Quick refresher on prob. theory

- My notation : $p(x) = \begin{cases} \underline{\text{probability}} \text{ for } x & \\ \underline{\text{probability density}} \text{ for } x & \end{cases}$

  Units :

  $[p(x)] = 1$

  $[p(x)] = \frac{1}{[x]}$

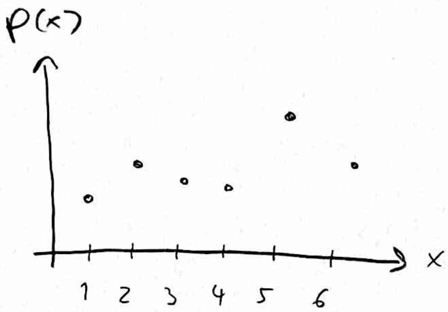
- With multiple variables :

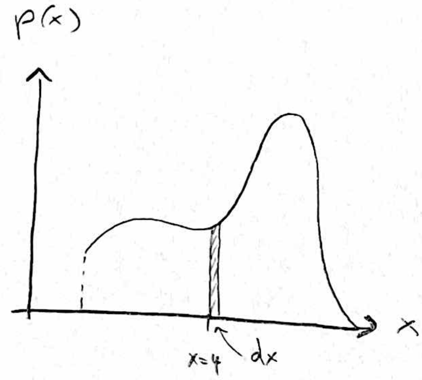
  Should do : $p_x(x)$, $p_y(y)$, $p_{x,y}(x,y)$, $p_{x|y}(x|y)$

  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.

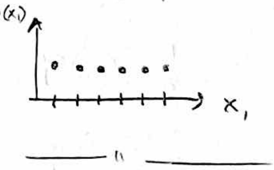
- Shorthand (but potentially confusing) notation:

$$X \sim p(x)$$

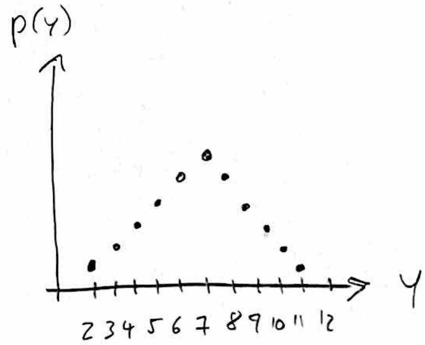
[Does <u>not</u> mean that "X is approximately equal to $p(x)$" or that "X is proportional to $p(x)$"!]

- 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 $p(x, y)$

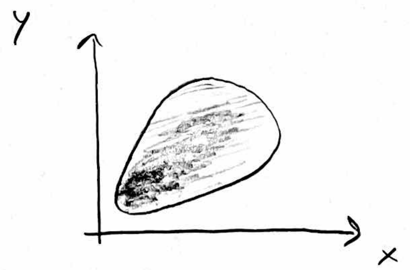
- Will use 2D pdf as example

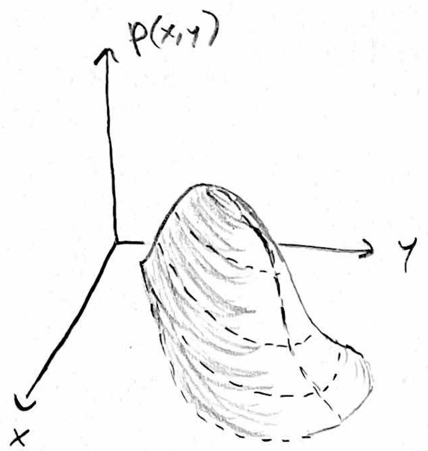
- Need to distinguish

  - 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:

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





[Normalisation: $\iint p(x,y)\,dx\,dy = 1$]

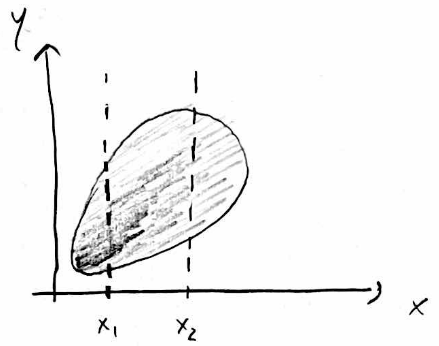
- Conditional pdfs

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

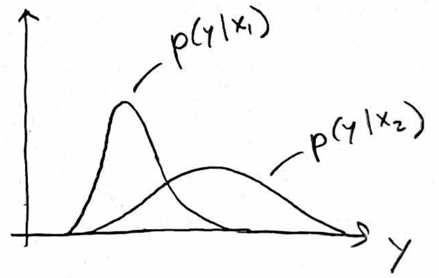
    [and similarly for $p(x|y)$]

