f(X)).
\end{aligned}
$$

We obtain equality if and only if $H(X|f(X)) = 0$, which happens if and only if X is constant given $f(X)$.

7. Same proof as property 5.

$\square$