CS3236 Introduction to Information theory

Lecture 2: Probability and Entropy

Course given by Pierre Senellart
Material by Stephanie Wehner, with additions by P. Senellart

Let's Recap…

Class survival guide

- Lecture on Monday 2-4pm
- Tutorial on Wed 4-5pm and Thu 2-3pm, starting Week 3
  - Choose one session, no need to go to both
  - Discussion, QA, help with homework exercises, projects
- Recess week September 20-28
- No physical lecture/tutorial on Sep 8-12, e-Learning material
- No lecture on October 6 (Hari Raya Haji)
- No tutorial on week 10, October 22-23 (Deepavali)
- Available to meet for answering individual questions. Use http://www.doodle.com/psenellart to schedule a meeting. Office ICube #03-09.
- Grading
  - 50% Exam (Friday November 28)
  - 50% Continuous Assessment
    - 10% Mid-Term exam (October 13)
    - 20% Homework (assigned each week, due next Monday)
    - 20% Small project (handed out September 1, due November 14)
- If you have any doubt about the schedule, check IVLE

Today: Probability 101

Probability 101

- When talking about probabilities we must say
  - What is the set of possible events: The alphabet A
  - What is the probability that an event $a \in A$ occurs: $P(a)$
    - In the literature, this is also denoted as
      - $P(x = a)$ where x is a random variable
      - $P_X(x)$ where X is a random variable and $x \in A$
      - $P(x)$

- Example: 6 sided die
  - Alphabet $A = \{1,2,3,4,5,6\}$
  - Probability $P(1) = P(2) = \ldots . = P(6) = \frac{1}{6}$

Ensemble: summarizing the parameters

- Ensemble: $E = (x, A, P)$
  - x - what we'll call the random variable
  - A - the alphabet
  - P – the (list of) probabilities

JOINT PROBABILITY DISTRIBUTION

- What is the probability that two (or more) events occur together?

- Event (x, y) occurs with probability $P(x, y)$
- Takes values $(x, y) \in A \times B$
- Joint ensemble $E = \{(x, y), A \times B, P\}$

- Example:
  - A = {curly hair, straight hair}
  - B = {Asian, European}
  - P(curly, Asian) – probability of having curly hair AND being Asian ☺

Marginal distribution

- What is the probability of x alone?
- Marginal distribution $P(x) = \sum_y P(x, y)$
- Example:
  - P(curly) = P(curly,Asian) + P(curly, European)

New from old: functions of events

- **Example:**
  - x: outcome of a fair 6 sided die
  - y: outcome of a fair 4 sided die
  - z = x + y: sum of outcomes dies
- What is the probability of $P(z)$?
- $P(z) = \sum_{x,y,s.t.z=x+y} P(x,y)$
- In our example

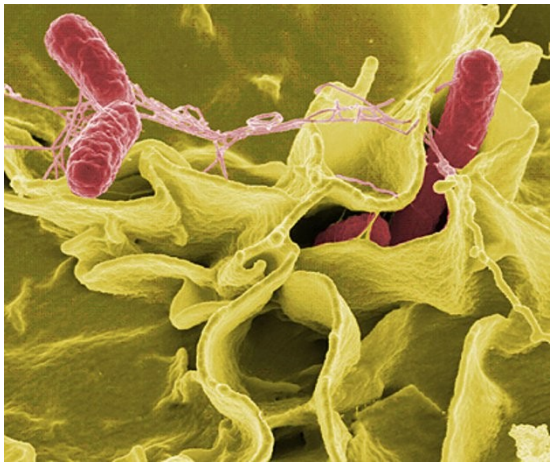
Conditional distributions

- Probability of observing event y, given event x has happened $P(y|x)$
- Product rule for the joint distribution
  - $P(x,y) = P(y|x)P(x)$
  - $P(x,y) = P(x|y)P(y)$
- Conditional probability $P(x|y) = \frac{P(x,y)}{P(y)}$
- Bayes theorem $P(x|y) = \frac{P(y|x)P(x)}{P(y)}$

Power of Bayes' theorem

- Guess my die ☺
  - 1 fair 6 sided die $P(x) = \frac{1}{6}$
  - 1 unfair die $P(1) = P(6) = \frac{1}{2}$

