Introduction to Statistics
Session 2

Grzegorz Chrupała

Saarland University

October 13, 2011

Random variables

- Function $X : \Omega \to \mathbb{R}^n$ (typically $n = 1$)
- It may be more convenient to work with real number than directly with events
- Coin toss: $X : \{H, T\} \to \{0, 1\}$
- Sum of two dice throws: $\{1..6\}^2 \to \{2..12\}$
- Probability mass function:

$$\mathrm{p}(x) = P(X = x) = P(A) \text{ where } A = \{\omega \in \Omega : X(\omega) = x\}$$

Expectation

- Expectation is a mean (weighted average) of a random variable

$$E(X) = \sum_x \mathrm{p}(x) \cdot x$$

- Example: rolling a dice:

$$E(X) =$$

Expectation

- Expectation is a mean (weighted average) of a random variable

$$E(X) = \sum_x \mathrm{p}(x) \cdot x$$

- Example: rolling a dice:

$$E(X) = \sum_{x=1}^{6} \mathrm{p}(x)x = \sum_{x=1}^{6} \frac{x}{6} = 3.5$$

- A function $g(X)$ defines new random variable. In this case:

$$E(g(X)) = \sum_x \mathrm{p}(x)g(x)$$

  Example?
- Also for two random variables:

$$E(X + Y) = E(X) + E(Y)$$

  and if independent

$$E(XY) = E(X)E(Y)$$

Variance

- Variance measures how much values of a random variable vary

$$Var(X) = E[(X - E(X))^2]$$

- Standard deviation σ is the square root of the variance
- What is the variance of a random variable describing a single throw of a dice?

Entropy

- Entropy is a measure of degree of uncertainty.
- The most important concept in information theory
- Entropy is a property of a random variable X distributed according the pmf p

$$H(X) = H(\mathrm{p}) = E(-\log_2(x)) = -\sum_x \mathrm{p}(x)\log_2(\mathrm{p}(x))$$

- For $\log_2(x)$ units are bits, for $\ln(x)$, nats

Entropy as amount of information

- You can think of entropy as measuring the cost of transmitting information about the result of an experiment
- Fair coin toss:

Entropy as amount of information

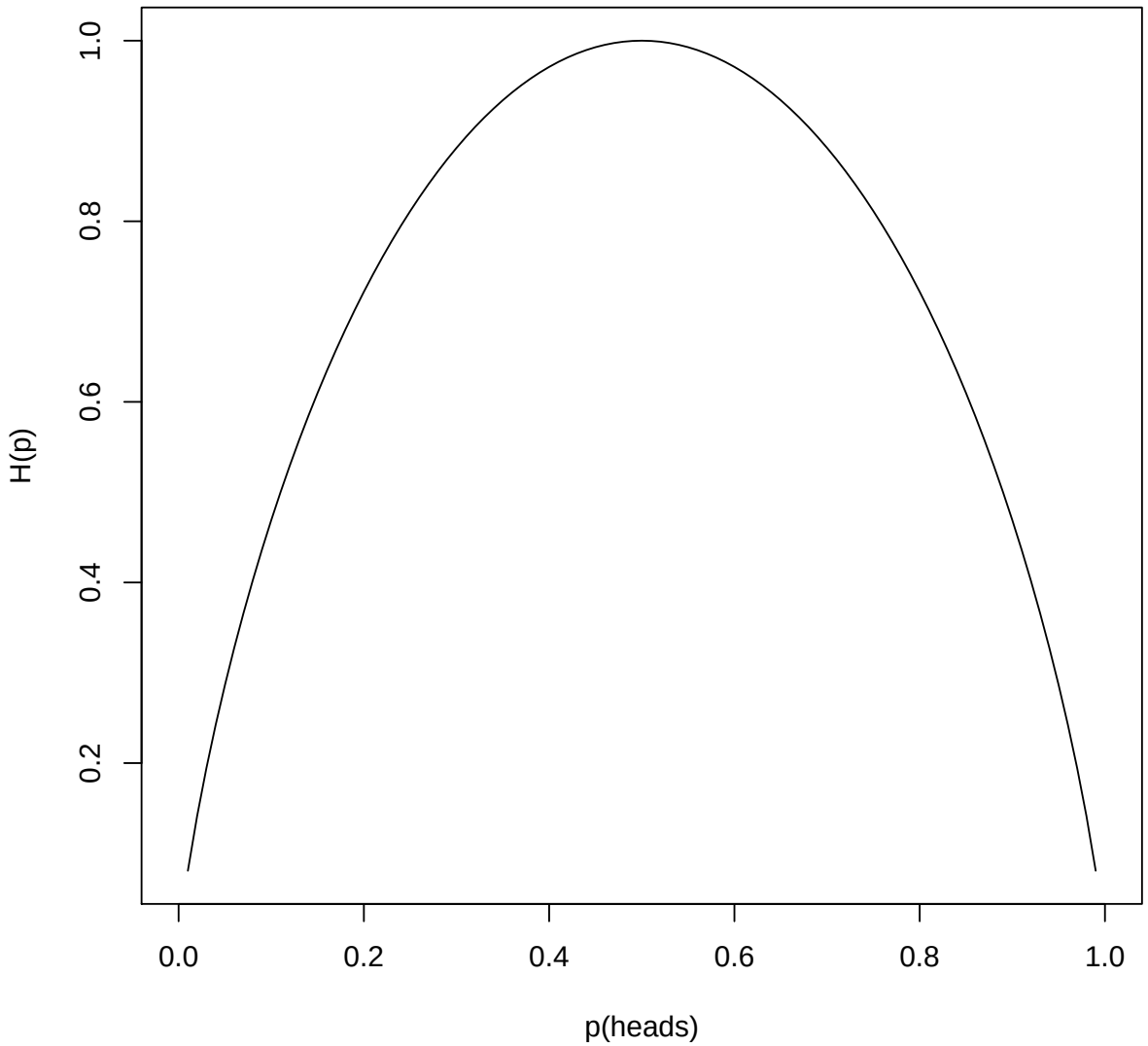
- You can think of entropy as measuring the cost of transmitting information about the result of an experiment
- Fair coin toss:

$$H(X) = -\sum_{x=0}^{1} \mathrm{p}(x) \log_2(\mathrm{p}(x)) \tag{1}$$

$$= \frac{1}{2}[-\log_2\left(\frac{1}{2}\right) - \log_2\left(\frac{1}{2}\right)] \tag{2}$$

$$= \frac{1}{2} \cdot 2 \tag{3}$$

Entropy of an unfair coin



Properties of entropy

- $H(p) \geq 0$
- When is it $H(p) = 0$?
- The highest entropy corresponds to the most uniform distribution

Entropy: joint and conditional

- For two variables X and Y , the amount of information needed to specify values of both

$$H(X, Y) = -\sum_x \sum_y \mathrm{p}(x, y) \log_2(\mathrm{p}(x, y))$$

- Conditional entropy: if we know the value of X, how much does to cost to transmit the value of Y?

$$H(Y|X) = \sum_x \mathrm{p}(x) H(Y|X = x) \tag{4}$$

$$= \sum_x \mathrm{p}(x) \left[-\sum_y \mathrm{p}(y|x) \log(\mathrm{p}(y|x)) \right] \tag{5}$$

$$= -\sum_x \sum_y \mathrm{p}(y|x)\mathrm{p}(x) \log(\mathrm{p}(y|x)) \tag{6}$$

$$= -\sum_{x,y} \mathrm{p}(x, y) \log(\mathrm{p}(y|x)) \tag{7}$$

Conditional entropy: example

- X: number of heads in two tosses of a fair coin
- Y: is at least one of two tosses of a fair coin a heads?

	X	Y
HH	2	1
HT	1	1
TT	0	0
TH	1	1

What is $H(X)$? What is $H(X|Y)$?

Chain rule for entropy

$$H(X, Y) = H(X|Y) + H(Y)$$
$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1, \ldots, X_{n-1})$$

