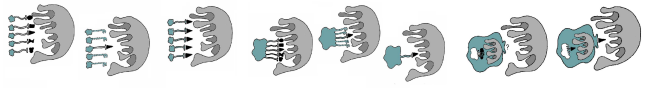
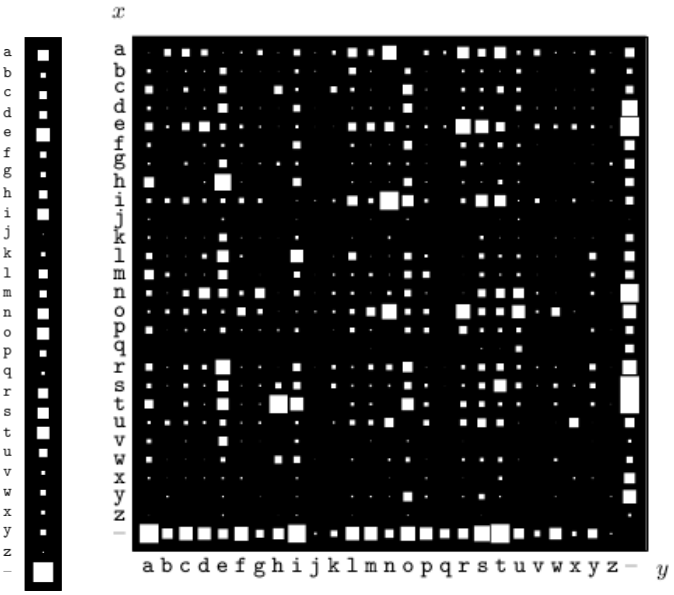
uncertainty

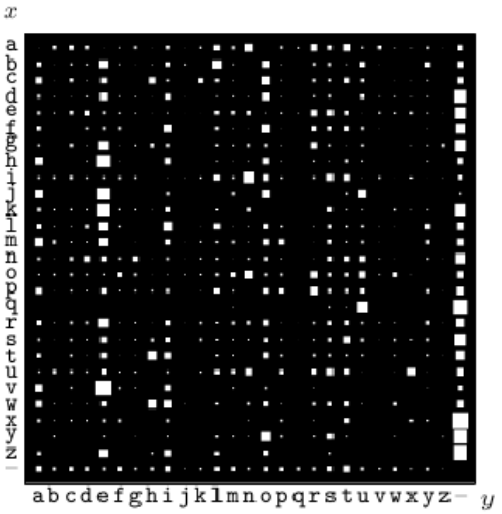
What's the right way to deal with uncertainty? Is there an ideal way to make predictions from limited data? Solutions to inference problems always seem to have aspects that appear heuristic or *ad hoc* in nature. Can this *odd hack*-ery be eliminated, or are we stuck with it?



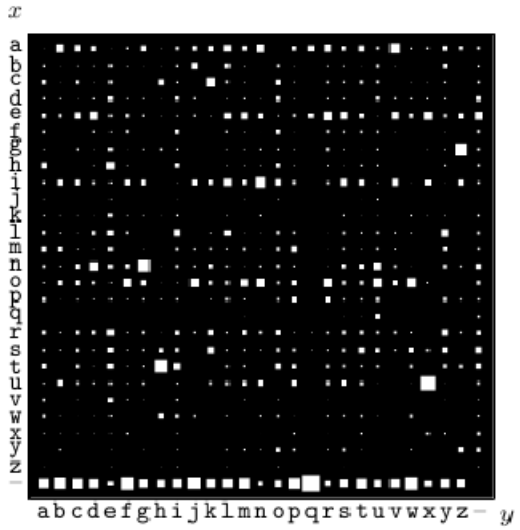
PROBABILITIES

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928





(a) $P(y \mid x)$



(b) $P(x \mid y)$

the sum rule
Where there are a number of mutually exclusive possibilities, their probabilities sum to one. *E.g.* if I roll a die,

$$P_1 + P_2 + P_3 + P_4 + P_5 + P_6 = 1$$

And if some event x either happens or doesn't happen,

$$P(x) + P(\neg x) = 1$$

If there are two events x and y, and the outcome of y is unknown,

$$P(x) \;=\; \sum_y P(x, y)$$

If the space of possible y is continuous, that sum will become an integral.

the product rule

The probability that both x and y are true is the probability that x is true *times* the probability that y is true, given that x is true already:

$$P(x, y) = P(x)\,.\,P(y|x)$$

A special case of this is where y and x are completely *independent* events, in which case $P(y|x) = P(y)$ since the truth or otherwise of x has no effect on whether y is true. For example if I throw a coin and then a die, the probability I get a head and a 6 is $P_{\text{head}}\, P_6$.

