PMR Introduction

Probabilistic Modelling and Reasoning

Amos Storkey

School of Informatics, University of Edinburgh

Outline

1 Welcome

2 Probability Refresher

3 Some Distributions

Welcome

to the Probabilistic Modelling and Reasoning Course

What is this about?

- Probabilistic Models for Unsupervised Machine Learning

One course among many:

- Introductory Applied Machine Learning
- **Machine Learning and Pattern Recognition**
- Information Theory
- Reinforcement Learning
- Data Mining and Exploration
- Neural Information Processing

Just a few courses that are relevant to machine learners on the

MSc (Master of Science) in Informatics and MSc in Artificial Intelligence at the School of Informatics, University of Edinburgh.

Why?

Exciting area of endeavour

- Coherent.
- Interesting (and some unsolved...) problems.
- Now ubiquitous...
- Relevant to data analytics, business analytics, financial modelling, medical systems, signal processing, condition monitoring, brain science, the scientific method, image analysis and computer vision, language modelling, speech modelling, handwriting recognition, risk management, medical imaging, web analytics, recommender engines, computer games engines, geoinformational systems, intelligent management, operational research, etc. etc. etc.
- In great demand.

What is the Point of Studying this Course?

What should you be able to do after this course?

- Understand the foundations of Bayesian statistics.
- Understand why probabilistic models are the appropriate models for reasoning with uncertainty.
- Know how to build structured probabilistic models.
- Know how to learn and do inference in probabilistic models.
- Know how to apply probabilistic modelling methods in many domains.

How to succeed

- We will use David Barber's book as our 'Hitchhikers guide to Probabilistic Modelling'.

Don't Panic

- Hard course. Decide now whether to do it. Then don't panic.
- Work together...
- Practice... Use tutorials. Hand in tutorial questions for marking.
- Ask me questions. Yes. Even in the lectures. Use the nota bene.
- Keep on top. Don't let it slip.

- Do this and I promise you will learn more on this course than any other course you have done or are doing!

How to succeed

- Working together is critical. Working on your own is critical for effective working together.
  - Tutorials are vital. They are the only way to survive the course!
  - Form pairs or small groups within your tutorial.
  - Arrange a meeting a day or two prior to the tutorial to go through the tutorial questions together.
  - Try the questions yourself first. Then get together to discuss. Work out where you are uncertain. Work out what you don't follow. Prepare questions for the tutorial.
  - Use the online nota bene site: Ask questions. Answer others questions. Minimize the time being stuck and maximize understanding.
  - Don't try to rote learn this course. Focus on understanding.

