Lecture 1

Entropy

1.1 What is Information?

Average number of bits to describe an random variable. Horse race

1.2 Entropy

H(X) is measure of uncertainty about a random variable X

1.3 Conditional Entropy

H(Y|X), Given you know X, the uncertainty of Y

1.4 Properties and Chain Rules for Conditional Entropy

Entropy and conditional entropy have many properties

1.1 Entropy: What is Information?

The Surreal Poet
http://bit.ly/1fU5Oqt

The Front Porch
http://bit.ly/1fU5Oqt

Information in a coin flip

50% Heads

50% Tails

1 Bit of Information

Information Is Represented by Random Variables

Random Variable Notation

For a random variable X:

- The sample space is $\mathcal{X}$. It is a set.
- $|\mathcal{X}|$ means "the number of elements in the set $\mathcal{X}$" or "the cardinality of $\mathcal{X}$".
- The probability distribution is $p_{\mathsf{X}}(x)$. Also written:

$$p_{\mathsf{X}}(x) = \Pr[\mathsf{X} = x]$$

- $p_{\mathsf{X}}(x)$ must satisfy:

$$0 \leq p_{\mathsf{X}}(x) \leq 1 \text{ and } \sum_{x \in \mathcal{X}} p_{\mathsf{X}}(x) = 1$$

Random Variable Example

Let Y be a random variable denoting the weather.

- The sample space $\mathcal{Y} = \{\text{sunny}, \text{cloudy}, \text{rainy}\}$.
- The cardinality is $|\mathcal{Y}| = 3$.
- The probability distribution is:

$$p_Y(y) = \begin{cases} \frac{1}{4} & \text{if } y = \text{sunny} \\ \frac{1}{2} & \text{if } y = \text{cloudy} \\ \frac{1}{4} & \text{if } y = \text{rainy} \end{cases}$$

which satisfies $\sum_{y \in \mathcal{Y}} p_Y(y) = 1$.

Who Won The Horse Race?

Imagine a race of 8 horses.

We want a code to send the winner of the race.

- Use as few bits as possible
- Consider the **average** number of bits to send the message

Adios
Big Brown
Cigar
Deep Impact
Easy Goer
Funny Cide
Go Man
Hyperion

Simple Code to Transmit the Winner

Adios	000
Big Brown	001
Cigar	010
Deep Impact	011
Easy Goer	100
Funny Cide	101
Go Man	110
Hyperion	111

3 bits to transmit winning horse

RACE 11
6:05
MTP 29
RACE 10
YIELD RESULTS
4
2
5
8
FAST
WIN ODDS
1 3 2 5 9 9 5
2 7 6 1 0 10 4 2
3 1 2 7 1 1 1 1 9/2

Odds of Winning

Suppose we use the following Variable-length code instead:

Adios	$\frac{1}{2}$	0
Big Brown	$\frac{1}{4}$	10
Cigar	$\frac{1}{8}$	110
Deep Impact	$\frac{1}{16}$	1110
Easy Goer	$\frac{1}{64}$	111100
Funny Cide	$\frac{1}{64}$	111101
Go Man Go	$\frac{1}{64}$	111110
Hyperion	$\frac{1}{64}$	111111

Adios	$\frac{1}{2}$	0
Big Brown	$\frac{1}{4}$	10
Cigar	$\frac{1}{8}$	110
Deep Impact	$\frac{1}{16}$	1110
Easy Goer	$\frac{1}{64}$	111100
Funny Cide	$\frac{1}{64}$	111101
Go Man Go	$\frac{1}{64}$	111110
Hyperion	$\frac{1}{64}$	111111

Expected Length

$$1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3\frac{1}{8} + 4\frac{1}{16} + 4\left(6 \cdot \frac{1}{64}\right) = 2 \text{ bits}$$

Reduced expected length from 3 bits to 2 bits

1.2 Entropy

Entropy $H(X)$ is the amount of uncertainty about a random variable X.

$$H(X) \quad = \quad -\sum_x p_X(x) \log p_X(x)$$

"log" means $\log_2$ base 2

$0 \log 0 = 0$, by assumption

Intuition for Entropy

Entropy can be described as the smallest number of bits required to describe the outcome of an experiment.

