Introduction to Neural Networks
U. Minn. Psy 5038

Bayesian inference, graphical models

Initialize standard library files:

```
Off[General::spell1];
```

Last time

Basic rules of probability imply Bayes theorem

Basic rules of inference: Given a joint distribution, p(H,data), condition on what you know (product rule), and integrate out what you don't care about (sum rule).

The product rule implies Bayes rule:

p(H|data) = p(data|H)p(H)/p(data)

posterior = likelihood x prior x k

1/k can be found using the sum rule. If the data is given, and only H varies, then to infer H, we don't need to know k.

Rules look simple, but computations in real life problems get complicated quickly.

Multivariate distribution: conditioning, marginalizing, and sampling in Mathematica

Conditioning and marginalization preserve gaussianity.

Hand-constructing an energy function and update rule

If one understands the generative process, then you don't need learning. Instead construct an energy function, and use gradient descent to calculate an update rule.

Again we saw that minimizing energy can be viewed as maximizing *a posteriori* probability, given some nodes whose values are fixed by the data.

Examples illustrating smoothing given sparse data in early vision.

Demo of specular rigid motion.

Today

Bayesian inference and conjugate priors: coin flipping

Why study Bayesian inference for neural networks?

Rationale: The tools of probabilistic modeling can provide a deeper understanding of neural networks. They also provide quantitative descriptions at the "right" level of abstraction to model functional human behaviors. At the same time Bayesian models are extensible, in that there is a logical connection to neural network and circuitry models.

The problem

While Bayesian statistical methods are most commonly used for hypothesis testing given data, the Bayesian conceptual and mathematical "toolbox" is also useful for modeling human behavior. Applications include modeling perception, cognition, and motor behavior.

Here's perhaps the simplest example. Imagine asking someone the questions:

Given that you have observed 1 coin flip, say it comes up "heads", what is your best guess as to the coin's probability of coming up heads in all subsequent flips?

Given that you have observed 5 coin flips, and you see "heads" come up 3 times, what is your best guess as to the coin's probability of coming up heads in all subsequent flips?

We'll start of by using maximum likelihood estimation, and then go on to Bayesian MAP estimation.

When we are done, we'll have a model that can answer the question:

Given that you have observed k coin flips, what is your best guess as to the coin's bias?

and more generally:

Given that you have observed k outcomes of a binary process, what is your best guess as to the generative process, i.e. the underlying distribution?

One flip: Bernoulli Distribution & maximum likelihood

Let the probability of the coin generating heads be θ. And the probability of tails is 1-θ.

```
PDF[BernoulliDistribution[θ], k]
```

$$\begin{cases} 1-\theta & k == 0 \\ \theta & k == 1 \\ 0 & \text{True} \end{cases}$$

Let's define a random variable X = {0 or 1} ↔ {tails, heads}. The probability of X conditional on the probability that coin has a bias of favoring heads or tails, pc(x|p) can be written as:

```
pc[x_, θ_] := If[θ ≠ 1 , θ^x (1 - θ)^(1 - x), 1]
```

```
pc[x, θ]
```

$$\text{If}\left[\theta \neq 1, \theta^{x}\ (1-\theta)^{1-x}, 1\right]$$

Given an observation, say X = heads = 1, what is the probability that the coin will always generate heads? Let's use maximum likelihood estimation:

```
{pc[1, 0], pc[1, 1]}
```

$$\{0, 1\}$$

θ=1 (heads) is the most likely state of the coin.

Maximum likelihood says that, based on the data, we should guess that the coin will always produce heads. This strongly violates our intuition based on our experience with coins. We should get more data.

More flips: Binomial distribution & maximum likelihood

For a coin whose probability of heads is p, the binomial distribution gives the probability of getting k heads in n tosses:

$$\text{pb}\ (k \mid \theta, n) = \binom{n}{k} \theta^{k}\ (1-\theta)^{n-k}$$

where:

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

```
{{5}, {3}} == 10
```

is the binomial coefficient -- the number of combinations of "n take k".

Here's the Mathematica form:

```
PDF[BinomialDistribution[n, θ], k]
```

$$\begin{cases} (1-\theta)^{-k+n}\,\theta^{k}\,\text{Binomial}[n, k] & 0 \le k \le n \\ 0 & \text{True} \end{cases}$$

The probability of 3 heads in 5 throws would be:

```
PDF[BinomialDistribution[5, θ], 3]
```

$$10\ (1-\theta)^{2}\,\theta^{3}$$

The probability of k heads in n throws would be:

```
PDF[BinomialDistribution[n, θ], k]
```

$$\begin{cases} (1-\theta)^{-k+n}\,\theta^{k}\,\text{Binomial}[n, k] & 0 \le k \le n \\ 0 & \text{True} \end{cases}$$

Lets use maximum likelihood again. We can simplify by not considering the constant terms, and only the factors that depend on p:

$$pb(k \mid \theta, n) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \propto \theta^k (1-\theta)^{n-k}$$

To further simplify, we can look for the maximum in the Log of $\theta^k (1-\theta)^{n-k}$, rather than $\theta^k (1-\theta)^{n-k}$ itself.

```
PowerExpand[Log[θ^k (1 - θ)^(n-k)]]
```

```
(-k + n) Log[1 - θ] + k Log[θ]
```

Take the derivative and find the stationary point:

```
D[PowerExpand[Log[θ^k (1 - θ)^(n-k)]], θ]
```

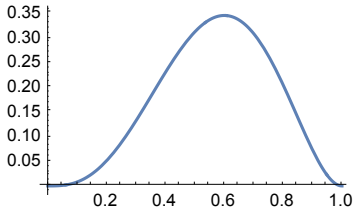
$$-\frac{-k+n}{1-\theta} + \frac{k}{\theta}$$

```
Solve[-(-k + n)/(1 - θ) + k/θ == 0, θ]
```

$$\left\{\left\{\theta \to \frac{k}{n}\right\}\right\}$$

For k =3 in n=5 tosses, $\theta \to 3/5 = 0.6$. we can also confirm by plotting:

```
Plot[PDF[BinomialDistribution[5, θ], 3], {θ, 0, 1}]
```



► 1. You can also use the general function FindMaximum[ ]:
FindMaximum[PDF[BinomialDistribution[5,θ],3],{θ,.41}]

```
{0.3456, {θ → 0.6}}
```

Let's again check our intuitions. Suppose 5 flips produced 3 heads. Should you guess that the coin always produce heads in the proportion $\theta = 3/5$? You probably wouldn't. Based on our experience, we believe that even before gathering data, a good guess would be that $\theta = 1/2$.