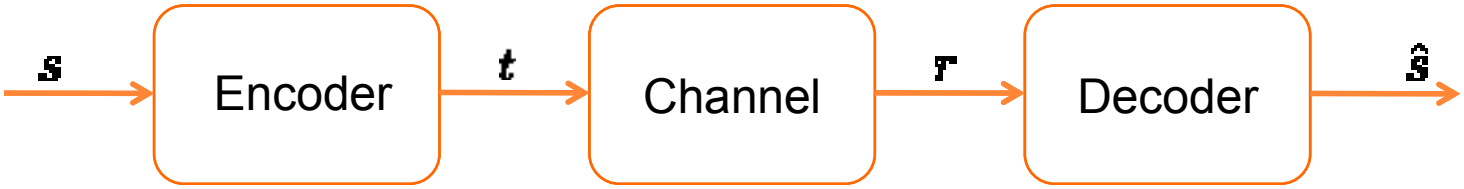
Significance of the prior



Several years after the first Mars explorations, it was discovered that Astronauts were exposed to a deadly space bug!

Luckily there is a test, which is 95% accurate.
Given a positive result, what is the probability an astronaut has the disease?

Bayes applied to coding



- Forward probability

- Inverse probability

The power of large numbers

Independent events

- Two events are independent iff for all x and y
  $P(x, y) = P(x)P(y)$
- Such a distribution is called a *product distribution*
- Our previous example was clearly not a product distribution ☺
- Example of a product distribution:
  - tossing a 6 sided die, and 4 sided die independently

AYE, PIE, CPE ….. IID!

- Imagine we toss a coin n times to obtain a string $x = x_1 x_2 \ldots x_n$
- In each toss we have
  - Probability of 0: $P(0)$
  - Probability of 1: $P(1)$

- Probability of the entire string is a product distribution $P_X(x) = P_{X_1}(x_1) \ldots P_{X_n}(x_n)$
  - Independently distributed
- In fact, the probabilities are the same each time
  - Identically distributed

- **IID: Independently and identically distributed**

How many 1's?

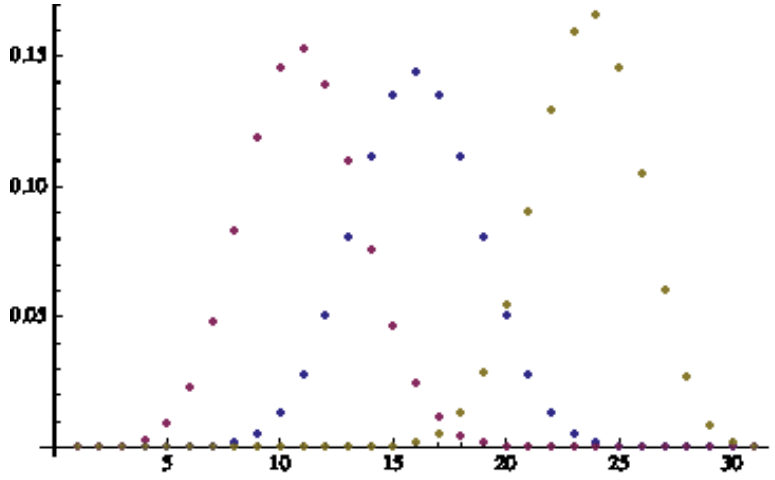
The binomial distribution

- Ways to pick k out of n

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- Probability of observing k 1's in a string of length n when the probability of 1 is p

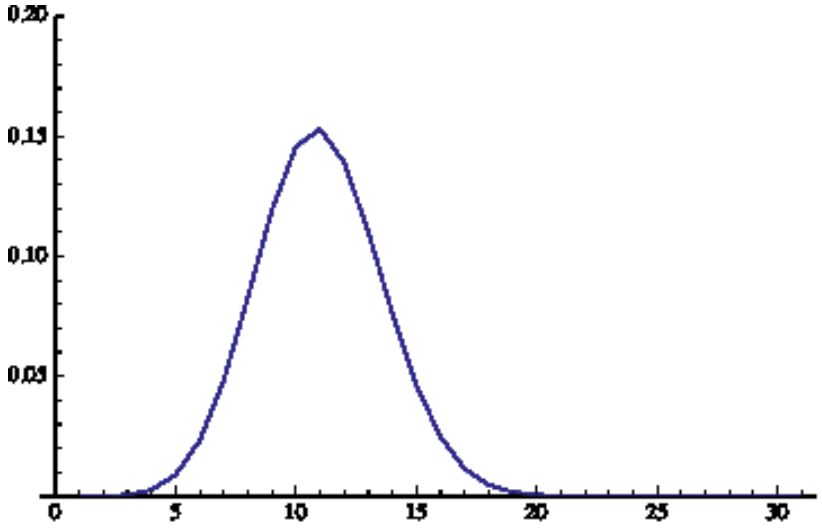
$$P(\#1's = k) = \binom{n}{k} p^k (1-p)^{n-k}$$