  - Example:

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



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



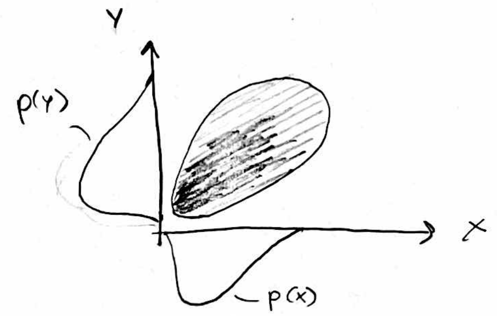
- Marginal pdfs

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

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

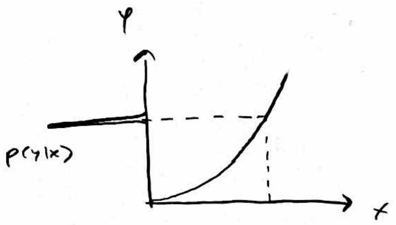
[analogous for discrete case]

- Sometimes a "deltafunction perspective" is useful:

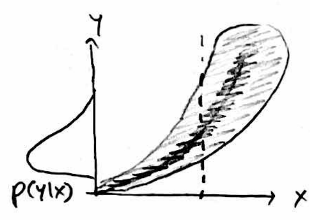
  - Instead of:
    - x is an uncertain variable with pdf $p(x)$
    - $y = x^2$ is a function of x

  - 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 $p(y)$

Aspects of Bayesian statistics

- Probabilities, probabilities, probabilities!
- Starting point: $P(X) \equiv$ degree of belief/knowledge that X is true
- Bayes theorem

$$\boxed{P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}}$$

Both frequentists and Bayesians use this

Let: H: hypothesis
D: data
I: any other information

Bayesians can discuss $P(H)$, $P(H|D)$, etc., so we can write

$$\boxed{P(H|D,I) = \frac{P(D|H,I)\,P(H|I)}{P(D|I)}}$$

- We often drop the ~~I~~ conditional on I for simplicity, but should remember it's always there!

-
  - $P(H|I)$: Prior prob. for H
  - $P(D|H,I)$: The probability for data D given that H is correct
  - $P(H|D,I)$: Posterior prob. for H, updated from prior in light of the new data D.
  - $P(D|I)$: The "Bayesian evidence"

- Given a set of mutually exclusive hypotheses $H_1, H_2, \ldots$

$$P(D|I) = \sum_{H_i} P(D, H_i|I) = \sum_{H_i} P(D|H_i,I)\,P(H_i|I)$$

$$= P(D|H_1,I)\,P(H_1|I) + P(D|H_2,I)\,P(H_2|I) + \ldots$$

- Typically distinguish two types of applications:

  - Parameter estimation : $p(\theta | D, M) = ?$
  - Model comparison : $\dfrac{P(M_1 | D)}{P(M_2 | D)} = \dfrac{P(D | M_1)}{P(D | M_2)} \dfrac{P(M_1)}{P(M_2)} = ?$

- Parameter estimation:

  $H \rightarrow \theta$, i.e. some value for a cont. parameter

$$\Rightarrow \quad p(\theta | D, M) = \frac{p(D | \theta, M)\, p(\theta | M)}{p(D | M)}$$

$$= \frac{L(\theta)\, \pi(\theta)}{Z} = \frac{L(\theta)\, \pi(\theta)}{\int L(\theta)\, \pi(\theta)\, d\theta}$$

{ will do param. estimation in more detail soon }

- The likelihood function, $L(\theta)$:

$$\boxed{p(D | \theta, M)}$$

  - Read as a function of D, given a fixed value for the parameter θ: a prob. distr. for possible data D.

  - If we insert the observed data $D = D_{obs}$ and read $p(D = D_{obs} | \theta, M)$ as a function of θ: the likelihood function $L(\theta)$, which is not a pdf!