Notice $P(x, y)$ can be written in either of two ways, so we have that $P(x)\, P(y|x) \;=\; P(y)\, P(x|y)$, or

$$P(y|x) \;=\; \frac{P(y)\, P(x|y)}{P(x)}$$

known as *Bayes theorem*. The denominator can be found using the sum rule.

THE COX-JAYNES AXIOMS

As philophers from at least as far back as Hume have pointed out, we can never *know* the "real" state of things with absolute certainty. Knowledge always comes down to *beliefs*: assertions about the real state of the world. How should beliefs behave? How should beliefs change in the light of evidence?

Denote the plausibility of[1] A being true, given that $\mathcal{H}$ is true, by $\text{plaus}(A|\mathcal{H})$.

Cox and Jaynes, middle of the last century, posed three axioms about plausibilities:

1. Transitivity:

$$\begin{aligned} \text{if} \quad & \text{plaus}(A|\mathcal{H}) \;> \text{plaus}(B|\mathcal{H}) \\ \text{and} \quad & \text{plaus}(B|\mathcal{H}) \;> \text{plaus}(C|\mathcal{H}), \\ \text{then} \quad & \text{plaus}(A|\mathcal{H}) \;> \text{plaus}(C|\mathcal{H}) \end{aligned}$$

   This implies plausibilities can be mapped onto real numbers. We can arbitrarily squash the reals into a given finite range, so map "totally positive it is true" to 1, and "absolutely ruled out" to 0. Incredibly, just two more axioms are needed to calculate a numerical value on $\text{plaus}(A|\mathcal{H})$.

[1] "plausibility of", or "preference for", or "confidence in", or "degree of belief in"...

2. Our confidence in a proposition being true is directly related to our confidence that it's false, and this *relationship* should be independent of the details of the actual proposition. That is:
$$\text{plaus}(\neg A|I) = F\left[\text{plaus}(A|I)\right]$$

3. If there are two ways of arriving at the plausibility of some proposition, both these ways should arrive at the same result, provided they use the same background information. Specifically, the truth of some compound proposition AB (meaning "both A and B are true") can be found be establishing A first and then B (given A), *or* by finding B first and then A given B.
$$\text{plaus}(AB|I) = G\left[\text{plaus}(A|I),\ \text{plaus}(B|AI)\right]$$
Note this is invariant to interchanging A and B.

These three axioms, together with the scaling choice of $\text{plaus} \in [0, 1]$, entirely determine how to calculate degrees of belief, for they imply that $F(x) = 1 - x$ and $G(x, y) = xy$. This leads to the *product rule*:
$$\text{plaus}(AB|I) = \text{plaus}(A|I)\,.\,\text{plaus}(B|A, I)$$
and the *sum rule*:
$$\text{plaus}(A|I) + \text{plaus}(\neg A|I) = 1$$
Notice that if all uncertainty is removed, these reduce to the two basic rules of Boolean algebra for the negation and conjunction of propositions.

Plausibilities are isomorphic (obey exactly the same rules as) probabilities. Thus all coherent beliefs and predictions can be mapped to probabilities, and the 2 laws of probability describe how to update these beliefs and predictions in the light of data. *So the language of inference is probability theory*.

Any inference scheme other than using these two laws of probability is inconsistent with at least one of the Cox-Jaynes axioms. As Bayesian physicist John Skilling has put it,

> 'In engineering we respect the laws of physics or court disaster, and in mathematics we respect those of arithmetic: in inference we should respect the laws of probability'.

TWO VIEWS OF PROBABILITY

As Jacob Bernoulli pointed out in about 1713, if a fair coin is tossed over and over, the ratio of heads to the total number of throws gets closer and closer to $\frac{1}{2}$. So if I say (of a bent coin) "the probability of a head is 31%", I really mean that
$$\lim_{N\to\infty} \frac{N_{\text{heads}}}{N} = 0.31$$
where $N = N_{\text{heads}} + N_{\text{tails}}$. That is, the term probability is refering to the expected number of observations of a certain event or outcome, given that we can repeat the experiment as often as we like.

This refers to **effects** given **causes** then: *if* the coin is fair (the cause, or hypothesis) *then* on average it'll come up heads 50% of the time (the observable effect). This use of probabilities to reason about effects given causes should be very familiar to you. We've already used it to calculate things like the likelihood of a learner getting a certain output correct. In that case,

we treated learning as heading towards parameter values that maximized the likelihood. This could be called 'forward' inference, since it goes from assumed causes to uncertain effects.

In what we could call the "frequentists" view of statistics, probabilities are only used to describe the *expected outcomes of repeated events*: they're used exclusively for inferences about new data, given an assumed cause.

How then should we reason about the causes themselves ('causal inference')? Afterall at some point we need to decide whether to treat the coin as fair or not. This kind of inference is sometimes called "inverse", since it goes from observed effects *back* to uncertain causes.