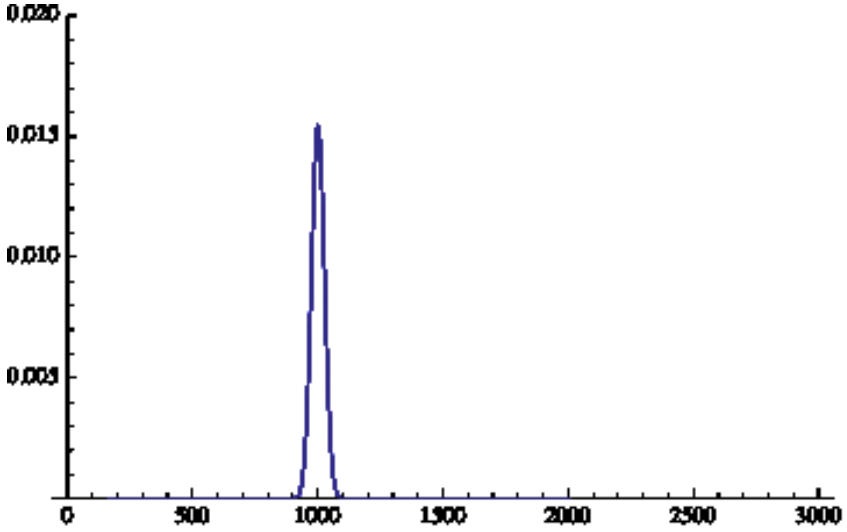
Examples: n=30
p = 1/3,1/2,3/4

Law of large numbers - intuition

- p=1/3, n = 30



- n = 3000



Expectation and variance

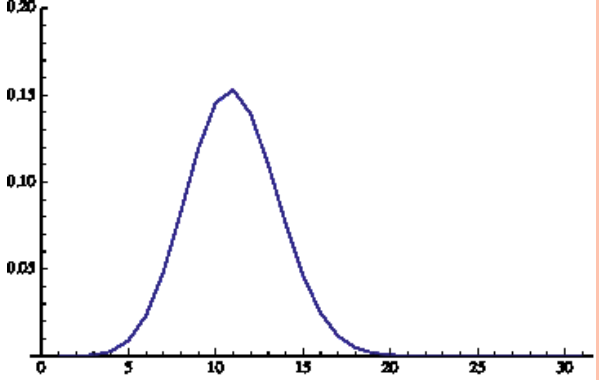
- Expectation of a random variable t

$$\mathbb{E}(t) = \bar{t} = \sum_{a \in A} P(t = a)a$$

- Variance

$$\mathrm{Var}(t) = \sigma_t^2 = \mathbb{E}((t - \bar{t})^2)$$

Weak law of large numbers



- Define a new random variable for the average

$$x = \frac{1}{N}\sum_{n=1}^{N} h_n$$

- Weak law of large numbers

$$P((x-\bar{h})^2 \geq \alpha) \leq \frac{\sigma_h^2}{\alpha N}$$

Implications

$$P((x-\bar{h})^2 \geq \alpha) \leq \frac{\sigma_h^2}{\alpha N} \qquad x = \frac{1}{N}\sum_{n=1}^{N} h_n$$

- As *N* goes to infinity, the sample average becomes as close to the expectation as required
- This is true even if the random variable has high variance (high variance will just require more samples)
- Distance to the mean is inversely proportional to the square root of the sample size