Mutual information

- From the chain rule we have

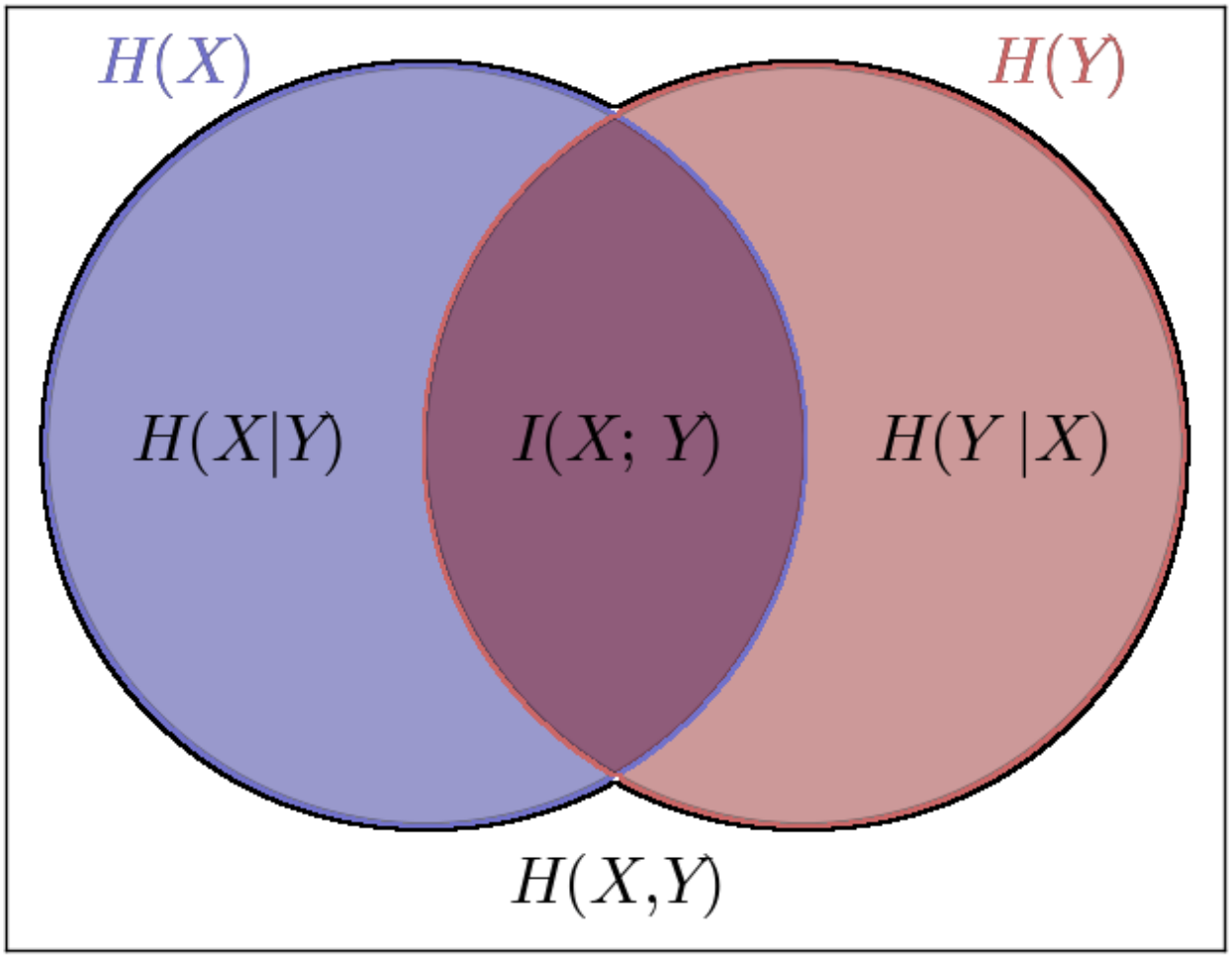
$$H(X) + H(Y|X) = H(Y) + H(X|Y)$$

- Therefore

$$H(X) - H(X|Y) = H(Y) - H(Y|X)$$

- This difference is known as Mutual information $I(X;Y)$

Joint and conditional entropy and mutual information



Mutual information

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

- What is $H(X|X)$?
- What is $I(X;X)$?

Kullback Leibler divergence

- A measure of the difference between two probability mass functions p and q is Kullback Leibler divergence (relative entropy)

$$D(p||q) = \sum_x \mathrm{p}(x) \log \left(\frac{\mathrm{p}(x)}{\mathrm{q}(x)} \right)$$

- Can be interpreted as an average number of bits wasted by encoding events distributed according to p with a code based on q
- We can define mutual information in terms of KL divergence:

$$I(X;Y) = D(\mathrm{p}(x,y)||\mathrm{p}(x)\mathrm{p}(y))$$