- Plan for the lectures:

Philosophy + math
- Interpretations of probability
- Quick refresher on prob. theory for many-dim. probability distributions

Statistics (not ML)
- Aspects of Bayesian statistics (parameter estimation and model comparison)

ML
- Gaussian processes (formalism + applications)

Main references:

- Sivia: "Data analysis, a Bayesian tutorial"
- Rasmussen & Williams: "Gaussian Processes for Machine Learning" (Available for free online!)

A brilliant and provocative gem:

- ET Jaynes: "Probability Theory: The Logic of Science"

Probability

- Q: How many have taken a course on prob. or statistics?

- Discussion: Discuss meaning of prob. using a coin flip or dice throw

- What does the statement $P(X) = 10\%$ mean?
- We don't know, or at least don't agree!
- Useful reference: "Interpretations of Probability", Stanford Encyclopedia of Philosophy

- Bertrand Russel, 1929: "Probability is the most important concept in modern science, especially as nobody has the slightest notion of what it means."

- Two main interpretations:
  - Frequentist: $\boxed{P(X) \equiv \lim_{n\to\infty} \frac{n_X}{n}}$ Prob. defined as long-run relative frequency
  - Bayesian: $\boxed{P(X) \equiv \text{degree of belief/knowledge that } X \text{ is true}}$
    - Degree of belief $\leftrightarrow$ subjective Bayesian
    - Degree of knowledge $\leftrightarrow$ objective Bayesian

- Formal/deductive logic: rules for reasoning with <u>certain</u> statements (Boolean logic)

- Cox and others: Find rules for <u>plausible reasoning</u>, i.e. <u>logic under uncertainty</u>

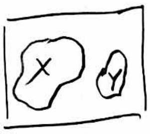
  ↳ "Rediscovered" the usual rules of prob. theory!

- Both freq. and Bayesian definitions of prob. agree with the Kolmogorov axioms that define the mathematical properties of the function $P(X)$ → Bayesians and frequentists use the same prob. rules, disagree about the interpretation.

Kolmogorov: $0 \leq P(X) \leq 1$

P is additive:

$$P(X \cup Y) = P(X) + P(Y) \quad \text{when } X \cap Y = \emptyset$$



- Frequentists: $\cancel{P(\text{hypothesis} \mid \text{data})}$ $\quad P(\text{data} \mid \text{hypothesis})$
- Bayesians: $P(\text{hypothesis} \mid \text{data})$ $\quad P(\text{data} \mid \text{hypothesis})$

- Subjective and objective Bayesians all happy with

$$P_{me}(X \mid I_1) \neq P_{you}(X \mid I_2)$$

- But objective Bayesians require that

$$P_{me}(X \mid I_1) = P_{you}(X \mid I_1)$$

Not required by subjective Bayesians!

- Usual rules for prob. theory does not tell us how to assign prob. in the first place, just how to <u>relate</u> probabilities in a consistent way! (Analogous with diff. eqs., which <u>relate</u> init. state to final state.)

- Objective Bayesians must introduce additional rules for assigning probabilities. Important example: "Maximum entropy"

Roughly saying: Given some information I, e.g. $X = 0.7 \pm 0.1$, choose the prob. distribution $P(x)$ that is the most uncertain but still consistent with I.

- Interpretations of prob. have important consequences:

1) Give rise to different approaches to statistics

Example: Bayesians can ask

$$P\left(\begin{matrix}\text{parameter}\\ \text{value}\end{matrix} \,\middle|\, \text{data}\right) = ?$$

Freq. cannot ask this, since prob. of a param. value does not make sense.

- Bayesian 95% credible interval for a parameter θ:

$$[0.1, 0.3]$$

"We have a 95% degree of belief that the true value of θ is between 0.1 and 0.3"

- Frequentist 95% confidence interval for θ:

$$[0.1, 0.3]$$

"If the experiment was repeated an infinite number of times, an interval constructed with this procedure should contain the true θ value in 95% of the repetitions"

2) Are probabilities necessarily linked to randomness?

- Is anything truly random? (Metaphysics, determinism, apparent vs. true randomness)

  Example: Is P(heads) in a coin flip necessarily 50%?

  Bayesian view: No necessary link between prob. and randomness. Can simply use prob. to quantify uncertainty. (But does not imply that randomness does not exist!)

- Probabilities in science, what do they mean?

  In particular: Interpretations of quantum mechanics.