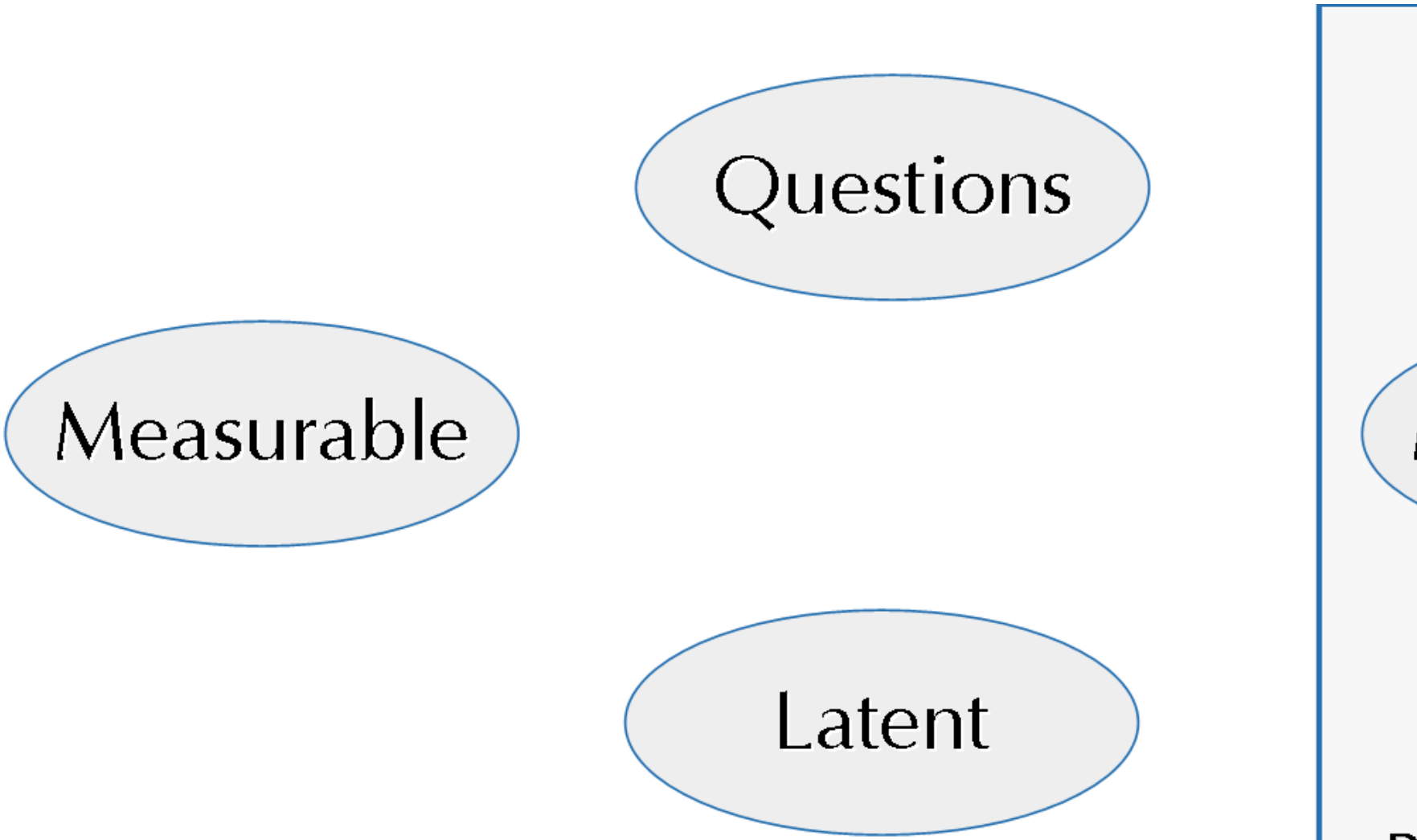
How to succeed

- Be bold
  - You are not supposed to understand everything immediately. Its hard! Its supposed to be.
  - Ask, Ask, Ask. Ask yourself. Ask other people. Ask me. Ask in lecturers.
  - Don't rely on lectures. You need more than lectures.
  - Make lectures better! Prepare beforehand. Ask during.
  - Again use nota bene when you get stuck. Ask a Q. Its okay to just say "I don't get it. Can you put it a different way."
  - Work through example questions. Lots of them. If you can't do a question you are missing something. Fix it.

Stop point

- Quick break.