Proof - Chebychev to the rescue: First inequality

For all random variables *t* and real parameters α with $t \geq 0, \alpha > 0$

$$P(t \geq \alpha) \leq \frac{\bar{t}}{\alpha}$$

Second chebychev inequality

$$P(t \geq \alpha) \leq \frac{\bar{t}}{\alpha}$$

For all random variables x and positive real values α

$$P((x - \bar{x})^2 \geq \alpha) \leq \frac{\sigma_x^2}{\alpha}$$

Proof - Weak law of large numbers

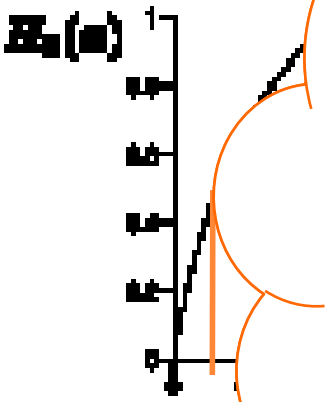
$$x = \frac{1}{N}\sum_{n=1}^{N} h_n$$

Want $$P((x-\bar{h})^2 \geq \alpha) \leq \frac{\sigma_h^2}{\alpha N}$$

Have 2nd Chebychev $$P((x-\bar{x})^2 \geq \alpha) \leq \frac{\sigma_x^2}{\alpha}$$

Shannon entropy

$C = 1 - H($



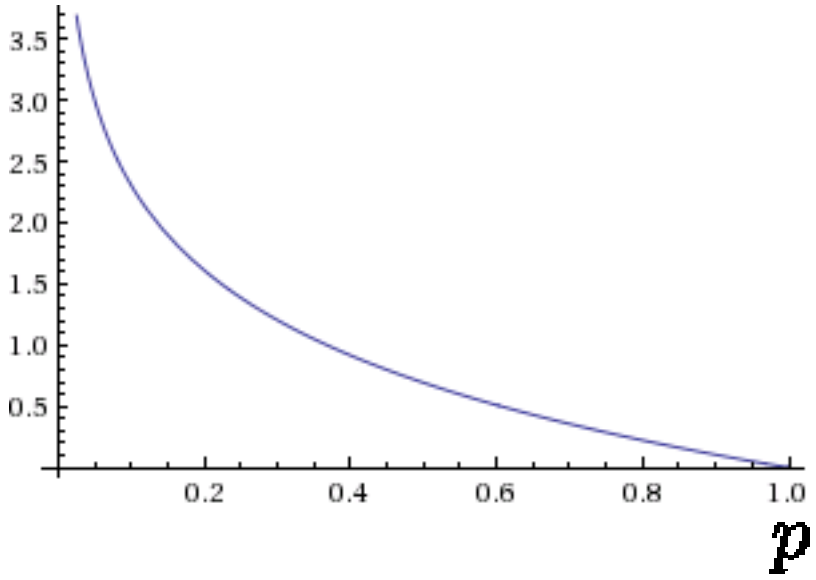
Why should this be a useful measure of information?

- Binary e py $H($ $= -f \log_2 f - (1-f)\log_2(1-f)$
- Shannon entropy $H(p) = -\sum_x p_x \log_2 p_x$

Shannon information content

- Information content of an event x $\quad P(x) > 0$

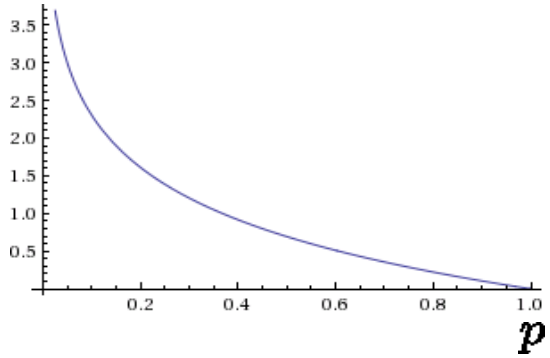
$$h(x) = \log_2 \frac{1}{P(x)}$$



Intuitively, measures how much we learn by observing *x*

Shannon's information content

- The more unlikely the outcome, the higher its information content



$$h(x) = \log_2 \frac{1}{P(x)}$$

- Information content is measured in bits (h(x) bits)

Why bits?