Quick refresher on prob. theory

- My notation :

$$p(x) = \begin{cases} \text{probability for } x & [p(x)] = 1 \\ \text{probability density for } x & [p(x)] = \dfrac{1}{[x]} \end{cases}$$

Units :

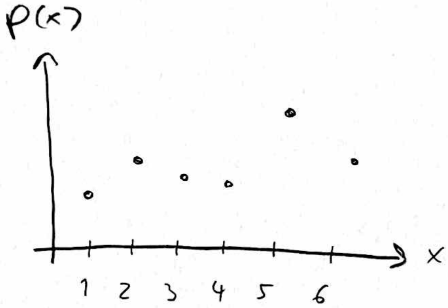
- With multiple variables :

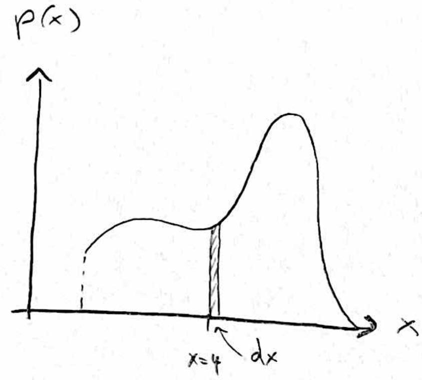
Should do : $p_x(x)$, $p_y(y)$, $p_{x,y}(x,y)$, $p_{x|y}(x|y)$

or alternatively: $f(x)$, $g(y)$, $h(x,y)$, $q(x)$

But I will be sloppy: $p(x)$, $p(y)$, $p(x,y)$, $p(x|y)$

- Discrete vs continuous :





$\text{Prob}(x=4) = p(4)$	$\text{Prob}(x \in [4, 4+dx]) = p(4)\,dx$
$0 \leq p(x) \leq 1$	$0 \leq p(x)\,dx \leq 1$
	Note : $p(x)$ can have arbitrarily large, positive value.
$\text{Prob}(2 \leq x \leq 4) = p(2) + p(3) + p(4)$	$\text{Prob}(2 \leq x \leq 4) = \int_2^4 p(x)\,dx$
$\sum_{\text{all allowed values}} p(x) = 1$	$\int_{\text{all allowed values}} p(x)\,dx = 1$

- We say that X "has a pdf $p(x)$", or "follows a pdf $p(x)$", or "is distributed as $p(x)$", etc.

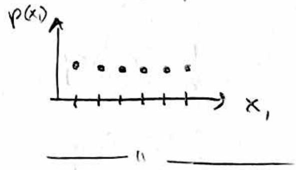
- Shorthand (but potentially confusing) notation:

$$X \sim p(x)$$

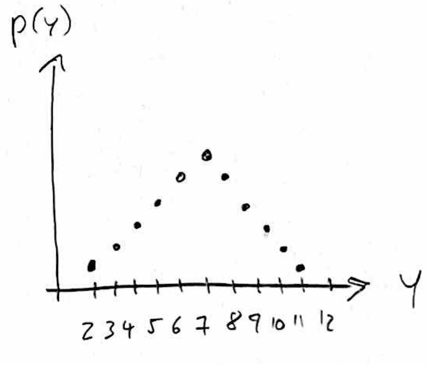
[Does <u>not</u> mean that "X is approximately equal to $p(x)$" or that "X is proportional to $p(x)$"!]

- Important reminder: A function of an uncertain/random variable, is itself a random variable!

Example: X_1: outcome of dice throw 1
X_2: outcome of dice throw 2



Let $Y \equiv X_1 + X_2$



Probability densities of many variables

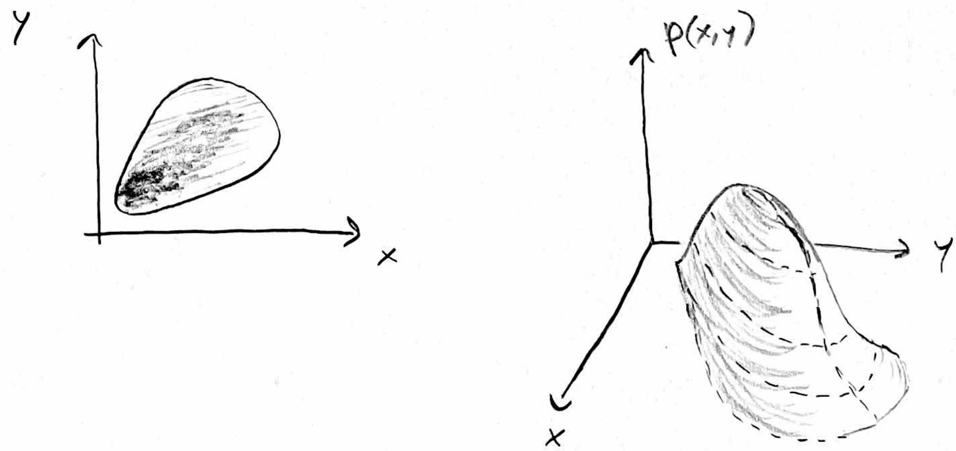
- Notation: $p(x_1, x_2, x_3, \ldots)$ or $p(\bar{x})$

  For two variables: will often use $p(x,y)$

- Will use 2D pdf as example

- Need to distinguish
  - *joint* prob. dens. $p(x,y)$ — 2D
  - *conditional* prob. dens. $p(x|y)$, $p(y|x)$ — 1D
  - *marginal* prob. dens. $p(x)$, $p(y)$ — 1D