In contrast to maximum likelihood, a Bayesian MAP guess would update your prior belief about θ as data comes in.

Note that because we identify beliefs with probabilities, we are now talking about probabilities of probabilities. I.e. the prior probability, $p(\theta)$, and its posterior updates given data. Let's see how that works.

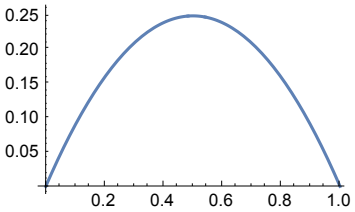
Beta distribution, Bayes & Maximum a posterior estimation

We need a prior probability distribution on probabilities θ. We could just assume $\theta = 1/2$, but we would still want an expression with parameters that could be conveniently updated as data comes in. After all, maybe the coin isn't symmetric. Or we'd like our theory to work with other binary processes in addition to coins.

Let's list some criteria for the general shape of the function: 1) The range of θ should be between 0 and 1; 2) We need some parameters to allow degrees of certainty in prior beliefs, and biases towards or away from $\theta = 0$ and 1; 3) We may want to avoid being dogmatic about $\theta = 1/2$ at the beginning (there are other binomial problems besides coin flipping), so certain parameter values should let $p(\theta)$ to be unity over the range $\theta \in [0,1]$ -- we may want to start with the principle of insufficient reason; 4) the expression should make the computations easy.

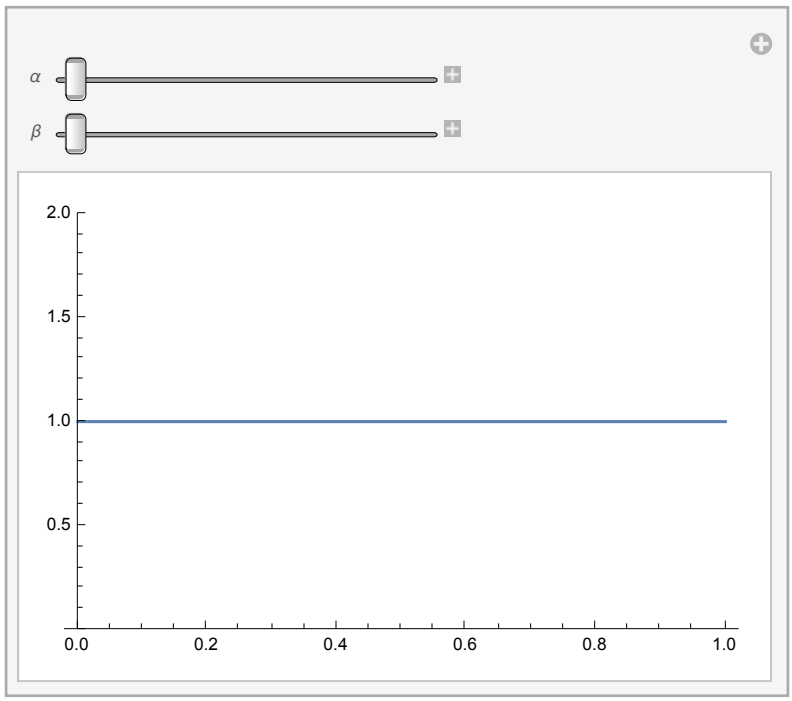
So here's one possibility:

```
Plot[θ * (1 - θ), {θ, 0, 1}]
```



This does reflect our intuitions about coins, but this doesn't satisfy 2) or 3). We can satisfy both 2) and 3) with a simple modification: $(1-\theta)^{-1+\beta}\,\theta^{-1+\alpha}$.

```
Manipulate[Plot[θ^(-1+α) * (1-θ)^(-1+β), {θ, 0, 1}, PlotRange → {0, 2}],
 {α, 1, 10}, {β, 1, 10}]
```



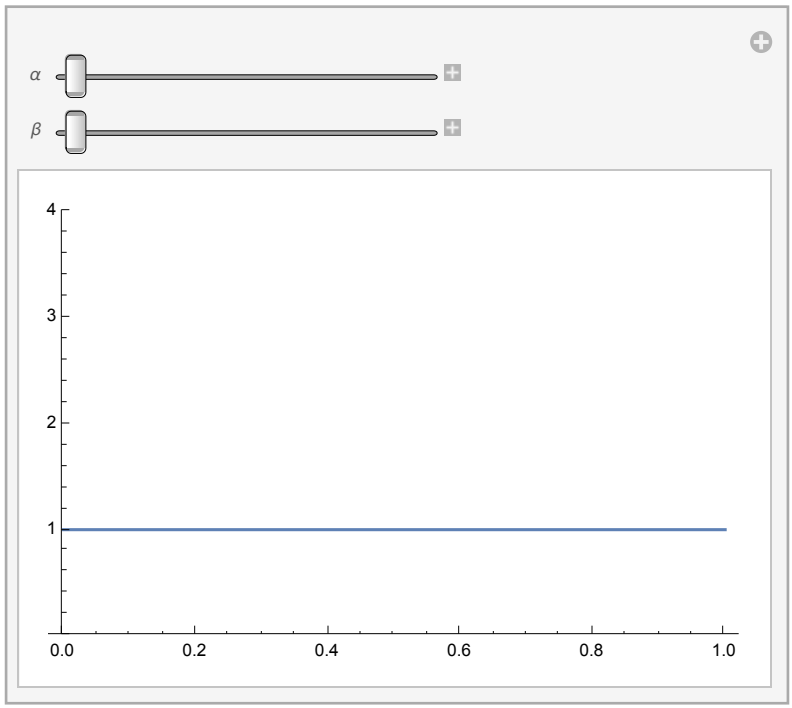
α and β determine the shape of the beta function.

What is missing is a normalizing constant so that $\int p(\theta)\,d\theta = 1$. The normalizing constant is Beta[α,β]. The Beta distribution does what we need:

```
Beta[α, β]
```

```
Beta[α, β]
```

```
Manipulate[Plot[PDF[BetaDistribution[α, β], θ], {θ, 0, 1}, PlotRange → {0, 4}],
 {α, 1, 10}, {β, 1, 10}]
```



Note: if α=β=1, we have a uniform prior: p(θ)=1. If α=β=2, we have a prior that favors θ=1/2. If α=β=5, the prior still favors θ=1/2, but with increased certainty, i.e. narrower width. You can think of α and β as ***1 plus "#pseudo-counts"*** of heads and tails, respectively. In other words, think of a prior as having seen say 10 heads and 10 tails (α=11,β =11) before gathering any "real" data.

If α<β, the mode favors θ<1/2, if α>β, the mode favors θ>1/2, so the expression can capture a bias either towards tails or heads.