Probabilistic Machine Learning



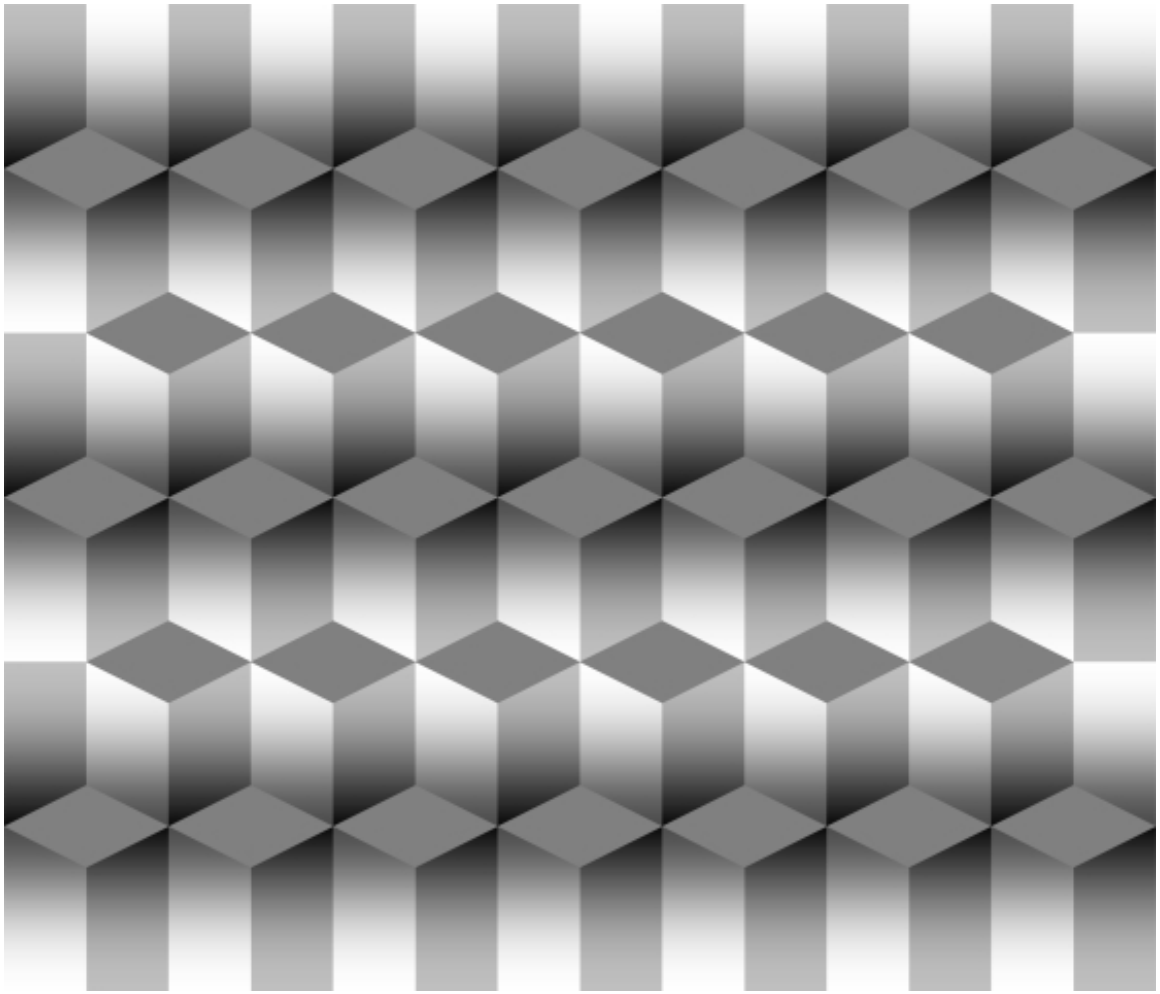
Thinking about Data

- Probabilistic Modelling does not give us something for nothing.
- Prior beliefs and model + data $\rightarrow$ posterior beliefs.
- Can do nothing without some a priori model - no connection between data and question.
- A priori model sometimes called the inductive bias.

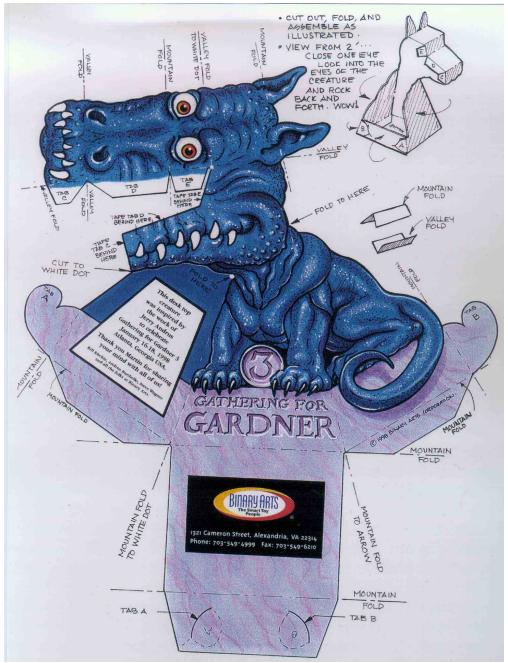
Illusions

- Logvinenko illusion
- Inverted mask illusion

Logvinenko Illusion



Inverted Mask Illusion



Example: Single Image Super-Resolution

Example

- Given lots of example images...
- Learn how to fill in a high resolution image given a single low resolution one.
- State of the art is hard to beat, and appears to be in widespread use.
- However state of art technology appears to be restricted to a small region in Los Angeles area.
- Technology used unknown but seems particularly pertinent at discovering reflections of people in windows, and is usually accompanied by series of short beeps in quick procession.