★1

Introduction to Self-Study Quiz (SSQ)

Lecture notes

SSQ 1.1. Let X be a ternary random variable with $\mathsf{X} = \{1, 2, 3\}$ with probability distribution $p_{\mathsf{X}}(x)$:

$$p_{\mathsf{X}}(x) = \begin{cases} \frac{1}{2} & \text{if } x = 1 \\ \frac{1}{4} & \text{if } x = 2 \\ \frac{1}{4} & \text{if } x = 3 \end{cases} \tag{1.7}$$

Calculate $H(\mathsf{X})$. *Go to the course website for solutions and more SSQs.*

Course Web Site

SSQ: Entropy

Let X be a ternary random variable with $\mathcal{X} = \{1, 2, 3\}$ with probability distribution $p_X(x)$:

$$p_X(x) = \begin{cases} \frac{1}{2} & \text{if } x = 1 \\ \frac{1}{4} & \text{if } x = 2 \\ \frac{1}{4} & \text{if } x = 3 \end{cases}$$

Calculate $H(X)$.

Answer:

Properties of Entropy

Proposition 1.1. *Uniform distribution maximizes entropy.* Let X take on values from $\mathcal{X}$. Then, $H(\mathsf{X}) \leq \log|\mathcal{X}|$, with equality if and only if X has a uniform distribution over $\mathcal{X}$.

This is proved on page **??**.

The entropy of a constant is 0. Since a constant is not random, it has no uncertainty and the entropy is 0.

Entropy is non-negative. The lower bound $H(\mathsf{X}) \geq 0$ can be shown by noting that $p_{\mathsf{X}}(x) \leq 1$ means $\log \frac{1}{p_{\mathsf{X}}(x)} \geq 0$.

★2

Binary Random Variable X

Sample space $\mathcal{X} = \{0, 1\}$. For $0 \leq p \leq 1$:

$$\Pr[\mathsf{X} = 0] = 1 - p$$
$$\Pr[\mathsf{X} = 1] = p$$

Samples of a binary random variable with p = 0

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Binary Random Variable X

Sample space $\mathcal{X} = \{0, 1\}$. For $0 \leq p \leq 1$:

$$\Pr[\mathsf{X} = 0] = 1 - p$$
$$\Pr[\mathsf{X} = 1] = p$$

Samples of a binary random variable with p = 0.1

0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0		

Binary Random Variable X

Sample space $\mathcal{X} = \{0, 1\}$. For $0 \leq p \leq 1$:

$$\Pr[\mathsf{X} = 0] = 1 - p$$
$$\Pr[\mathsf{X} = 1] = p$$

Samples of a binary random variable with p = 0.5

1 0 1 0 1 0 0 1 1 1 0 1 0 0 0 1 0 0 1 1 0 0 0 1 0 0 1 1 0 1 0 0 1 1 0 0 1 0 1

Binary Random Variable X

Sample space $\mathcal{X} = \{0, 1\}$. For $0 \leq p \leq 1$:

$$\Pr[\mathsf{X} = 0] = 1 - p$$
$$\Pr[\mathsf{X} = 1] = p$$

Samples of a binary random variable with p = 1.0

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Entropy of Binary Random Variable

In general:

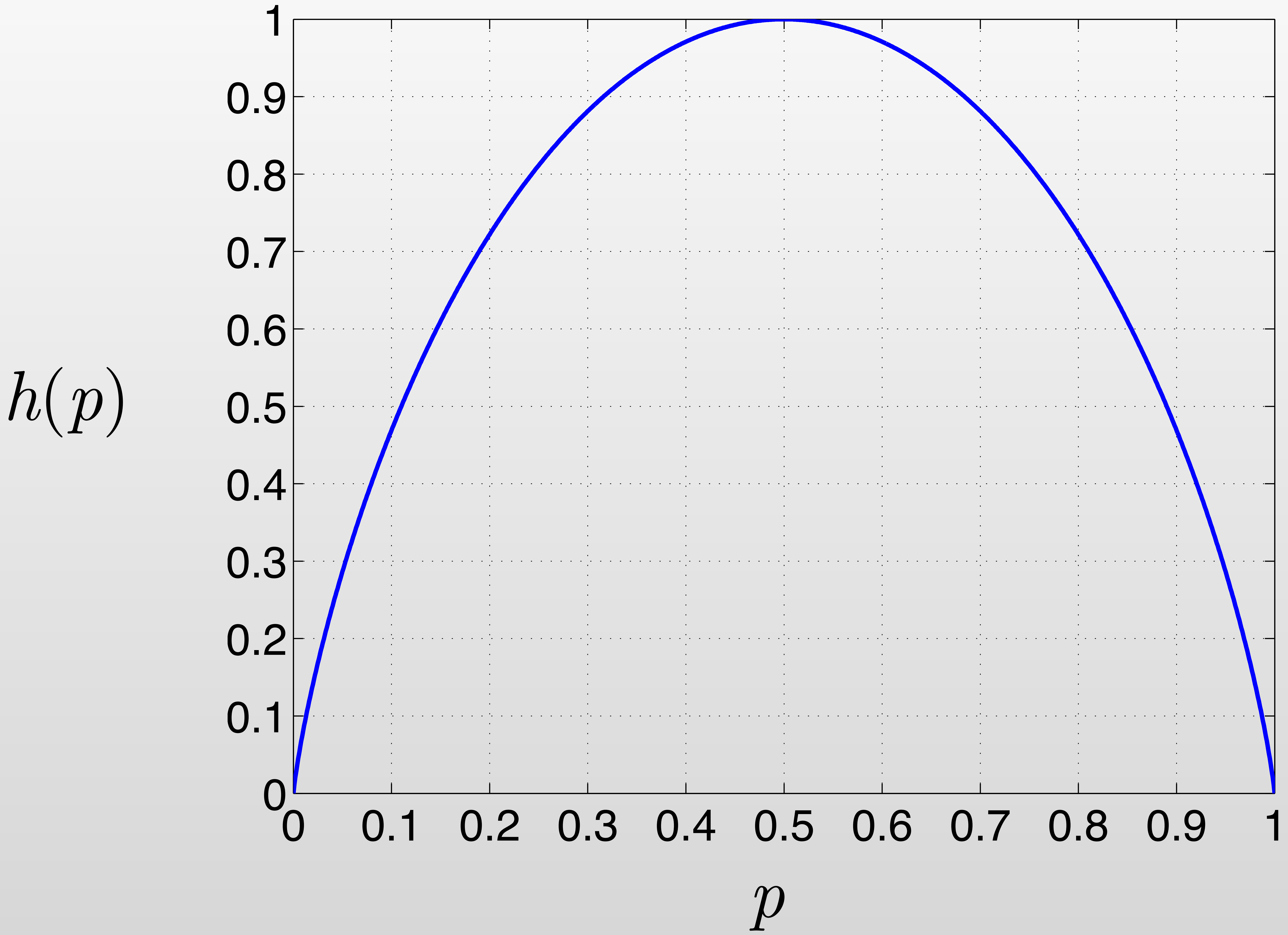
$$H(\mathsf{X}) = -\sum_{x \in \mathcal{X}} p_{\mathsf{X}}(x) \log p_{\mathsf{X}}(x)$$

If X is binary :

$$H(\mathsf{X}) = h(p)$$

$$h(p) = -p \log p - (1-p) \log(1-p)$$

Binary Entropy Function $h(p)$



Joint Entropy

$P_{X,Y}$		Y	
		Bicycle	Train
X	Sunny	1/3	1/3
	Rainy	0	1/3

Joint Entropy

Joint entropy $H(X,Y)$ is
the amount of uncertainty about a
random variable pair X,Y.

$$H(X,Y) = -\sum_x \sum_y p_{X,Y}(x,y) \log p_{X,Y}(x,y)$$

Joint Entropy

$P_{X,Y}$		Y	
		Bicycle	**Train**
X	**Sunny**	**1/3**	**1/3**
	Rainy	**0**	**1/3**

$$
\begin{aligned}
H(X,Y) \\
&= -\sum_x \sum_y p_{X,Y}(x,y) \log p_{X,Y}(x,y) \\
&= -3 \cdot \left(\frac{1}{3} \log \frac{1}{3} \right) - 0 \log 0 \\
&= \log 3 \approx 1.585 \text{ bits}
\end{aligned}
$$