Regarding 4), computational convenience, the Beta distribution is a good choice because it is the *conjugate distributio*n to the Binomial. Conjugate means that the posterior has the same mathematical form as the prior after updating, which keeps things simple. Such a prior is also called a conjugate prior for the likelihood function.

It isn't too hard prove that the beta function is the binomial's prior.

```
Clear[α, β];
PDF[BetaDistribution[α, β], θ]
```

$$\begin{cases} \frac{(1-\theta)^{-1+\beta}\,\theta^{-1+\alpha}}{\text{Beta}[\alpha,\beta]} & 0 < \theta < 1 \\ 0 & \text{True} \end{cases}$$

We don't need to worry about the normalization constant, which doesn't affect the shape of the distribution.

So note that the beta distribution is $\propto\ (1-\theta)^{-1+\beta}\,\theta^{-1+\alpha}$.

Suppose we've observed k heads in n counts, then the *binomial likelihood* is $\propto \theta^k (1-\theta)^{n-k}$.

So by Bayes, posterior $\propto$ prior x likelihood:

$$= (1-\theta)^{-1+\beta}\,\theta^{-1+\alpha} \text{ x } \theta^k (1-\theta)^{n-k} = (1-\theta)^{-1+\beta+n-k}\,\theta^{-1+\alpha+k} =$$

$$(1-\theta)^{(\beta+n-k)-1}\,\theta^{(\alpha+k)-1}$$

which has the same form as the prior. And after normalizing, it is again a Beta distribution.

Regarding computational convenience, note that after observing k heads in n tosses, the updated parameters are:

$\beta \to \beta$ + n - k = β + #tails
and
$\alpha \to \alpha$ + k = α + #heads

Assuming we have 0 pseudo-counts, our prior is α=α0=1; β = β0=1. We can see that updating is just a matter of incrementing the α and β parameters. So criterion 4) is satisfied.

Lets look at how the posterior distribution for the coin bias changes as we make progressively more coin flips, up to 5 total. Assume the specific case where we get 3 heads out of 5 tosses.

```
heads = {1, 0, 1, 1, 0}; (*3 heads out of 5 flips*)
tails = BitNot[{1, 0, 1, 1, 0}] + 2;

Accumulate[heads]
Accumulate[tails]
```

```
{1, 1, 2, 3, 3}
```

```
{0, 1, 1, 1, 2}
```

```
Clear[β1, α1];
α0 = 1;
β0 = 1;
β1[n_, k_] := β0 + n - k;
α1[k_] := α0 + k;

α2 = Table[α1[Accumulate[heads][[i]]], {i, 1, 5}]
```

```
{2, 2, 3, 4, 4}
```

```
β2 = Table[β1[i, Accumulate[heads][[i]]], {i, 1, 5}]
```

```
{1, 2, 2, 2, 3}
```

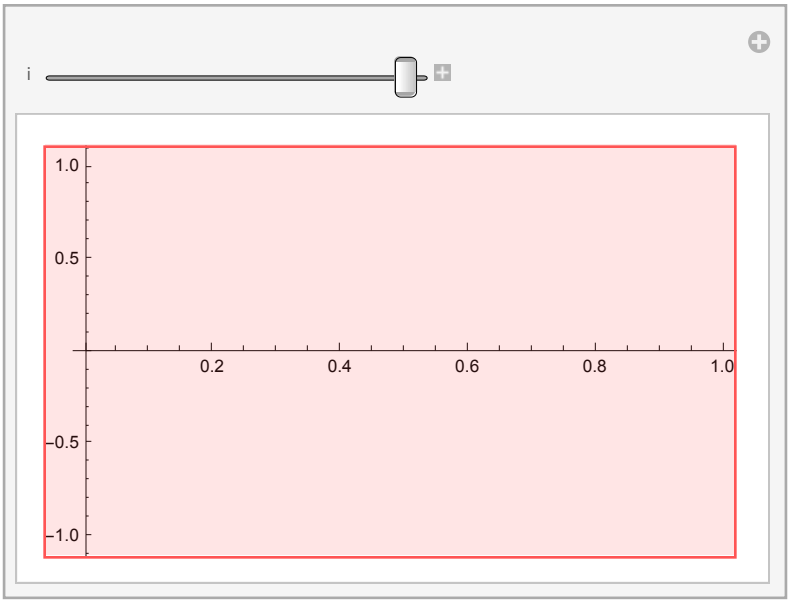
Insert the prior values of α and β, i.e. 1 and 1:

```
αl = Prepend[α2, α0]
βl = Prepend[β2, β0]
```

```
{1, 2, 2, 3, 4, 4}
```

```
{1, 1, 2, 2, 2, 3}
```

```
Manipulate[Plot[PDF[BetaDistribution[αl[[i]], βl[[i]]], x],
  {x, 0, 1}, Filling → Axis], {i, 1, 6, 1}]
```



Part: Part specification αl[[6]] is longer than depth of object.

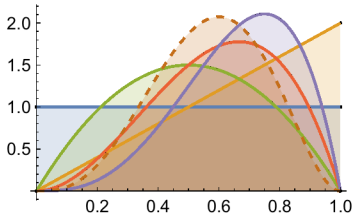
Part: Part specification βl[[6]] is longer than depth of object.

Part: Part specification αl[[6]] is longer than depth of object.

General: Further output of Part::partd will be suppressed during this calculation.

Here's a summary in one graph. The dashed line is the posterior after the last, 5th, coin toss.

```
Plot[Table[PDF[BetaDistribution[αl[[i]], βl[[i]]], x], {i, 1, 6}] // Evaluate,
  {x, 0, 1}, Filling → Axis, PlotStyle → {, , , , , Dashed}]
```



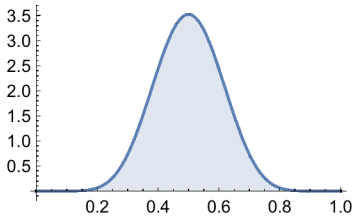
We can again use FindMaximum to calculate the mode, our best bet of the coin's θ:

```
FindMaximum[PDF[BetaDistribution[αl[[6]], βl[[6]]], x], {x, .41}]
```

```
{2.0736, {x → 0.6}}
```

If your pseudo-counts were 9 heads and 9 tails:

```
Plot[PDF[BetaDistribution[10, 10], x], {x, 0, 1}, Filling → Axis]
```



► 2. Use the above Manipulate[] function to show how the posterior evolves if the pseudo-counts are 9 and 9, for heads and tails respectively. What if the the pseudo-counts were 90 and 90?