Example: Single Image Super-Resolution

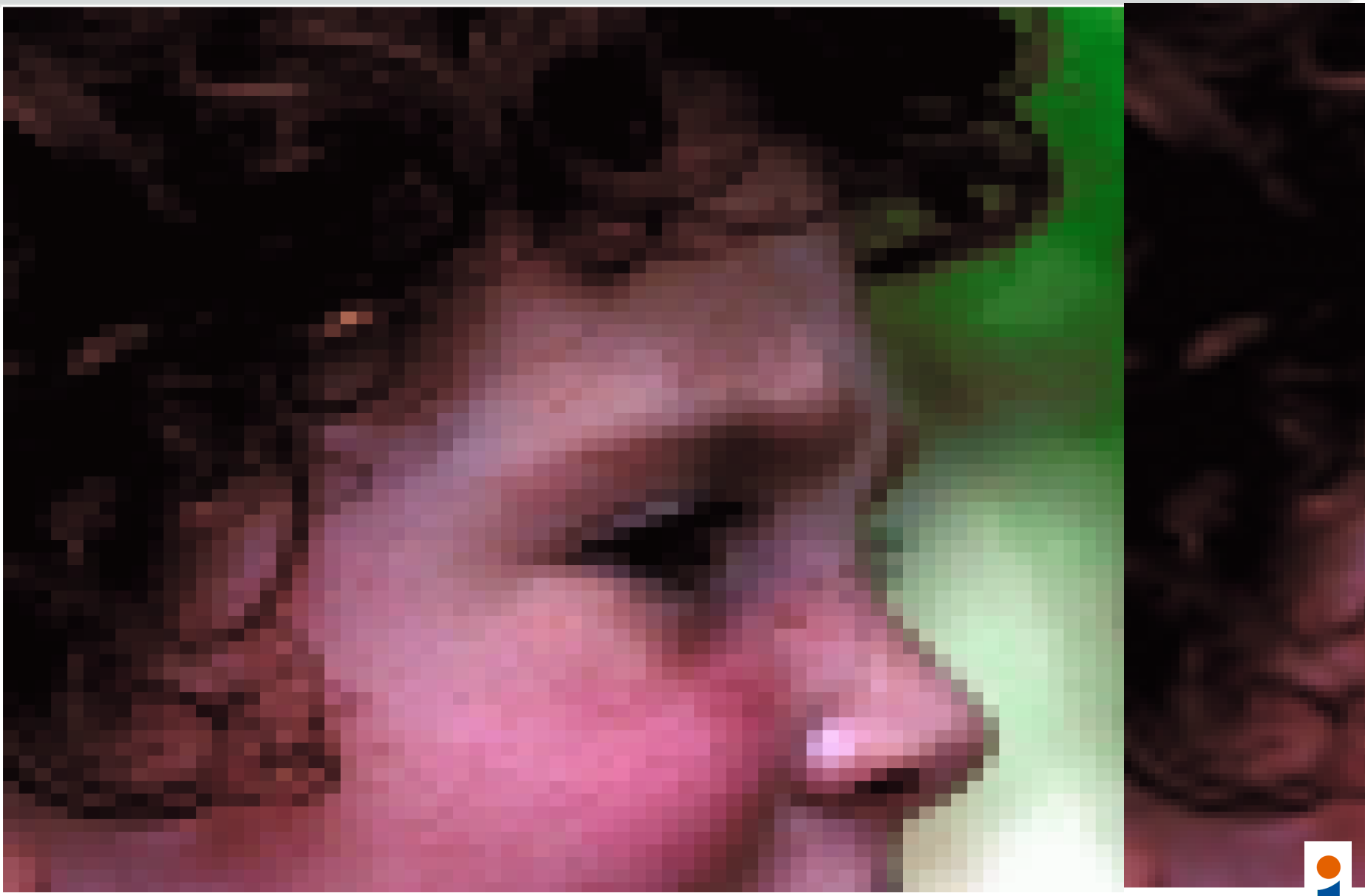
Ancknowledgements: Duncan Robson

Example: Single Image Super-Resolution

Example

- Rest of us stuck with more standard approaches.
- One approach:
  - Build model for image patches. Build model for map from high to low dimensional space.
  - Refine model to fit example data.
  - Invert model to give high dimensional region for low dimensional data.

Example: Single Image Super-Resolution



Reading

- Please read David Barber's book Chapter 1: Probability Refresher.
- Try some of the exercises at the end.

Probability Introduction

- Event Space, Sigma Field, Probability measure.
- Prior beliefs and model + data $\rightarrow$ posterior beliefs.
- Can do nothing without some a priori model - no connection between data and question.
- A priori model sometimes called the inductive bias.

Event Space

- The set of all possible future states of affairs is called the event space or sample space and is denoted by Ω.
- A σ-field $\mathfrak{F}$ is a collection of subsets of the event space which includes the empty set $\emptyset$, and which is closed under countable unions and set complements.
- Intuitively... The event space is all the possibilities as to what could happen (but only one will). σ-fields are all the bunches of possibilities that we might be collectively interested in.
- σ-fields are what we define probabilities on.
- A probability measure maps each element of a σ-field to a number between 0 and 1 (ensuring consistency).

Random Variables

- Random variables assign values to events in a way that is consistent with the underlying σ-field. ($\{x \leq x\} \in \mathfrak{F}$) – the bunch of possibilities we might be interested in.
- We will almost exclusively work with random variables. We implicitly assume a standard underlying event space and σ-field for each variable we use.
- $P(x \leq x)$ is then the probability that random variable x takes a value less than or equal to x.
- Don't worry. Be happy.

Rules of Probability

- Axioms for events: $0 \leq P(A)$ for all events $A \in \Omega$. $P(\Omega) = 1$. For countable disjoint $A_1, A_2, \ldots$ we have $P(A_1 \cup A_2 \cup A_3 \ldots) = \sum_{i=1} P(A_i)$.
- Consequences
  - Normalization: $\sum_y P(\mathrm{y} = y) = 1$. ($\sum \rightarrow \int$ for densities).
  - Joint distributions: work in the product space: $P(\mathrm{x} < x, \mathrm{y} < y)$.
  - Marginalisation: $\sum_x P(x, y) = P(y)$
  - Conditioning: $P(x|y) = P(x, y)/P(y)$.
  - Chain rule: Repeated conditioning.
  - Factorizing: $P(x, y|z) = P(x|z)P(y|z)$ iff x and x are independent.