- A game of yes and no questions
  - I choose a number between 0 and 63 $P(x) = \frac{1}{64}$
  - You can ask yes/no questions to find out my number
- Our strategy gives a binary encoding of the number

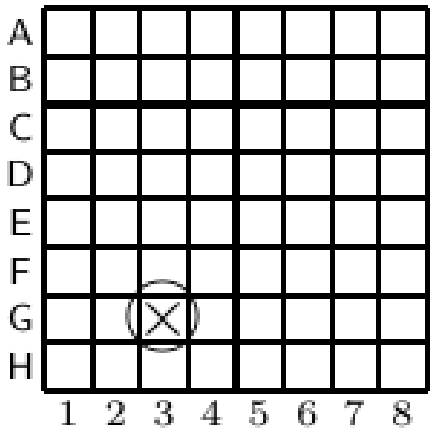
$$0 \leq x \leq 64$$

- Length of this encoding is determined by the information content

$$h(x) = \log \frac{1}{64} = \log 2^6 = 6 \text{ bits}$$

A FURTHER EXAMPLE

- Let's put the numbers in a square (64 squares)



Outcome x = n (no submarine)
x = y (submarine!)

What's the information content of x=n and x=y at the start?

If we hit the sub on the first try we gain bits of information:

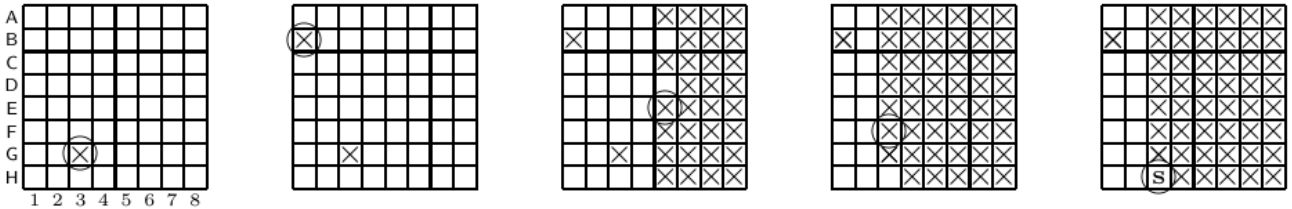
$$h(x=y) = -\log \frac{1}{64} = 6 \text{ bits}$$

If we do not hit the sub on the first try we gain bits of information:

$$h(x=n) = -\log \frac{63}{64} = 0.023 \text{ bits}$$

Information content of "shots"

- Suppose we hit the sub on the 5th attempt

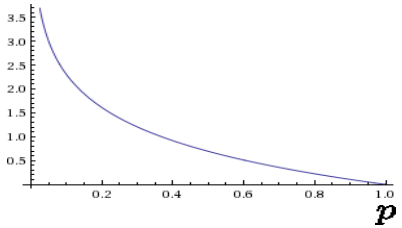


move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

- Regardless of when we hit the sub, the overall information content (added up) is 6 bits once we know where it is.

Intuitive interpretation

- Unlikely outcomes give more information
  - It's unlikely we hit the sub, but if we do we know the exact position



$$h(x) = \log_2 \frac{1}{P(x)}$$

- Intuitively, quantifies how "many" of the bits we learn in one question/shot on the board

Shannon entropy

- Average information content

$$H(X) = \sum_x P(x)h(x) = \sum_x P(x) \log \frac{1}{P(x)} = -\sum_x P(x) \log P(x)$$