Introduction to Graphical Models

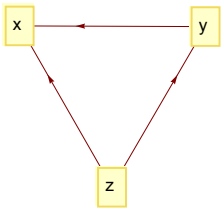
From joint distributions to graphs and back

The coin model was extremely simple, only involving a single scalar variable to be estimated and a single, scalar piece of data. Graphical models provide a powerful tool to characterize higher dimensional joint distributions. Suppose we have a joint probability: p(x,y,z). We can use the product rule to factor the joint into conditionals expressing the dependence of each variable on the others:

p(x,y,z) = p(x,y | z) p(z) = p(x | y, z) p(y|z) p(z).

This factorization suggests a graph connecting the variables, and with arrows showing conditional dependence:

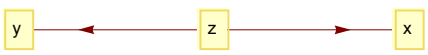
```
GraphPlot[{"z" → "y", "z" → "x", "y" → "x"},
 VertexLabeling → True, DirectedEdges → True]
```



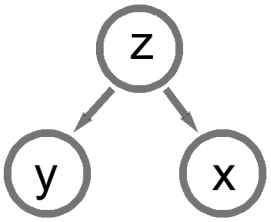
Of course in this example, the order of the variables in the joint doesn't matter, so by symmetry, we could have started with x (or y) and factor the distribution as: p(z|y,x) p(y|x) p(x), etc.. So far, the graph hasn't bought us anything. Any distribution that is a function of three variables could be described by this graph. But suppose, we have a different joint with the factorization:

p(x,y,z) = p(x | z) p(y|z) p(z), then the graph looks like:

```
GraphPlot[{"z" → "y", "z" → "x"},
 VertexLabeling → True, DirectedEdges → True, SelfLoopStyle → .1]
```



Usually portrayed as:



The fact that there is a missing link--no arrow between y and x--tells us something--it restricts the space of probabilities. Given a fixed value of z,

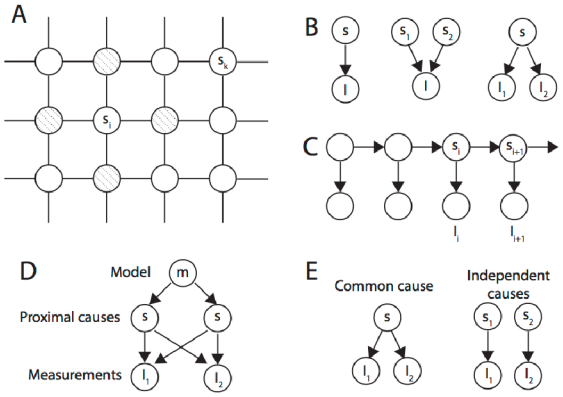
$$p(x,y|z) = p(x \mid z)\, p(y \mid z).$$

In other words given z, x and y are independent. x and y are said to be "*conditionally independent*" given z. In the language of graphical models, z is the *parent*, and x and y are "children".

Here's an example of conditional independence to help build intuition. When corn prices drop in the summer, hay fever (allergies) incidence goes up. However, if the joint on corn price and hay fever is conditioned on ``ideal weather for corn and ragweed'', the correlation between corn prices and hay fever drops. This is because given ideal weather, corn price and hay fever symptoms are conditionally independent--knowing corn price doesn't help predict hay fever incidence.

Causal structure and conditional independence

To recap, in general we'd like to be able to specify natural patterns such as images by a high-dimensional joint probability. This is usually too difficult in practice, and requires simplification and approximations. If we have some idea of the conditional relationships between various causes and data, this knowledge can be represented in a directed graph, which might be much more complicated than those above. The idea is to represent the probabilistic structure of the joint distribution by a graphical model that expresses how variables influence each other. Random variables are represented by nodes, and the links or arrows between the nodes represent conditional relationships.



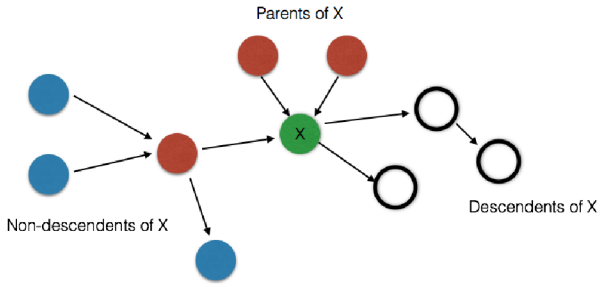
The above figure illustrates several types of graphical structures. A. undirected graph doesn't have arrows. This is called a Markov Random Field. B-E illustrate directed graphs: B. Simple influence

relationships. C. Markov process. D. Simple hierarchical model. The "hidden" variables, s are sometimes called "latent variables". E. Common vs. independent causes (cf. Körding et al.).

Undirected graphs, as in A, can be used to represent the dependencies that we saw in the Boltzmann machine, where the influence relationship between any two units is symmetric. The graph above would correspond to a network of random variables where each node is conditionally dependent only on its four nearest neighbors. C is used for processes that evole in time, where arrows make sense. It also illustrates a Markov property - the state of any node s is conditionally dependent only on the state of the previous node: $p(S_i \mid S_1, S_2, ..., S_{i-1}) = p(S_i \mid S_{i-1})$.

The Markov assumption greatly simplifies computations. We are going to use this Markov property (in an acausal way) in the next lecture when we revisit smoothing in the context of Belief Propagation.

While not restricted to causal interpretations, the arrows in directed graphs are often based on assumptions about what causes what (Pearl, 1988; 1996).

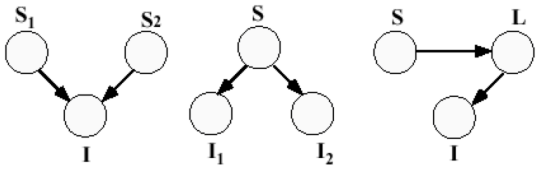


The directed acyclic graph above represents the case where X is independent of non-descendants, given parents. This is an example of a causal Markov condition. X has two children, but three descendants.

Directed graphs can be acyclic or cyclic (i.e. the arrows loop). It can be shown that probabilistic computations, e.g. involving conditioned variables, and marginalization, on directed acyclic graphs converge, whereas there is no guarantee for cyclic graphs.

In a later lectures, we will look at how one can update the probability distributions at each of the nodes through belief propagation. We will also see how to generate samples on potentially complex graphs (using MCMC), samples which can then be used to estimate means and modes of node values.

► 3. Given the graphs below, write out factorizations of the joint distributions.