Distribution Functions
of Random Variables

- *The Cumulative Distribution Function* $F(x)$ is $F(x) = P(\mathrm{x} \leq x)$
- *The Probability Mass Function* for a discrete variable is $P(x) = P(\mathrm{x} = x)$
- *The Probability Density Function* for a continuous variable is the function $P(x)$ such that

$$P(\mathrm{x} \leq x) = \int_{-\infty}^{x} P(\mathrm{x} = u)du$$

- Think in terms of probability mass per unit length.
- We will use the term *Probability Distribution* to informally refer to either of the last two: it will be obvious which from the context.

Distribution Functions
of Random Variables

- We write $P(\mathrm{x} = x)$ for the probability distribution (density or mass) that random variable x takes value x.

Sloppier notations for brevity

$P(\mathrm{x})$ Sometimes we conflate the notation for the random variable and the value it takes.

$P(x)$ Sometimes we implicitly assume the underlying random variable.

$P(\mathrm{x} = x)$ If there is any doubt we will specify the full form.

Notation

- See notation sheet. Notation follows Barber mostly. However I will use capital P for probability, and overload it.
- Simply put: work out the random variable that P has as an argument, and P makes sense.

Interpretation

- Probability Theory is a mathematical abstraction. To use it (i.e. apply it) you need an interpretation of how “real world” concepts relate to the mathematics.
- Probability as degree of belief. $P = 1$ is certainty. $P = 0$ is impossibility.
- See Cox’s axioms

Cox's Axioms

- Assume measure of plausibility f, and map of negation $c(.)$ If
  1. Plausibility of proposition implies plausibility of negation: $c(c(f)) = f$
  2. Plausibility of A and B: We can write $f(A \text{ and } B) = g(A, B|A)$. Then g is associative.
  3. Order independence: the plausibility given information is independent of the order that information arrives in.
- Cox's theorem: plausibility satisfying the above is isomorphic to probability.
- See also Dutch Book arguments of Ramsey and de Finetti.

Stop point

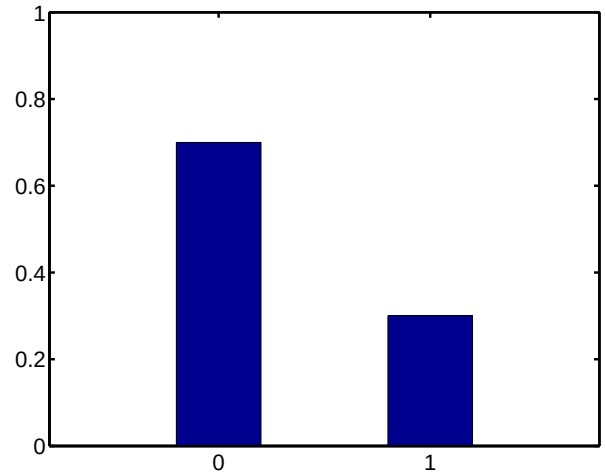
- Quick break.

Distributions

- Briefly introduce some useful distributions.
- Note: Probability mass functions or probability density functions.

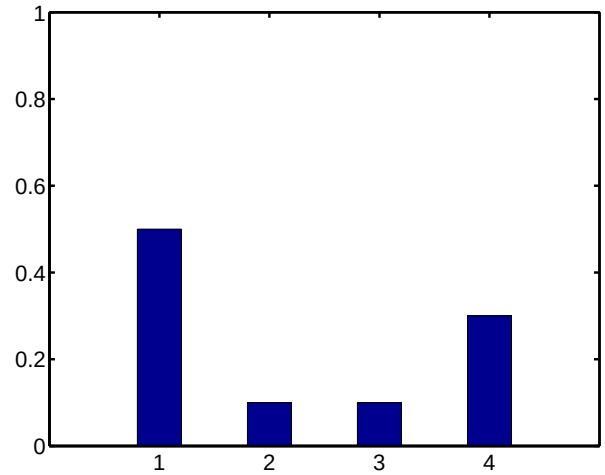
Bernoulli Distribution

- x is a random variable that either takes the value 0 or the value 1.
- Let $P(\mathrm{x} = 1|p) = p$ and so $P(\mathrm{x} = 0|p) = 1 - p$.
- Then x has a Bernoulli distribution.



