Lesson 4
Entropy
4-2: The Definition of Entropy

Jan Reimann

Math 574, Topics in Logic
Penn State, Spring 2014

Entropy = Expected Information

Suppose X is a random variable taking values in a finite set A with distribution P.

We define

$$\begin{aligned} H(X) &= \sum_{a \in A} P(X = a)(-\log P(X = a)) \\ &= -\sum_{a \in A} P(X = a) \log P(X = a) \\ &= -\sum_{a \in A} P(a) \log P(a) \end{aligned}$$

We put $0 \log 0 = 0$. *This is consistent as* $x \log x \to 0$ *for* $x \to 0$.

In the last lecture we saw that $-\log P(X = a)$ can be seen as the information we gain from knowing $X = a$.

Hence $H(X)$ gives us the expected gain in information with respect to the distribution of the random variable X.

Properties of Entropy

$-\log P. \quad 0 \leq P \leq 1$

We have $\underline{H(X) \geqslant 0}$ and $H(X) = 0$ iff $P(X = a) = 1$ for some $a \in A$.

$H(X)$ depends only on the distribution of X. It is hence a function defined for any finite probability vector $p = (p_1, \dots, p_n)$, $p_i \geqslant 0$, $\sum p_i = 1$:

$$\underline{H(p)} = -\sum_i p_i \log p_i.$$

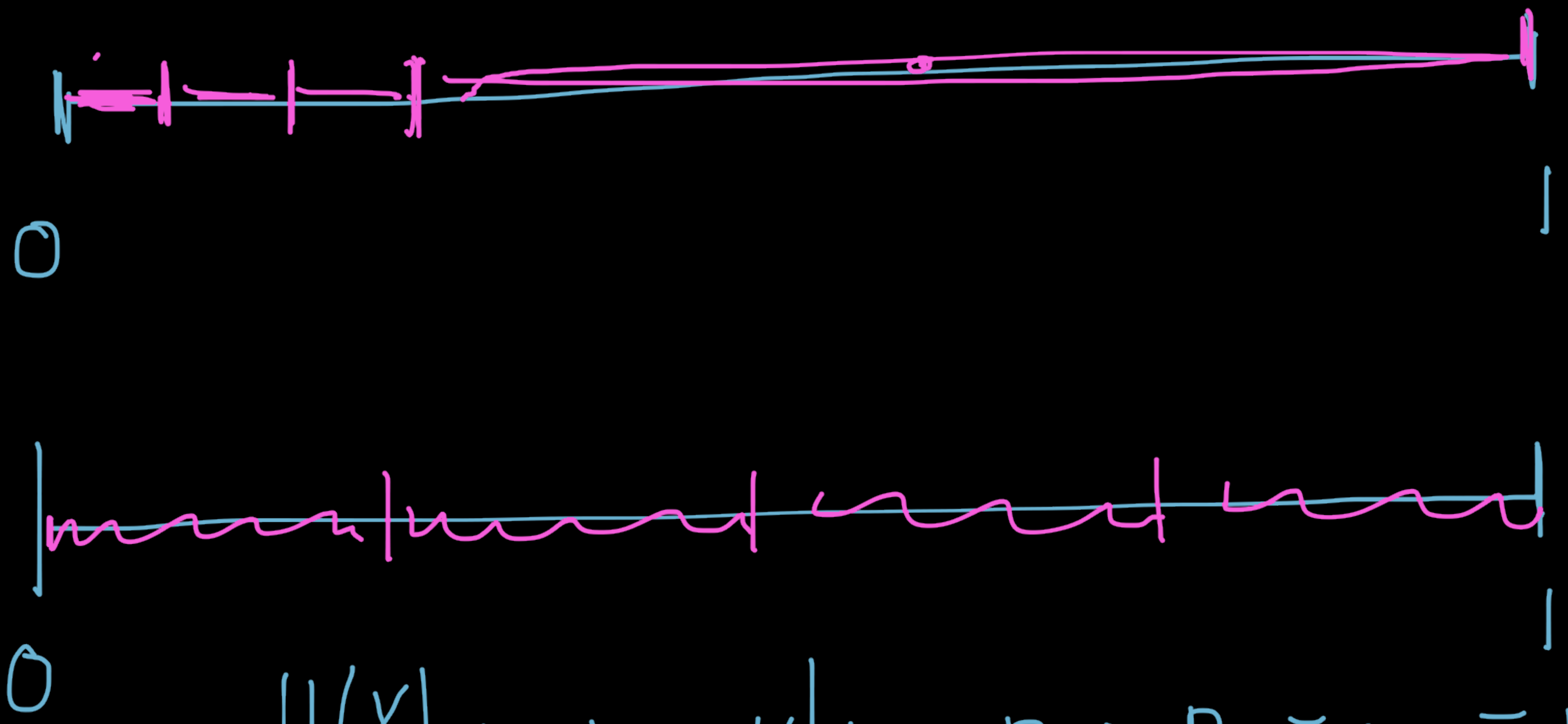
Then $H(X) = H(p)$ where $p = (P(X = a_1), \dots, P(X = a_n))$ for $A = \{a_1, \dots, a_n\}$.