The arrows tell us how to factor the joint probability into conditionals. So for the three examples above, we have:

p(S1, S2, I) = p(I | S1, S2) p(S1) p(S2)
p(S, I1, I2) = p(I1 | S) p(I2 | S) p(S)
p(S, L, dI) = p(I | L) p(L|S) p(S)

► 4. You might have heard the saying “correlation is not causation”. There is a correlation between eating ice cream and drowning. Why? What event could you condition on to make the dependence go away?

Marginalization

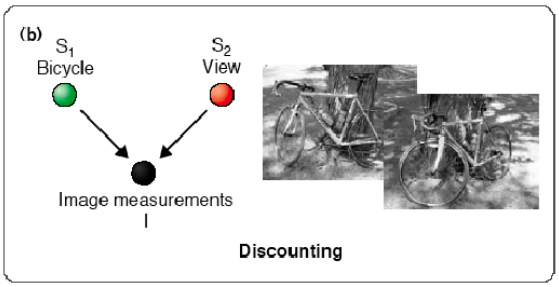
In general, there are going to be some random variables that we want to estimate, some that get fixed by data (or by other processes), and some variables that we don’t care about so need to be integrated out. The next section illustrates these with some examples from perceptual processes.

Graphical models and problems of perception

The color code is: green means we want to estimate these values, black means fixed data, and red means we want to integrate them out.

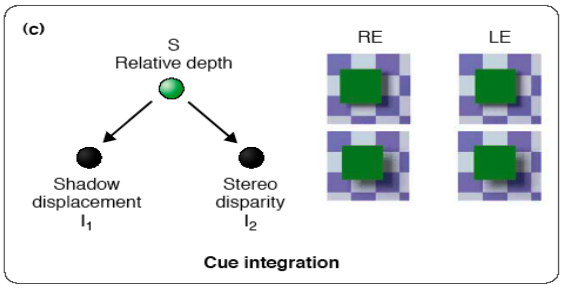
Discounting

This is the core problem of object recognition. Discounting is the flip side of invariance.



$p(S_1 \mid I) = \sum_{S_2} p(S_2, S_1 \mid I)$

Cue integration

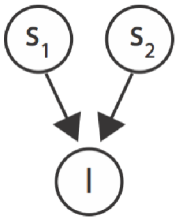


$p(S \mid I_1, I_1), = p(S_2, S_1, I) / p(I_1, I_2)$

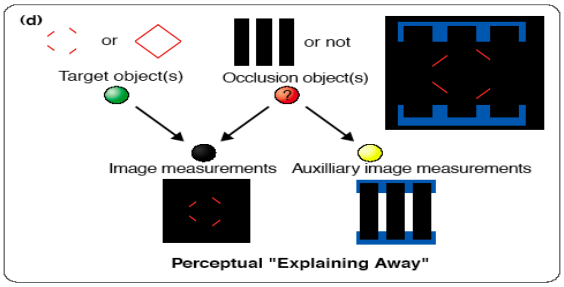
Here two measurements (shadow displacement and stereo disparity) may be correlated. However, if S is fixed, then they become *conditionally independent*. This is the basis for optimal cue weighting. There is a large literature showing that human perception often (but not always) approximates optimal cue weighting.

Explaining away and conditional dependence

An simple but interesting graph with arrows converging on to one node is characteristic of the common occurrance of two competing or interacting causes, S_1 and S_2 which could both influence the value of I.



"Explaining away" is a characteristic of many perceptual inferences, for example when there are alternative perceptual groupings consistent with a set of identical or similar sets of local image features (Kersten, 2003). In theory, explaining away can be accomplished through a generative process (e.g. recall the Boltzmann machine). This kind of inference may have a consequence on measurable neural activity (Murray et al., 2002).



In general, human reasoning is particularly good at "explaining away" inferences (Pearl, 1988).

Explaining away as *conditional dependence*: Another coin flipping example

"Explaining away" is a general phenomenon that occurs in probabilistic belief networks in which two (or more) variables influence a third variable whose value can be measured. Once measured, it provides evidence to infer the values of the influencing variables. Imagine two coins that can be flipped independently, and the results (S_1 = heads or tails, and S_2 = heads or tails) have an influence on a third variable, I. For concreteness, assume the third variable's value is 1 if both coins agree, and 0 if not (NOT-XOR). If we are ignorant of the value of the third variable, knowledge of one influencing variable doesn't help to guess the value of the other—the two coin variable probabilities are independent. (This is called marginal independence, "marginal" with respect to the third variable.) But if the value of the third variable is measured (suppose it is 1), the two coin variables become coupled, and they are said to be conditionally dependent. Now knowing that one coin is heads guarantees that the other one is too. The phrase "explaining away" arises because coupling of variables through shared evidence arises often in human reasoning, when the influences can be viewed as competing causes. Suppose that the evidence is 0. If our interpretation is that "heads" in either coin can cause such a "suppression" of the NOT-XOR output, then which coin did the suppressing? One of the coins is heads and one tails, but not both. Any auxiliary evidence that tips the balance toward one coin being "to blame", reduces our belief that the other caused the observed 0. The other coin's possible influence is explained away by the new evidence supporting the true-culprit coin's value of heads.

Optimal Inference and task dependence

Recall the question: What is noise? Noise could be defined as those variables you don't care to estimate, but contributes to the data. These variables are also called nuisance variables or confounding

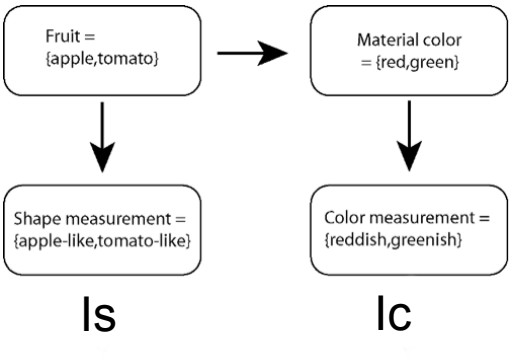
variables.

But changes in the task can determine what is important and what is not. So what is “noise” for one task, is “signal” for another. Yuille and Kersten describe variables as primary and secondary in the context of visual inference. If the task doesn’t require you to estimate a random variable (secondary), then it gets integrated out of the joint distribution to produce a marginal distribution over the variable that you are interested in (primary).

One counter-intuitive consequence of this is that the most probable (MAP) estimates are not necessarily consistent across tasks. This is illustrated in the next two sections.

Fruit example

(due to James Coughlan; see Yuille, Coughlan, Kersten & Schrater).