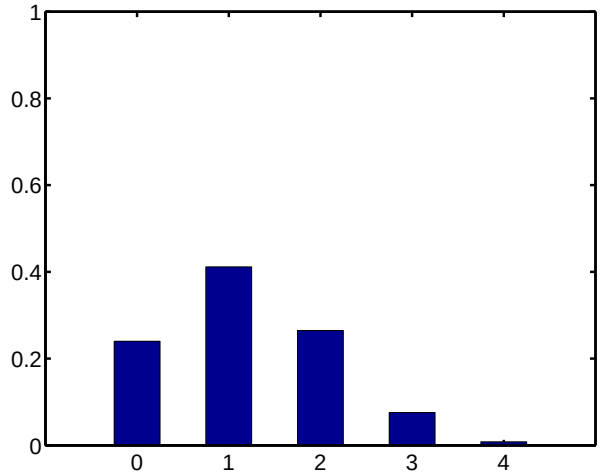
Multivariate Distribution

- x is a random variable that takes one of the values $1, 2, \ldots, M$.
- Let $P(\mathrm{x} = i|\mathbf{p}) = p_i$, with $\sum_{i=1}^{m} p_i = 1$.
- Then x has a multivariate distribution.



Binomial Distribution

- The binomial distribution is obtained from the total number of 1's in n independent Bernoulli trials.
- x is a random variable that takes one of the values $0, 1, 2, \ldots, n$.
- Let $P(\mathrm{x} = r|p) = \binom{n}{r} p^r (1-p)^{(n-r)}$.
- Then x is binomially distributed.



Multinomial Distribution

- The multinomial distribution is obtained from the total count for each outcome in n independent multivariate trials with m possible outcomes.
- $\mathbf{x}$ is a random vector of length m taking values $\mathbf{x}$ with $x_i \in \mathbb{Z}^+$ (non-negative integers) and $\sum_{i=1}^{m} x_i = n$.
- Let

$$P(\mathbf{x} = \mathbf{x}|\mathbf{p}) = \frac{n!}{x_1! \dots x_m!} p_1^{x_1} \dots p_m^{x_m}$$

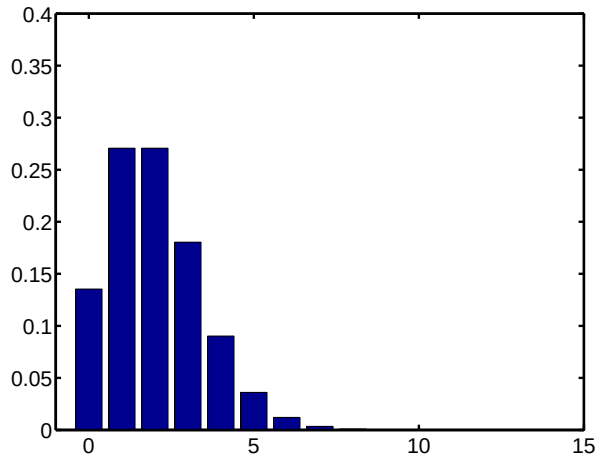
- Then $\mathbf{x}$ is multinomially distributed.

Poisson Distribution

- The Poisson distribution is obtained from binomial distribution in the limit $n \to \infty$ with $p/n = \lambda$.
- x is a random variable taking non-negative integer values $0, 1, 2, \ldots$.
- Let

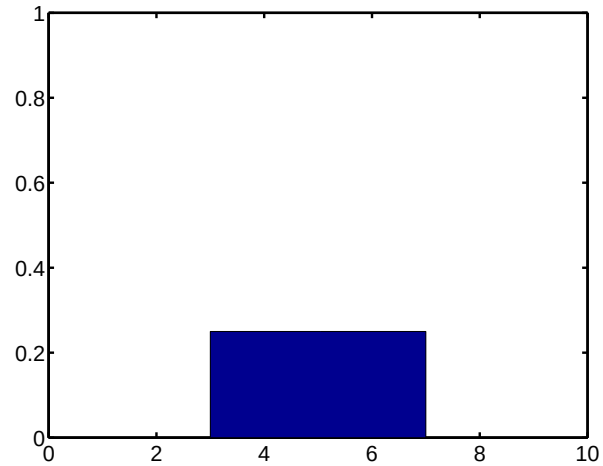
$$P(\mathrm{x} = x|\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

- Then x is Poisson distributed.



Uniform Distribution

- x is a random variable taking values $x \in [a, b]$.
- Let $P(\mathrm{x} = x) = 1/[b - a]$
- Then x is uniformly distributed.