It is clear that H does not change when we $\underline{\text{permute the}}$ p_i. It is a symmetric function.

Properties of Entropy

When does assume $H(X)$ the maximum value?

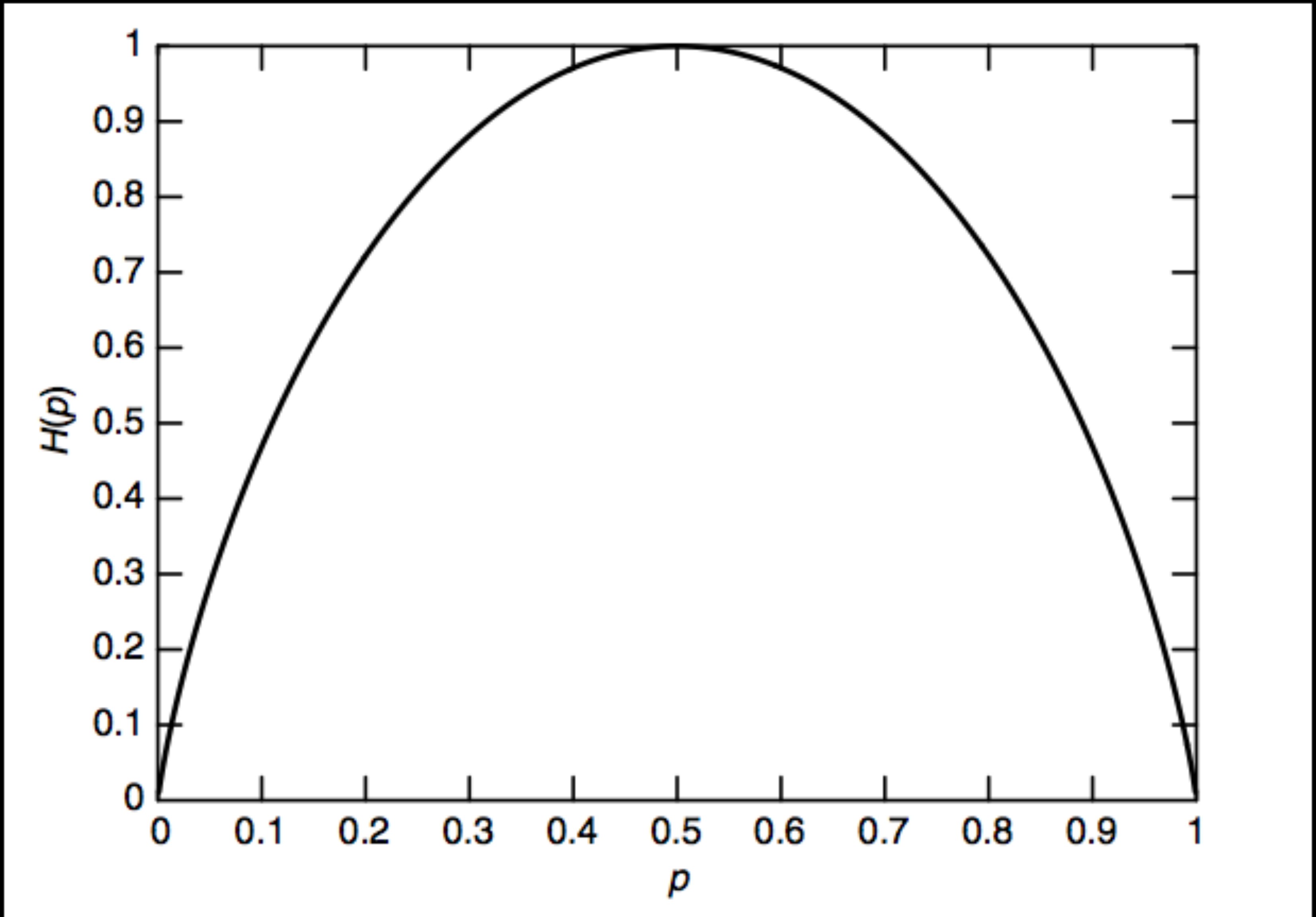
↑ continuous ↑ for fixed A



$H(X)$ max when $p_1 = p_2 = \cdots = p_n$

The Entropy Graph

Recall: $\log = \log_2$



$X \qquad A = \{0, 1\}$

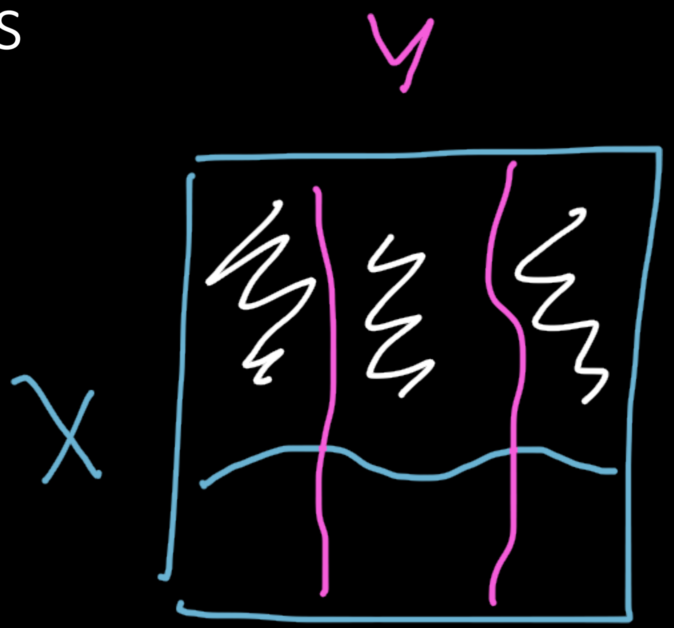
$P(X=0) = p \qquad P(X=1) = 1-p$

Joint Entropy

Assume now *X*, *Y* are random variables (with values in finite sets *A* and *B*, respectively).

The joint distribution of (X, Y) is given by the values

$$P(X = a, Y = b).$$



The joint entropy of *X*, *Y* is then defined as

$$\boxed{H(X,Y) =} -\sum_{a\in A}\sum_{b\in B} P(X = a, Y = b) \log P(X = a, Y = b)$$

which we can write simply as

$$\mathbb{E}(-\log P(X, Y))$$

Conditional Entropy

We can also define the conditional entropy $H(X|Y)$:

$H(X|Y) =$ entropy of $H(X|Y = b)$, averaged over all possible values for Y.

Formally,

$$H(X|Y) = \sum_{b \in B} P(Y = b) H(X|Y = b),$$

where the term $\underline{H(X|Y = b})$ denotes the entropy of the distribution of X conditioned on $Y = b$:

$$H(X|Y = b) = -\sum_{a \in A} P(X = a|Y = b) \log P(X = a|Y = b).$$

We put this together and obtain