- *Joint* pdf:
  - $p(x,y)\,dx\,dy = \text{Prob}\left(X \in [x, x+dx] \text{ and } Y \in [y, y+dy]\right)$



$$\left[\text{Normalisation: } \iint p(x,y)\,dx\,dy = 1 \right]$$

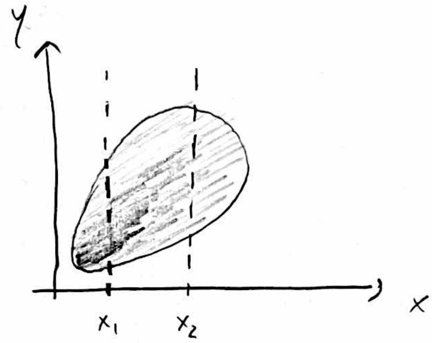
- ## Conditional pdfs

  - $p(y|x)dy = \text{Prob}(Y \in [y, y+dy]$ given a specific $X = x)$

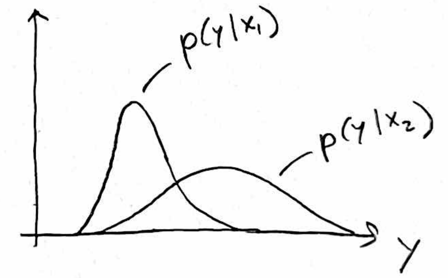
    [and similarly for $p(x|y)$]

  - Example :

    If the joint pdf $p(x,y)$ looks like this ...



... we can get conditional pdfs looking like this :



$p(x,y) = p(x)p(y)$

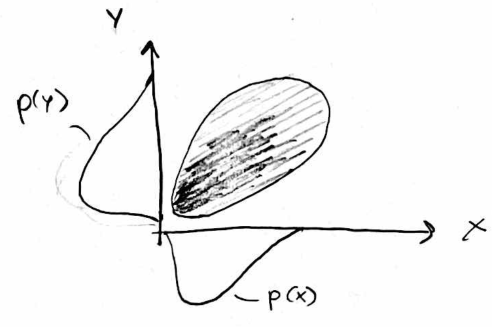
- ## Marginal pdfs

  - $p(x)dx = \text{Prob}(X \in [x, x+dx]$, irrespective of $Y)$

    [and similarly for $p(y)dy$]

$$p(x) = \int p(x,y)\,dy \qquad \text{"marginalise over } y\text{"}$$

$$p(y) = \int p(x,y)\,dx \qquad \text{" ——— '' ——— } x\text{"}$$



- Useful relations:

Bayes' theorem $\quad P(A|B) = \dfrac{P(B|A)\,P(A)}{P(B)}$

1) $p(x,y) = p(x|y)\,p(y) = p(y|x)\,p(x)$

2) $p(x) = \int p(x,y)\,dy = \int p(x|y)\,p(y)\,dy$

$p(y) = \int p(x,y)\,dx = \int p(y|x)\,p(x)\,dx$

[The conditional pdfs weighted according to the other marginal pdf]

Discrete case:
$p(x) = \sum_y p(x,y) = \sum_y p(x|y)\,p(y)$

Analogous for $p(y)$

- With 1) and 2) we can express Bayes' theorem as

$$p(y|x) = \frac{p(x|y)\,p(y)}{p(x)} = \frac{p(x|y)\,p(y)}{\int p(x|y)\,p(y)\,dy}$$

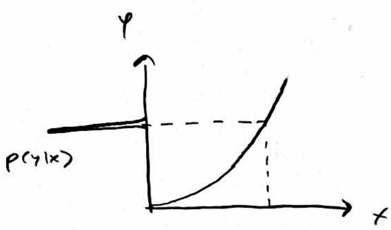
[analogous for discrete case]

- Sometimes a "deltafunction perspective" is useful:

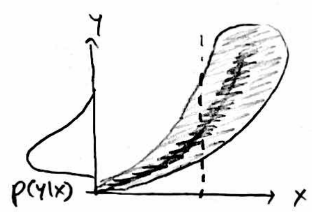
  - Instead of:
    - x is an uncertain variable with pdf $p(x)$
    - $y = x^2$ is a function of x

  - Rather:
    - x and y are uncertain variables
    - The statement $y = x^2$ is just saying that, given an x value, we are 100% certain what y is. In other words: $p(y|x) = \delta(y - x^2)$

deltafunction pdf!



is a limit of the general case, e.g. this $\longrightarrow$



- Correctly relates the probabilities $p_Y(y)$ and $p_x(x)$:

$$p_Y(y) = \int p(x,y)\, dx$$

$$= \int p(y|x)\, p_x(x)\, dx$$

$$= \int \delta(y-x^2)\, p_x(x)\, dx$$

$$p_Y(y) = p_x(x=\sqrt{y})$$

One way of understanding why the procedure

1) Sample $x \sim p(x)$
2) Evaluate $y = x^2$ for all samples
3) Histogram y samples

gives a histogram that approximates $p(y)$