Note

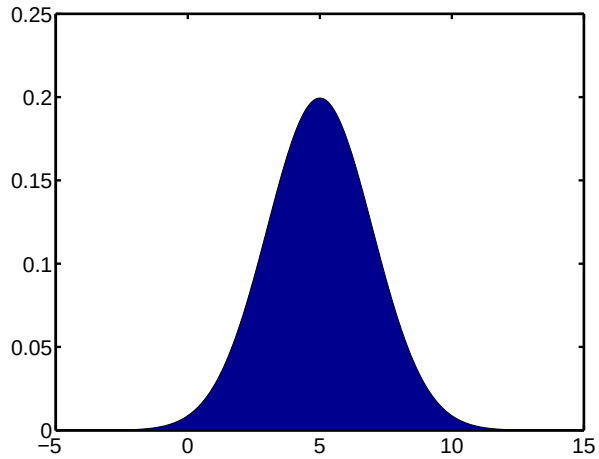
Cannot have a uniform distribution on an unbounded region.

Gaussian Distribution

- x is a random variable taking values $x \in \mathbb{R}$ (real values).
- Let $P(\mathrm{x} = x|\mu, \sigma^2) =$

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Then x is Gaussian distributed with mean μ and variance σ^2.

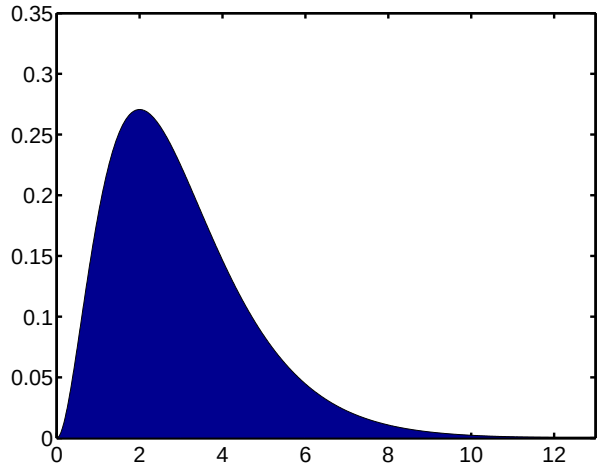


Gamma Distribution

- The Gamma distribution has a rate parameter $\beta > 0$ (or a scale parameter $1/\beta$) and a shape parameter $\alpha > 0$.
- x is a random variable taking values $x \in \mathbb{R}^+$ (non-negative real values).
- Let

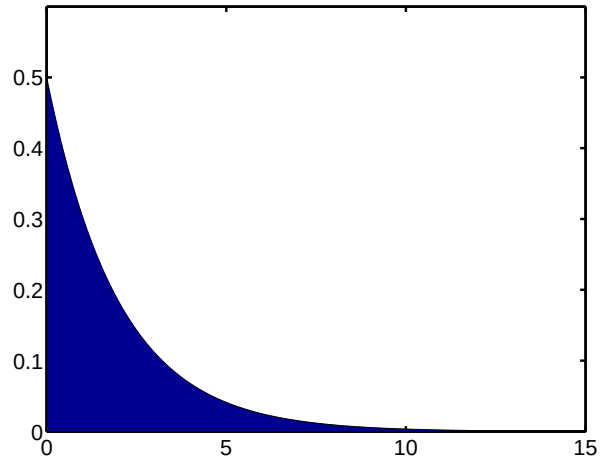
$$P(\mathrm{x} = x|\lambda) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \beta^{\alpha} \exp(-\beta x)$$

- Then x is Gamma distributed.
- Note the Gamma function.



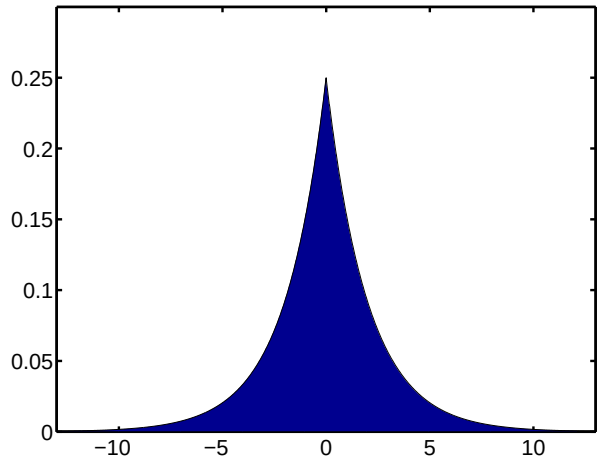
Exponential Distribution

- The exponential distribution is a Gamma distribution with $\alpha = 1$.
- The exponential distribution is often used for arrival times.
- x is a random variable taking values $x \in \mathbb{R}^+$.
- Let $P(\mathrm{x} = x|\lambda) = \lambda \exp(-\lambda x)$
- Then x is exponentially distributed.



Laplace Distribution

- The Laplace distribution is obtained from the difference between two independent identically exponentially distributed variables.
- x is a random variable taking values $x \in \mathbb{R}$.
- Let $P(\mathrm{x} = x|\lambda) = (\lambda/2)\exp(-\lambda|x|)$
- Then x is Laplace distributed.