Frequentist statistics has many techniques for deciding between competing causes given some data - confidence intervals, Student t tests, chi-squared tests, the acceptance/rejection of null hypotheses and so on and so forth. Lots of distinct methods. But in the light of the Cox axioms, we'll instead be following the Bayesian school of thought on probabilities, which is that they can (in fact should) also describe degrees of belief about unknown causes.

Bayesian statistics

In 1763 the Reverend Thomas Bayes established the theorem that bears his name, immediately igniting a fierce debate that has raged through the centuries virtually unabated. It wasn't so much the formula (which is simple and obvious when applied to letter frequencies) as the fact that the Reverend applied it to causes, not just effects:

$$P(\text{cause}|\text{effect}) \;=\; \frac{P(\text{effect}|\text{cause})\,P(\text{cause})}{P(\text{effect})}$$

We can all interpret things like $P(\text{effect}|\text{cause})$, but what does it mean to have a probability for a cause? The idea that this is the limit of some statistic when the universe is re-run thousands of times seems inappropriate.

Also the answer on the left hand side depends on the probability of the cause alone, *prior* to any evidence of its effects becoming known. Such an object can't have anything to do with the outcomes of hypothetical repeated experiments.

The Bayesian view is that probabilities should be used to describe *all* uncertainty, not just the outcomes of observable experiments.

The terms in Bayes theorem are used so much that they're given names. If we have some hypothesis H (*i.e.* possible cause) and data D (its observable effect) then $P(D|H)$ is the **likelihood** given the hypothesis, $P(H)$ is the **prior** probability of the hypothesis, and $P(H|D)$ is the **posterior** probability of the hypothesis once we've seen the data. Finally $P(D)$ is sometimes called the **evidence** - you can think of it as "just" normalisation that ensures $\sum_H P(H|D) = 1$, where the sum is over all competing hypotheses.

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

$$\text{posterior} \;=\; \frac{\text{likelihood . prior}}{\text{evidence}}$$

The 'evidence' can always be found by applying the sum rule, $P(D) = \sum_H P(H, D)$, which is often re-written in the following way[2] so as to match the form of the numerator in Bayes Theorem:

$$P(D) \;=\; \sum_H P(D|H)\,P(H)$$

[2]by using the product rule.

And if H came from a continuous space of possible 'causes', that sum would be replaced by an integral.

Looking at the sum rule formula more closely, essentially we get rid of the thing we're uncertain about (here it happens to be H) by summing the likelihood over all the values the unknown could possibly be, with each one weighted by its prior probability. This is called *marginalising over* or *integrating out* the unknown quantity H - it is an operation that is extremely common in what follows.

WORKING WITH DEGREES OF BELIEF

posterior versus likelihood

It is extremely common for confusion to arise between a probability expressing a likelihood and one expressing a posterior belief. Some examples:

- For the Springboks, $P(\text{lostgame}|\text{inDunedin})$ is very high, whereas $P(\text{inDunedin}|\text{lostgame})$ is quite low.
- There are many windy places in the world so $P(\text{Wellington}|\text{windy})$ is negligible. Does this mean $P(\text{windy}|\text{Wellington})$ is also low?
- A murder suspect tests positive in a DNA test for which $P(+\text{ve}|\text{innocent})$ is 1 in a million. So is it a good guess that the guy is guilty?
- A prize is hidden behind one of three doors labelled A, B, C. You're thinking of choosing door A when the presenter (who knows where the prize is) goes to C and opens it to show you it's not there. Should you switch from A to B?
- Outside my door there used to be a poster sent to us from Oxford University's Statistics department. It has a long list of names of illustrious people who studied at Oxford. Below this it says "Recognise some of these and you'll realise the probability of a successful career having studied at Oxford". They're saying that high $P(\text{Oxford}|\text{excellent})$ implies high $P(\text{excellent}|\text{Oxford})$.

comparing models

When there are two or more models for the data, which one should we prefer to use, *i.e.* which is more plausible? Bayesians compare models[3] based on their posterior probabilities.

Take the case of two models H_1 and H_2 for example. A simple way to compare them is to take their ratio and see if it's greater than one:

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)\ P(H_1)}{P(D|H_2)\ P(H_2)}$$

Notice that the 'evidence' term has been cancelled out, which is very convenient, since it involves a sum over all the models being considered. Notice also that if we have no prior reason to prefer one model over another, all we have to do is look at their likelihoods on the training set.

More generally then, if we want to choose the best out of say K different models then it's sufficient to look at their likelihoods times their priors, since the evidence is the same for all of them. The most plausible model will have the highest value of $P(D|H)\ P(H)$.

[3] as well as theories, hypotheses, unknown parameter values, and so on.

priors

Prior distributions reflect our uncertainty about models before the data arrives, and so they are essentially *subjective* - and yet different results follow from using different priors!