1.3 Conditional Entropy

Given you know X,
conditional entropy $H(\mathsf{Y}|\mathsf{X})$ is
the amount of uncertainty about Y

Two Types of Conditional Entropy

1. Conditional entropy $H(\mathsf{Y}|\mathsf{X}=x)$ given $\mathsf{X}=x$:

$$H(\mathsf{Y}|\mathsf{X}=x) = -\sum_{y\in\mathcal{Y}} p_{\mathsf{Y}|\mathsf{X}}(y|x)\log p_{\mathsf{Y}|\mathsf{X}}(y|x)$$

2. Conditional Entropy $H(\mathsf{Y}|\mathsf{X})$:

$$H(\mathsf{Y}|\mathsf{X}) = \sum_{x\in\mathcal{X}} p_{\mathsf{X}}(x)H(\mathsf{Y}|\mathsf{X}=x) \text{ or}$$

$$H(\mathsf{Y}|\mathsf{X}) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{\mathsf{XY}}(x,y)\log p_{\mathsf{Y}|\mathsf{X}}(y|x)$$

Note that $H(\mathsf{Y}|\mathsf{X})$ is a number, but $H(\mathsf{Y}|\mathsf{X}=x)$ is a function.

Conditional Entropy Example

$p_{Y\|X}(x\|y)$		Y	
		Bicycle	**Train**
X	**Sunny**	1/2	1/2
	Rainy	0	1

And let:

$$p_X(\text{Sunny}) = \frac{2}{3}$$
$$p_X(\text{Rainy}) = \frac{1}{3}$$

★3

1.3.2 Properties of Conditional Entropy

- Conditioning reduces entropy: $H(\mathsf{X}|\mathsf{Y}) \leq H(\mathsf{X})$.
  - If X, Y are independent then $H(\mathsf{X}|\mathsf{Y}) = H(\mathsf{X})$
- Conditional entropy of a function g: $H(g(\mathsf{X})|\mathsf{X}) = 0$
- Entropy conditioned on a bijective function g: $H(\mathsf{X}|g(\mathsf{X})) = 0$. Does not hold if g is not bijective.

★4

Chain Rule for Entropy

- Chain rule for entropy: $H(\mathsf{X},\mathsf{Y}) = H(\mathsf{X}) + H(\mathsf{Y}|\mathsf{X})$.
- Generalized chain rule for entropy:

$$H(\mathsf{X}_1,\mathsf{X}_2,\ldots,\mathsf{X}_n) = \sum_{i=1}^{n} H(\mathsf{X}_i|\mathsf{X}_{i-1}\ldots,\mathsf{X}_1)$$

- Independence bound on entropy:

$$H(\mathsf{X},\mathsf{Y}) \leq H(\mathsf{X}) + H(\mathsf{Y})$$

$$H(\mathsf{X}_1,\mathsf{X}_2,\ldots,\mathsf{X}_n) \leq \sum_{i=1}^{n} H(\mathsf{X}_i)$$

★5

Entropy — What You Should Have Learned

- Information is represented using random variables
- Entropy is a measure of uncertainty
- Entropy is the number of bits to describe the outcome of an experiment
- Single-variable entropy and properties:

  definition, non-negativity, uniform distribution upper bound
- Conditional Entropy and properties:

  conditioning reduces entropy, chain rule, entropy of a function

Class Info

- Next lecture: Monday April 17 at 9:00. Probability theory
- Tutorial hours: Monday, April 17 at 13:00. Probability exercises
- Homework 1 on LMS. Deadline: Monday, April 23 at 18:00

Just for Fun

Mr. Smith has two children who are playing in a garden. We can see that one of the children is a girl, but we cannot see the gender of the other child. What is the probability both children are girls?

Assume Pr(G) = Pr(B) = 0.5

AI-generated art in response to the prompt ``Two children playing in a garden. One of the children's face is turned away from the camera. black and white, line art."
starryai.com