The graph specifies how to decompose the joint probability:
p[F, C, Is, Ic] = p[Ic | C] p[C | F] p[Is | F] p[F]

Generative model: The prior model on hypotheses, F & C

More apples (F=1) than tomatoes (F=2), and:

```
ppF[F_] := If[F == 1, 9 / 16, 7 / 16];
TableForm[Table[ppF[F], {F, 1, 2}], TableHeadings -> {{"F=a", "F=t"}}]
```

F=a	$\frac{9}{16}$
F=t	$\frac{7}{16}$

The conditional probability **cpCF[C|F]**:

```
cpCF[F_, C_] := Which[F == 1 && C == 1, 5 / 9,
   F == 1 && C == 2, 4 / 9, F == 2 && C == 1, 6 / 7, F == 2 && C == 2, 1 / 7];
TableForm[Table[cpCF[F, C], {C, 1, 2}, {F, 1, 2}],
 TableHeadings -> {{"C=r", "C=g"}, {"F=a", "F=t"}}]
```

	F=a	F=t
C=r	$\frac{5}{9}$	$\frac{6}{7}$
C=g	$\frac{4}{9}$	$\frac{1}{7}$

The above conditional is a probability distribution on **C**. So would you expect the sum over a row to be 1? Over a column?

So by the product rule the joint is:

```
jpFC[F_, C_] := cpCF[F, C] ppF[F];
TableForm[Table[jpFC[F, C], {F, 1, 2}, {C, 1, 2}],
 TableHeadings -> {{"F=a", "F=t"}, {"C=r", "C=g"}}]
```

	C=r	C=g
F=a	$\frac{5}{16}$	$\frac{1}{4}$
F=t	$\frac{3}{8}$	$\frac{1}{16}$

Note that now all the entries sum to 1.

We can marginalize to get the prior probability on color alone:

$$\text{ppC[C\_] := } \sum_{F=1}^{2} \text{jpFC[F, C]}$$

- 5. Which color is a priori more probable?
- 6. Is fruit identity independent of material color--i.e. is F independent of C?

Check whether the joint probability on Fruit and Color can be factored into the product of the prior probabilities on Fruit and Color.

Answer

No.

```
TableForm[Table[jpFC[F, C], {F, 1, 2}, {C, 1, 2}],
 TableHeadings -> {{"F=a", "F=t"}, {"C=r", "C=g"}}]
TableForm[Table[ppF[F] ppC[C], {F, 1, 2}, {C, 1, 2}],
 TableHeadings -> {{"F=a", "F=t"}, {"C=r", "C=g"}}]
```

	C=r	C=g
F=a	$\frac{5}{16}$	$\frac{1}{4}$
F=t	$\frac{3}{8}$	$\frac{1}{16}$

	C=r	C=g
F=a	$\frac{99}{256}$	$\frac{45}{256}$
F=t	$\frac{77}{256}$	$\frac{35}{256}$

Generative model: The likelihood model--probabilities of measurements, i.e. some features given hypotheses

Suppose that we have gathered some "image statistics" which provides us knowledge of how the image measurements for shape (Is), and for color (Ic) depend on the type of fruit F, and material color, C. For simplicity, our measurements are discrete and binary , say Is = {am, tm}, and Ic = {rm, gm}. (In a more realistic case, they would have continuous values).

P(I_S= am, tm | F=a) = {11/16, 5/16}
P(I_S= am, tm | F=t) = {5/8, 3/8}
P(I_C= rm, gm | C=r) = {9/16, 7/16}
P(I_C= rm, gm | C=g) = {1/2, 1/2}
We use the notation am, tm, rm, gm because the measurements are already suggestive of the likely cause. So there is a correlation between apple and apple-like shapes, am; and between red material, and "red" measurements, rm.
(This may sound too artificial, but a red apple can result in a greenish measurement if illuminated with a greenish light.)
In general, there may not be an obvious correlation like this.

We define a function for the probability of Ic given C, **cpIcC[Ic | C]**:

```
cpIcC[Ic_, C_] := Which[Ic == 1 && C == 1, 9 / 16,
   Ic == 1 && C == 2, 7 / 16, Ic == 2 && C == 1, 1 / 2, Ic == 2 && C == 2, 1 / 2];
TableForm[Table[cpIcC[Ic, C], {C, 1, 2}, {Ic, 1, 2}],
 TableHeadings -> {{"Ic=rm", "Ic=gm"}, {"C=r", "C=g"}}]
```

	C=r	C=g
Ic=rm	$\frac{9}{16}$	$\frac{1}{2}$
Ic=gm	$\frac{7}{16}$	$\frac{1}{2}$

The probability of Is conditional on F is **cpIsF[Is | F]**:

```
cpIsF[Is_, F_] := Which[Is == 1 && F == 1, 11 / 16,
   Is == 1 && F == 2, 5 / 8, Is == 2 && F == 1, 5 / 16, Is == 2 && F == 2, 3 / 8];
TableForm[Table[cpIsF[Is, F], {Is, 1, 2}, {F, 1, 2}],
 TableHeadings -> {{"Is=am", "Is=tm"}, {"F=a", "F=t"}}]
```

	F=a	F=t
Is=am	$\frac{11}{16}$	$\frac{5}{8}$
Is=tm	$\frac{5}{16}$	$\frac{3}{8}$

The total joint probability

We now have enough information to put probabilities on the 2x2x2 "universe" of possibilities, i.e. all possible combinations of fruit, color, and image measurements. Looking at the graphical model makes it easy to use the product rule to construct the total joint, which is:
p[F, C, Is, Ic] = p[Ic | C] p[C | F] p[Is | F] p[F]:

```
jpFCIsIc[F_, C_, Is_, Ic_ ] := cpIcC[ Ic, C ] cpCF[F, C ] cpIsF[Is, F ] ppF[F ]
```

Usually, we don't need the probabilities of the image measurements (because once the measurements are made, they are fixed and we want to compare the probabilities of the hypotheses. But in our simple case here, once we have the joint, we can calculate the probabilities of the image measurements through marginalization p(Is,Ic)=$\sum_C \sum_F p(F, C, \text{Is}, \text{Ic})$, too:

```
jpIsIc[Is_, Ic_] := Sum[jpFCIsIc[F, C, Is, Ic ], {C, 1, 2}, {F, 1, 2}]
```

Three MAP tasks

We are going to show that the best guess depends on the task.
In other words, given measurements Is, Ic, the most probable choices of fruit and/or color depend on what which combination of hypotheses we care about.

Define argmax[] function:

```
argmax[x_] := Position[x, Max[x]];
```

This returns the position of the biggest value in a list.