Example: Poisson dist.: $p(n | \lambda) = \dfrac{\lambda^n e^{-\lambda}}{n!}$

- Discrete prob. distr. for discrete variable n
- Continuous likelihood function for cont. parameter λ

- The prior, $\pi(\theta)$:

  - Most controversial (and most useful?) aspect of Bayesian statistics

  - The formalism requires us to quantify our a priori assumptions using probabilities

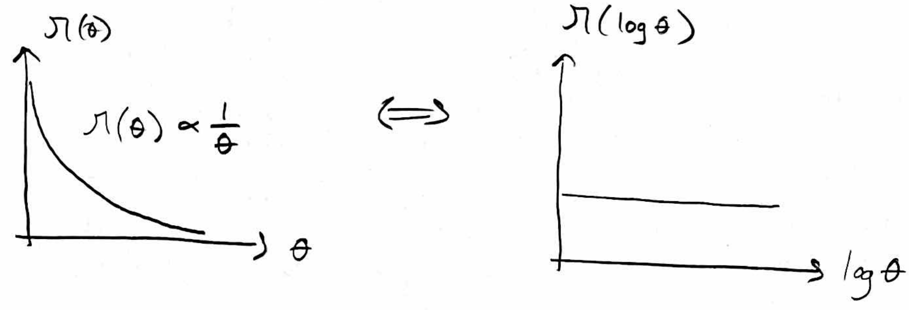
  - $\pi(\theta)$ = our degree of belief in value θ, before seeing the data D.

  - How to choose $\pi(\theta)$?

    - Subjective vs objective

    - Often want to express "complete uncertainty", but in what variable?

      "Log prior"



    - How to encode existing information?

- The "marginal likelihood" / "Bayesian evidence", Z :

$$Z = p(D|M) = \int p(D,\theta|M)\,d\theta$$
$$= \int p(D|\theta,M)\,p(\theta|M)\,d\theta$$
$$= \int L(\theta)\,\Pi(\theta)\,d\theta$$

[likelihood × prior integrated across the model parameter space.]

- In general: Difficult to compute Z! High-dim integral and $L(\theta)$ can be sharply peaked with long tails, multimodal, etc.

- Not important for parameter estimation, since all θ-dependence is integrated out. Plays the role as norm. constant:

$$p(\theta|D) = \frac{L(\theta)\Pi(\theta)}{Z} \propto L(\theta)\Pi(\theta)$$

- Z is the key quantity for model comparison :

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(D|M_1)}{p(D|M_2)}\frac{p(M_1)}{p(M_2)} = \frac{Z_1}{Z_2}\;\frac{\Pi(M_1)}{\Pi(M_2)}$$

posterior odds

Bayes factor/ evidence ratio

prior odds (often set to 1)

- How to interpret Bayes factor (and/or posterior ratio)?

Common to use <u>Jeffrey's scale</u> (or Kass & Raferty)
(Convention, similar to convention of doing frequentist hypothesis tests with a p-value threshold of 5%)

$$\text{Bayes factor} \quad B_{12} = \frac{P(D|M_1)}{P(D|M_2)} = \frac{Z_1}{Z_2}$$

$\lvert \ln B_{12} \rvert$	Odds	Strength of evidence in favor of M_1 when compared to M_2	Post. prob. $P(M_1\|D)$ if $P(M_1) = P(M_2) = 0.5$
< 1.0	$\lesssim 3:1$	Inconclusive	< 0.75
1.0	$\approx 3:1$	Weak evidence	0.75
2.5	$\approx 12:1$	Moderate evidence	0.923
5.0	$\approx 150:1$	Strong evidence	0.993

- Note on prior dependence:

$$B_{12} = \frac{Z_1}{Z_2} = \frac{\int \mathcal{L}_{M_1}(\theta_1)\,\Pi_{M_1}(\theta_1)\,d\theta_1}{\int \mathcal{L}_{M_2}(\theta_2)\,\Pi_{M_2}(\theta_2)\,d\theta_2}$$

From point of view of <u>Bayesian model</u> comparison:
A model specification <u>includes</u> the choice of parameter priors!