On the other hand, how can one conceive of doing inference without making any assumptions? The very definition of "assumption" admits that we cannot determine such a thing beforehand[4]. The Bayesian view is that orthodox statisticians also make assumptions, often without realising it. A great strength of the Bayesian approach is that it virtually forces you to make your assumptions explicit - you simply can't do the calculations otherwise.

Priors do seem to be irreparably subjective, and yet there are good arguments for guiding our prior beliefs. The best known of these is **maximum entropy**: in the absence of better information, we should use the prior distribution that intrinsically assumes the least, *i.e.* is 'maximally non-committal'. The concept of entropy is used to correctly measure the amount of uncertainty present in a distribution. For example a Gaussian distribution with a high variance assumes less than one with a low variance, and indeed the entropy of a Gaussian is proportional to its variance. Two examples:

- if all we know about some random variable x is that it is between a and b, the maximum entropy prior turns out to be the uniform distribution between a and b.
- if all we know about some random variable x is its mean and variance, the maximum entropy prior turns out to be the Gaussian distribution with that mean and variance.

Despite its intuitive appeal there seem to be cases where Maximum Entropy seems to give the wrong answer. In fact it seems unlikely that *any* general principle can be found for the determination of priors.

improving models

Recall that we've looked at maximum likelihood as a way of deriving learning rules that lead to improved models. For example, think of the particular weight values of a single linear neuron as being a *model* of[5] the world generating the data. By training the neuron via the delta rule, we find a point in weight space which is a local maximum of the likelihood, in other words a model under which the *data* is most likely. Now we can see this isn't quite the right thing to do: if we're going to take a model and improve it, shouldn't we aim to make the model more plausible given the data? This maximizing the posterior $P(H|D)$ rather than the likelihood $P(D|H)$, and is known as **maximum a posteriori** inference (MAP).

First we'll need to specify a prior distribution over the weights. So, before you've seen any data, what do you think plausible values for the weights are?! This seems a daft question, but we're *required* to come up with a prior distribution - there's no way around it. A very common choice is a broad Gaussian prior, *e.g.*

$$P(\mathbf{w}) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{w_i^2}{2\sigma^2}\right]$$

and for future reference, here is its logarithm:

$$\log P(\mathbf{w}) = -\frac{1}{2}\log 2\pi\sigma^2 \; - \; \frac{1}{2\sigma^2}\sum_i w_i^2$$

[4] If we could, it wouldn't need to be assumed!...

[5] or 'hypothesis about', if you prefer.

So now let's see what MAP inference entails for a simple linear neuron then. Maximizing the posterior probability $P(\mathbf{w}|D)$ is the same as maximizing its logarithm:

$$\log P(\mathbf{w}|D) = \log P(D|\mathbf{w}) \; + \; \log P(\mathbf{w}) \; - \; \log P(D)$$

Take the gradient:

$$\nabla_{\mathbf{w}} \log P(\mathbf{w}|D) = \nabla_{\mathbf{w}} \log P(D|\mathbf{w}) \; + \; \nabla_{\mathbf{w}} \log P(\mathbf{w})$$

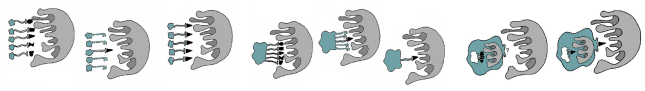
Notice we (thankfully) lost the evidence term as it's an integral over *all* weight vectors and doesn't depend on our particular choice of $\mathbf{w}$. Now we move the weights vector in the direction of this gradient. The first term gives us the delta rule, and the second is...?

$$\Delta^{\mu} w_i \quad \propto \quad \left[\sum_{\mu} (Y^{\mu} - y^{\mu}) \, x_i^{\mu} \right] \quad - \quad \beta \, w_i$$

where $\beta = \sigma^{-2}$ and μ indexes the training example. Weight decay! That is, we've shown that the heuristic of decaying weights towards zero corresponds to doing MAP inference of the most plausible weights, under a Gaussian prior. And this immediately gives us a procedure for making other learning rules, corresponding to different prior beliefs about what weight values seem plausible.

Notice that the MAP procedure comes down to the maximum likelihood (ML) one plus a 'correction' for model complexity then. The inclusion of the subjective prior distribution has (a) been *required* in the Bayesian formulation, and (b) resulted in a complexity control.

So one way of putting it is that the Bayesian formulation clarifies where **max likelihood** comes from: it's an approximation to **max a posteriori** in the case where we have a 'flat' prior.



Uncertainty about the world should be represented in the form of probability distributions, and updated in the light of data by following the rules of probability theory. However this raises serious issues of tractability. Background assumptions and tractability are intricately linked.

reading

There's lots to read in the textbook on this:

- chapter 2, but not section 2.4
- chapter 3
- chapter 28
- section 35.1
- And if time... chapters 41 and 45