TASK 1: Pick most probable fruit AND color--Answer "red tomato"

We are given some data--i.e. values of Is and Ic and want to draw some conclusions about what kind of fruit we are looking at, and its material color. First, suppose the task is to make the best bet as to the fruit AND material color from the measurements given. To make it concrete, suppose that we see an "apple-ish shape" with a reddish color, i.e., we measure Is=am=1, and Ic = rm=1. The measurements suggest "red apple", but to find the most probable, we need to take into account the priors too in order to make the best guess.
p(F,C | Is, Ic) is given by:

```
FCcIsIc[F_, C_, Is_, Ic_] := jpFCIsIc[F, C, Is, Ic ] / jpIsIc[Is, Ic]
```

```
TableForm[FCcIsIcTable = Table[FCcIsIc[F, C, 1, 1 ], {F, 1, 2}, {C, 1, 2}],
 TableHeadings -> {{"F=a", "F=t"}, {"C=r", "C=g"}}]
Max[FCcIsIcTable]
argmax[FCcIsIcTable]
```

	C=r	C=g
F=a	$\frac{55}{157}$	$\frac{308}{1413}$
F=t	$\frac{60}{157}$	$\frac{70}{1413}$

$$\frac{60}{157}$$

```
{{2, 1}}
```

So looking at the table we can see that our best bet is “red tomato”.

TASK 2: Pick most probable color--Answer "red"

Same measurements as before. But now suppose we only care about the true material color, and not the identity of the object. Then we want to integrate out or marginalize with respect to the shape or fruit-type variable, F. In this case, we want to maximize the posterior:

p(C | Is=1, Ic=1)=$\sum_{F=1}^{2} p(F, C \mid \text{Is} = 1, \text{Ic} = 1)$

```
pC[C_, Is_, Ic_] := Sum[FCcIsIc[F, C, Is, Ic ], {F, 1, 2}]
```

```
pCTable = Table[pC[C, 1, 1], {C, 1, 2}];
TableForm[pCTable, TableHeadings → {{"C=r", "C=g"}}]
Max[pCTable]
argmax[pCTable]
```

C=r	115/157
C=g	42/157

$$\frac{115}{157}$$

{{1}}

Answer is that the most probable material color is C = r, "red".

TASK 3: Pick most probable fruit--Answer "apple"

Same measurements as before, Is=am=1, and Ic = rm=1. But now, we don't care about the material color, just the identity of the fruit. Then we want to integrate out or marginalize with respect to the material variable, C. In this case, we want to maximize the posterior:

p(F | Is, Ic)

$$\text{pF[F\_, Is\_, Ic\_]} := \sum_{C=1}^{2} \text{FCcIsIc[F, C, Is, Ic]}$$

```
pFTable = Table[pF[F, 1, 1 ], {F, 1, 2}];
TableForm[pFTable, TableHeadings -> {{"F=a", "F=t"}}]
Max[pFTable]
argmax[pFTable]
```

F=a	803/1413
F=t	610/1413

$$\frac{803}{1413}$$

{{1}}

The answer is "apple". So to sum up, for the same data measurements, the most probable fruit AND color is "red tomato", but the most probable fruit is "apple"!

Important "take-home message": *Optimal inference depends on the precise definition of the task*

Further, the above example shows that that the most probable answer from a marginal does not have to be consistent with the most probable answer from the joint distribution.

- 7. Try expressing the task-dependent consequences using the frequency interpretation of probability.
- 8. How would you sample from the fruit model?

► 9. Coin toss example

Consider seven of tosses of two coins that are strangely coupled: events[coin1,coin2] ={{1,0},{1,0},{0,1},{0,1}{0,1},{1,1},{1,1}}, where 1 is heads, and 0 is tails. Suppose these frequencies represent the true probabilities of heads (1) and tails (0) for each of the coins, *coin1* and *coin2*. In other words, if we repeat the tossing experiment of these two coins, millions of times, we'd find that the probabilities of the joint events occurs exactly as their frequency of occurrence in the first seven tosses.

What is the most probable value (heads or tails) for coin 1?
What is the most probable value (heads or tails) for coin 2?
What is the most probable value of the joint occurrence {coin1,coin2}?

```
events = {{1, 0}, {1, 0}, {0, 1}, {0, 1}, {0, 1}, {1, 1}, {1, 1}};
coin1 = events [[ ;; , 1]]; coin2 = events [[ ;; , 2]];
```

Calculate the probabilities of tosses of coin1 and coin2 coming up heads:

```
{"coin1:", TableForm[{Count[coin1, 1] / 7, Count[coin1, 0] / 7},
  TableHeadings -> {{"H=heads", "H=tails"}}], "    ",
 "coin2:", TableForm[{Count[coin2, 1] / 7, Count[coin2, 0] / 7},
  TableHeadings -> {{"H=heads", "H=tails"}}]}
```

$$\left\{\text{coin1:},\ \begin{array}{c|c} \text{H=heads} & \frac{4}{7} \\ \text{H=tails} & \frac{3}{7} \end{array},\quad \text{coin2:},\ \begin{array}{c|c} \text{H=heads} & \frac{5}{7} \\ \text{H=tails} & \frac{2}{7} \end{array}\right\}$$

We can see from above that the most probable value of coin1 is heads. And the same for coin2.

Now lets calculate the joint probabilities:

```
TableForm[{Count[events, {1, 0}] / 7,
  Count[events, {0, 1}] / 7, Count[events, {1, 1}] / 7, 0}, TableHeadings ->
  {{"H={heads,tails}", "H={tails,heads}", "H={heads,heads}", "H={tails,tails}"}}]
```

H={heads,tails}	$\frac{2}{7}$
H={tails,heads}	$\frac{3}{7}$
H={heads,heads}	$\frac{2}{7}$
H={tails,tails}	0

Observe that {tails, tails} never occurs. The most probable event is: "H={coin1=tails, coin2=heads}".

The main point is that you can't predict the most probable value of coin1 from the most probable value (i.e. event) of the joint.

► 10. Do the calculation again, starting from the joint distribution as an exercise in marginalization,

```
Transpose[events] // MatrixForm
```

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

```
joint = {{2, 2}, {3, 0}} / 7
```

$$\left\{\left\{\frac{2}{7}, \frac{2}{7}\right\}, \left\{\frac{3}{7}, 0\right\}\right\}$$

```
TableForm[joint,
 TableHeadings -> {{"coin1=heads", "coin1=tails"}, {"coin2=heads", "coin2=tails"}}]
```

	coin2=heads	coin2=tails
coin1=heads	$\frac{2}{7}$	$\frac{2}{7}$
coin1=tails	$\frac{3}{7}$	0

Most probable value of {coin1,coin2} ={tails,heads}.

Now lets sum over coin2, i.e. over the rows, to get the conditional on coin1:

```
Apply[Plus, joint]
```

$$\left\{\frac{5}{7}, \frac{2}{7}\right\}$$

The most probable value is heads.