Even if prior ratio for the <u>full</u> models is set to 1 (i.e. $P(M_1) = P(M_2) = 0.5$), the Bayes factor still <u>depends on the parameter priors within each model</u>

Not so much of a problem in <u>nested models</u>, where e.g. M_1 is a subset of the M_2 parameterspace

M_2: θ_a, θ_b

M_1: θ_a, $\theta_b = 0$

Then we might use similar prior for θ_a in the two models

- Bayes theorem as tool for consistent reasoning:

1) Reminds us that how plausible we should judge a hypothesis to be <u>depends on the alternatives</u>.

If we only had <u>one</u> hypothesis H:
$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} = \frac{P(D|H)P(H)}{P(D|H)P(H)} = 1$$

$$\left[\begin{array}{l}\text{i.e. prob. 1 for } H \text{ indep. of} \\ \text{the data } D.\end{array}\right]$$

Recall: $P(H_1|D) + P(H_2|D) + \ldots = 1$ so prob $P(H_i|D)$ dep. on prob. for $P(H_{j\neq i}|I)$

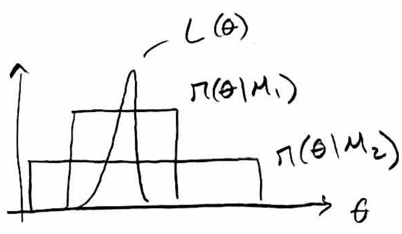
2) "<u>Extraordinary claims</u> require <u>extraordinary evidence</u>"

$$\frac{P(H|D)}{P(\bar{H}|D)} = \frac{P(D|H)}{P(D|\bar{H})}\frac{P(H)}{P(\bar{H})}$$

If $P(H)$ is tiny (H is an extraordinary claim), then $P(D|H)$ must be huge (extraordinary evidence) if we are to prefer H over $\bar{H}$.

3) <u>Occam's razor</u>

In a model comparison, models with fewer free parameters and more restrictive priors will be preferred, unless the data strongly prefers/requires a complex model



$$Z_1 = \int L(\theta)\,\pi(\theta|M_1)\,d\theta$$

$$Z_2 = \int L(\theta)\,\pi(\theta|M_2)\,d\theta$$

$\longrightarrow$ $Z_1 > Z_2$ by this Occam's razor effect.

Bayesian parameter estimation

- Starting point:
  - Assume a model $\mathcal{M}$ with parameters $\theta_1, \theta_2, \theta_3, \ldots$
  - Assign prior belief on parameter space

$$\pi(\theta_1, \theta_2, \ldots)$$

  - In practice, often choose $\pi(\theta_1, \theta_2, \ldots) = \pi_{\theta_1}(\theta_1)\, \pi_{\theta_2}(\theta_2) \ldots$ "separable prior"

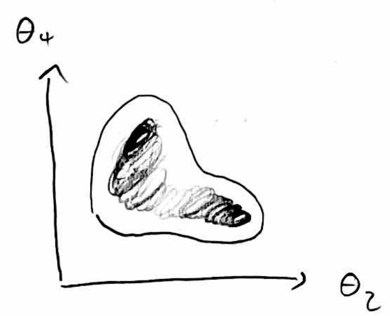
    ↑ 1D priors

  - Construct likelihood function $\mathcal{L}(\theta_1, \theta_2, \ldots) = \mathcal{L}(\bar{\theta})$ by formulating $p(D|\bar{\theta})$ and inserting $D = D_{obs}$.

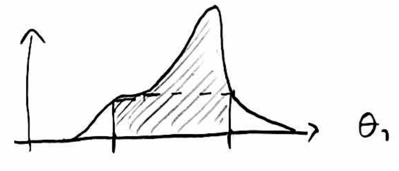
- Goal:
  - Theoretically: Obtain <u>posterior</u> $p(\bar{\theta}|D) = \dfrac{\mathcal{L}(\bar{\theta})\,\pi(\bar{\theta})}{Z}$
  - In practice: Obtain a set of <u>$\bar{\theta}$-samples from $p(\bar{\theta}|D)$</u> and use these to approximate properties of the posterior

- What to present?:
  - 1D / 2D marginalised posteriors, i.e.

$$p(\theta_2, \theta_4 | D) = \int\int p(\theta_1, \theta_2, \theta_3, \theta_4 | D)\, d\theta_1\, d\theta_3$$