$$\begin{aligned} H(X|Y) &= -\sum_{b \in B} P(Y = b) \sum_{a \in A} P(X = a|Y = b) \log P(X = a|Y = b) \\ &= -\sum_{b \in B} \sum_{a \in A} P(X = a, Y = b) \log P(X = a|Y = b) \\ &= \mathbb{E}(-\log P(X|Y)) \end{aligned}$$

Joint Entropy as Conditional Entropy

Interpreting entropy as information gain, the following equation makes sense intuitively:

Information gain from knowing *X* and *Y* =

Information gain from *X* + Information gain from *Y* given *X*.

THM: [Chain Rule] $H(X, Y) = H(X) + H(Y|X)$.

Proof: Straightforward, using $\log(xy) = \log(x) + \log(y)$, and observing that $H(X) = -\sum_A P(X = a) \log P(X = a)$ can be written as

$$H(X) = -\sum_A \sum_B P(X = a, Y = b) \log P(X = a).$$

Axiomatic Description of Entropy

Suppose H^* is defined for any A-valued random variable (A arbitrary finite set) that has the following properties:

1. $H^*(X) \geqslant 0$ and $H^*(X) = 0$ iff $P(X = a) = 1$ for some $a \in A$;
2. $\underrightarrow{H^*|_A}$ is continuous;

$A = \{a_1 \ldots , a_n\}$

$(p_1 — p_n)$

3. $H^*|_A$ is symmetric;
4. $H^*|_A$ takes its largest value for equidistributed X;
5. $H^*(X, Y) = H^*(X) + H^*(Y|X)$;
6. if $B = A \cup \{b\}$, X is A-valued, and Y trivially extends X in the sense that $P(Y = a) = P(X = a)$ for $a \in A$, $P(Y = b) = 0$, then $H^*(Y) = H^*(X)$.

Then there exists $\lambda > 0$ such that $H^* = \lambda H$.

If we moreover require that $H^*(1/2, 1/2) = 1$, then $\boxed{H^* = H}$.