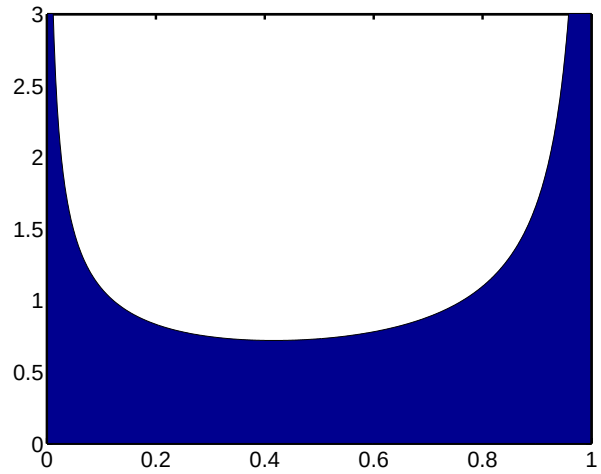


Beta Distribution

- x is a random variable taking values $x \in [0, 1]$.
- Let

$$P(\mathrm{x} = x|k) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

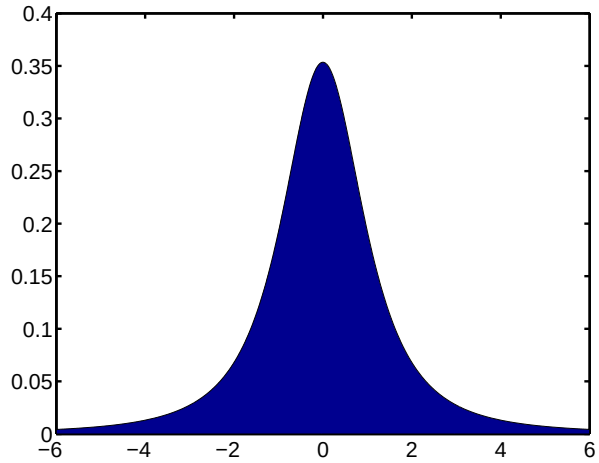
- Then x is $\beta(a, b)$ distributed.



Student t Distribution

- The Student t distribution is a *heavy tailed* distribution.
- x is a random variable taking values $x \in \mathbb{R}$.
- Let $P(\mathrm{x} = x|\nu) =$

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-(\frac{\nu+1}{2})}$$



- Then x is t distributed with ν degrees of freedom.
- The Cauchy distribution is a t distribution with $\nu = 1$.

The Kronecker Delta

- Think of a discrete distribution with all its probability mass on one value. So $P(\mathrm{x} = i) = 1$ iff (if and only if) $i = j$.
- We can write this using the Kronecker Delta:

$$P(\mathrm{x} = i) = \delta_{ij}$$

- $\delta_{ij} = 1$ iff $i = j$ and is zero otherwise.

The Dirac Delta

- Think of a real valued distribution with all its probability density on one value.
- There is an infinite density peak at one point (lets call this point a).
- We can write this using the Dirac delta:

$$P(\mathrm{x} = x) = \delta(x - a)$$

which has the properties $\delta(x - a) = 0$ if $x \neq a$, $\delta(x - a) = \infty$ if $x = a$,

$$\int_{-\infty}^{\infty} dx\ \delta(x - a) = 1 \text{ and } \int_{-\infty}^{\infty} dx\ f(x)\delta(x - a) = f(a).$$

- You could think of it as a Gaussian distribution in the limit of zero variance.

Other Distributions

- Chi-squared distribution with k degrees of freedom is a Gamma distribution with $\beta = 1/2$ and $k = 2/\alpha$.
- Dirichlet distribution: will be used on this course.
- Weibull distribution (a generalisation of the exponential)
- Geometric distribution
- Negative binomial distribution.
- Wishart distribution (a distribution over matrices). distributions.
- Use Wikipedia and Mathworld. Good summaries for distributions.

Things you must never (ever) forget

- Probabilities must be between 0 and 1 (though probability densities can be greater than 1).
- Distributions must sum (or integrate) to 1.
- Probabilities must be between 0 and 1 (though probability densities can be greater than 1).
- Distributions must sum (or integrate) to 1.
- Probabilities must be between 0 and 1 (though probability densities can be greater than 1).
- Distributions must sum (or integrate) to 1.
- Note probability densities can be greater than 1.

Summary

- I plan to challenge you.
- This is going to be hard. Keep up.
- Theoretical grounding is key.

To Do

Attending lectures is no substitute for working through the material! Lectures will motivate the methods and approaches. Only by study of the notes and bookwork will the details be clear. If you do not understand the notes then discuss them with one another. Ask your tutors.

Reading

These lecture slides. Chapter 1 of Barber.

Preparatory Reading

Barber Chapter 2.

Extra Reading

Cox's Axioms. Subjective Probability.