  - 68/95/99 % credible regions/intervals

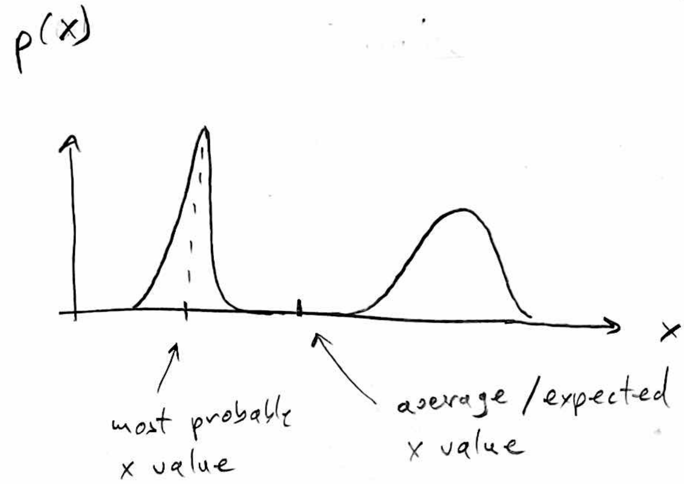


- Expectation values:

$$E[\theta] = \int \theta \, p(\theta|D) \, d\theta$$

$\approx$ average value of θ in set of posterior θ-samples

Note: Don't confuse expectation value with most probable value



[Example: Expected number of heads in a single coin toss: 0.5
~~[illegible]~~ But $P(0.5 \text{ heads}) = 0$.]

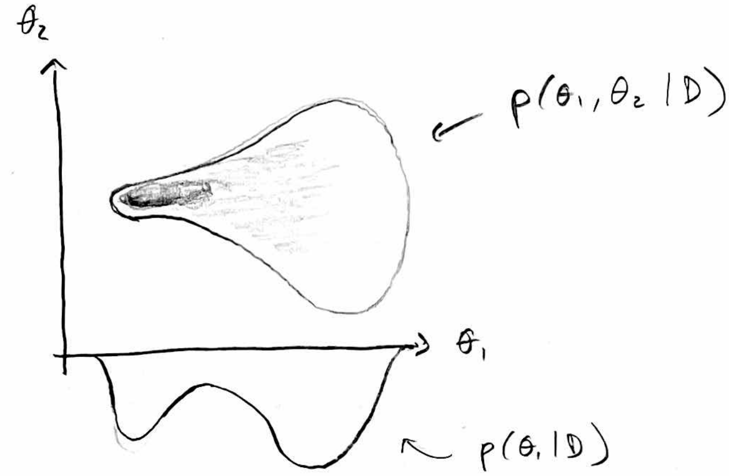
- [Look at arxiv: 2009.03286 as example of how to present many-dim posterior]

[Lecture ended here.]

- Keep in mind that integrating/marginalising out parameters can give equally large contributions in two different ways:

  1) $\mathcal{L}(\bar{\theta}) \times \pi(\bar{\theta})$ is large over some small region of $\bar{\theta}$ space

  2) $\mathcal{L}(\bar{\theta}) \times \pi(\bar{\theta})$ is small but non-zero over a large region of $\bar{\theta}$-space



$\Rightarrow$ Bayesian posteriors penalize "fine-tuning":
If $\mathcal{L}(\bar{\theta})$ is high along some narrow strip in $\bar{\theta}$-space, that will make a small impact on $p(\bar{\theta}|D)$

[will see example of this]

- What is regarded as "fine-tuned"?
  We implicitly choose this when choosing $\pi(\bar{\theta})$, in the way we distribute our probability across $\bar{\theta}$-space.

- [Look at GAMBIT paper as example of Bayesian param. est.]
  arXiv: 1705.07931 : $\boxed{\text{Difference between } L(\theta_1) \equiv L(\theta_1, \hat{\hat{\theta}}_2) \text{ and } p(\theta_1|D) = \int p(\theta_1,\theta_2|D)\, d\theta_2}$

  [Compare Fig. 1 (left) and Fig. 7 (left) to see example of fine-tuning, from arxiv: 1808.10465]

- Practical challenge:
  How to obtain a sufficiently dense set of $\bar{\theta}$-samples according to some high-dimensional and typically multimodal $p(\bar{\theta}|D)$?

  Alt 1) Some version of MCMC sampling

  Alt 2) Some version of nested sampling

  we'll look at this.