- Basic properties

$$\log |\mathcal{A}_X| \geq H(X) \geq 0$$

Alphabet

Entropy of independent R.V.s

- Consider two independent random variables X and Y

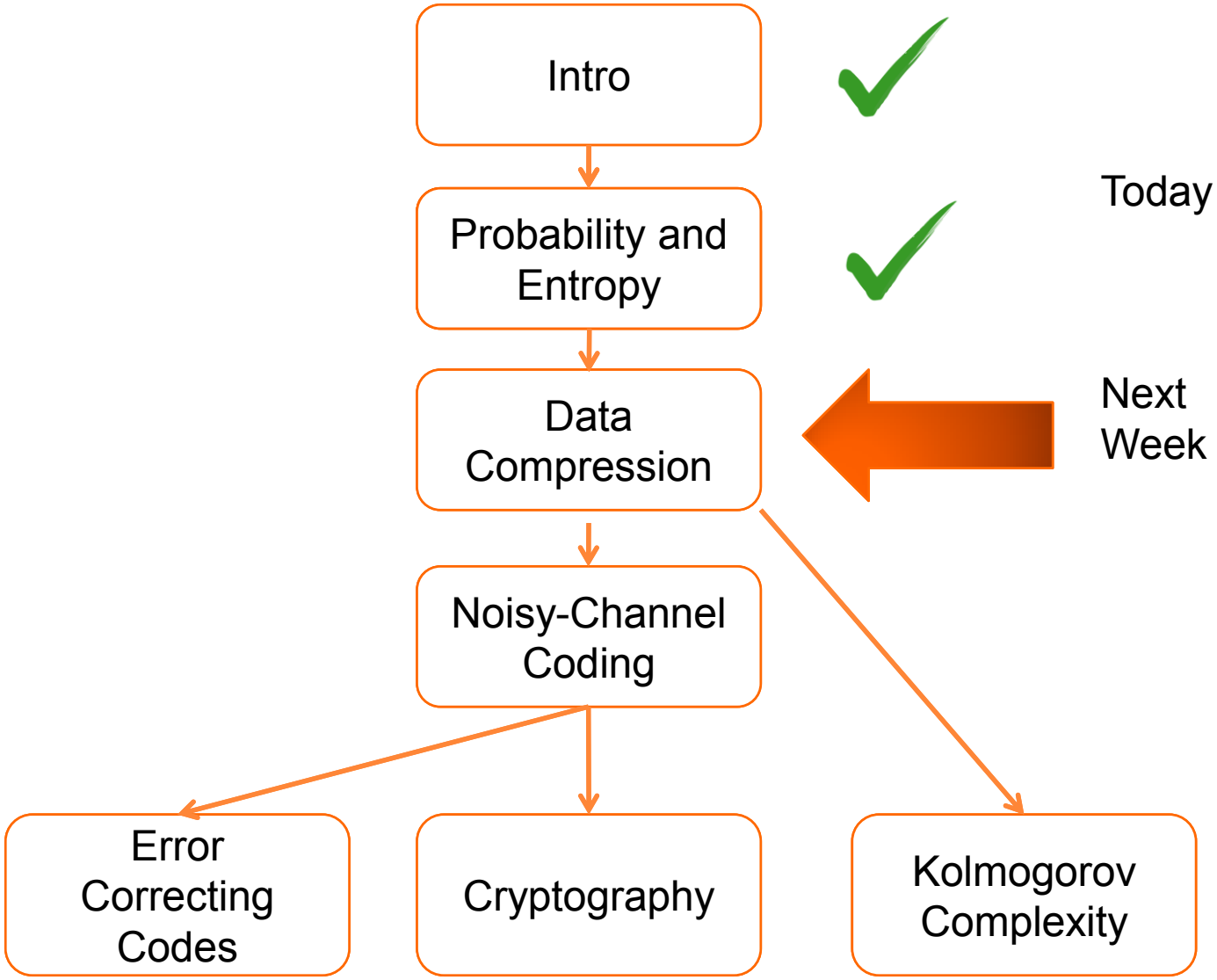
$$P_{XY}(xy) = P_X(x)P_Y(y)$$

- Entropy of X and Y

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

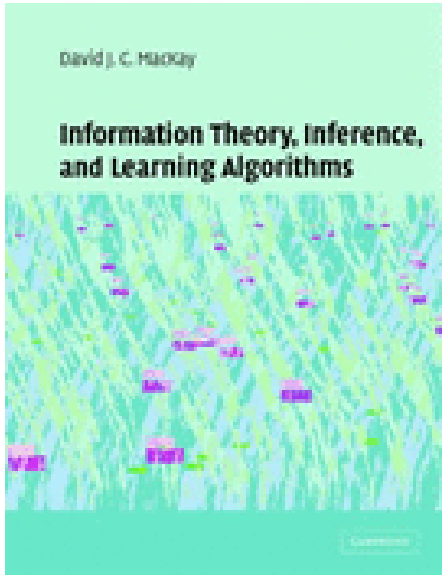
$$H(X,Y) = -\sum_{xy} P_{XY}(xy) \log P_{XY}(xy)$$

Where do we go from here?



Reading for this lecture

- Chapters 2, 4.1, and 4.5 in the book



Information Theory, Inference and Learning Algorithms
by David J. C. MacKay
Cambridge University Press, 2003

- Homework due by the beginning of next lecture, hand out in person at next lecture or upload to IVLE Workbin