Appendix

Exercises

► 11. Question: Is **p(F, C, Is=1, Ic=1)** a joint probability on F and C? (Answer: No.)

Redefine **jpFCcIsIcTable** and **jpFCIsIcTable**, without the TableForm[] wrapper and calculate:

```
Total[jpFCcIsIcTable, {2}]
Total[jpFCIsIcTable, {2}]
```

$$\left\{\frac{803}{4096}, \frac{305}{2048}\right\}$$

$$\left\{\frac{803}{4096}, \frac{305}{2048}\right\}$$

► 12. **Note that we don't always have to calculate the conditional. For example in Task 1 above.**

We can use the fact that p(F,C | Is, Ic) = $\frac{p(F,C,\text{Is},\text{Ic})}{p(\text{Is},\text{Ic})} \propto$ p(F, C , Is=1, Ic=1). I.e. the conditional is proportional to the function specified by the joint evaluated at Is=1, and Ic=1.

```
jpFCcIsIcTable = Table[jpFCIsIc[F, C, 1, 1], {F, 1, 2}, {C, 1, 2}];
TableForm[jpFCIsIcTable, TableHeadings -> {{"F=a", "F=t"}, {"C=r", "C=g"}}]
Max[jpFCIsIcTable]
argmax[jpFCIsIcTable]
```

	C=r	C=g
F=a	$\frac{495}{4096}$	$\frac{77}{1024}$
F=t	$\frac{135}{1024}$	$\frac{35}{2048}$

$$\frac{135}{1024}$$

$$\{\{2, 1\}\}$$

Using *Mathematica* lists to manipulate discrete priors, likelihoods, and posteriors

A note on list arithmetic

We haven't done standard matrix/vector operations above to do conditioning. We've take advantage of how *Mathematica* divides a 2x3 array by a 2-element vector:

```
M=Array[m,{2,3}]
X = Array[x,{2}]
```

$$\begin{pmatrix} m(1,1) & m(1,2) & m(1,3) \\ m(2,1) & m(2,2) & m(2,3) \end{pmatrix}$$

$$\{x(1), x(2)\}$$

```
M/X
```

$$\begin{pmatrix} \frac{m(1,1)}{x(1)} & \frac{m(1,2)}{x(1)} & \frac{m(1,3)}{x(1)} \\ \frac{m(2,1)}{x(2)} & \frac{m(2,2)}{x(2)} & \frac{m(2,3)}{x(2)} \end{pmatrix}$$

Putting the probabilities back together again to get the joint

```
Transpose[Transpose[pHx] px]
```

$$\left(\text{px pHx}^{\mathsf{T}}\right)^{\mathsf{T}}$$

```
pxH pH
```

$$\text{pH pxH}$$

Getting the posterior from the priors and likelihoods:

One reason Bayes' theorem is so useful is that it is often easier to formulate the likelihoods (e.g. from a causal or generativemodel of how the data could have occurred), and the priors (often from heuristics, or in computational vision empirically testable models of the external visual world). So let's use *Mathematica* to derive **p(H|x)** from **p(x|H**) and **p(H)** , (i.e. pHx from pxH and pH).

```
px2 = Plus @@ (pxH pH)
```

$$pH + pxH$$

```
Transpose[Transpose[(pxH pH)] / Plus @@ (pxH pH)]
```

$$\frac{(pH\ pxH)^{T\,T}}{pH + pxH}$$

Show that this joint probability has a uniform prior (i.e. both priors equal).

```
p = {{1 / 8, 1 / 8, 1 / 4}, {1 / 4, 1 / 8, 1 / 8}}
```

$$\begin{pmatrix} \frac{1}{8} & \frac{1}{8} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{pmatrix}$$

Marginalization and conditioning: A small dimensional example using list manipulation in *Mathematica*

A discrete joint probability

All of our knowledge regarding the signal discrimination problem can be described in terms of the joint probability of the hypotheses, **H** and the possible data measurements, **x.** The probability function assigns a number to all possible combinations:

p[H, x]

That is, we are assuming that both the hypotheses and the data are discrete random variables.

$$H = \begin{cases} S1 \\ S2 \end{cases}$$

```
x ∈ {1, 2, ...}
```

Let's assume that x can only take on one of three values, 1, 2, or 3. And suppose the joint probability is:

```
p = {{1/12, 1/12, 1/6}, {1/3, 1/6, 1/6}}
```

$$\begin{pmatrix} \frac{1}{12} & \frac{1}{12} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$$

```
TableForm[p, TableHeadings -> {{"H=S1", "H=S2"}, {"x=1", "x=2", "x=3"}}]
```

	x=1	x=2	x=3
H=S1	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{6}$
H=S2	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$

The total probability should sum up to one. Let's test to make sure. We first turn the list of lists into a singel list of scalars using **Flatten[]**. And then we can sum either with **Apply[Plus,Flatten[p]].**

```
Plus @@ Flatten[p]
```

1

We can pull out the first row of p like this:

```
p[[1]]
```

$\left\{\frac{1}{12}, \frac{1}{12}, \frac{1}{6}\right\}$

$\left\{\mathbf{\frac{1}{12}}, \mathbf{\frac{1}{12}}, \mathbf{\frac{1}{6}}\right\}$

$\left\{\frac{1}{12}, \frac{1}{12}, \frac{1}{6}\right\}$

Is this the probability of x? No. For a start, the numbers don't sum to one. But we can get it through the two processes of marginalization and conditioning.

Marginalizing

What are the probabilities of the data, p(x)? To find out, we use the *sum rule* to sum over the columns:

```
px = Apply[Plus, p]
```

$\left\{\frac{5}{12}, \frac{1}{4}, \frac{1}{3}\right\}$

"Summing over "is also called **marginalization** or **"integrating out".** Note that marginalization turns a probability function with higher degrees of freedom into one of lower degrees of freedom.

What are the prior probabilities? p(H)? To find out, we sum over the rows:

```
pH = Apply[Plus, Transpose[p]]
```

$\left\{\frac{1}{3}, \frac{2}{3}\right\}$

Conditioning

Now that we have the marginals, we can get use the *product rule* to obtain the conditional probability through conditioning of the joint:

$$p[x \mid H] = \frac{p[H,\ x]}{p[H]}$$

In the Exercises, you can see how to use *Mathematica* to do the division for conditioning. The syntax is simple:

```
pxH = p / pH
```

$\begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$

Note that the probability of x conditional on H sums up to 1 over x, i.e. each row adds up to 1. But, the columns do not. **p[x|H]** is a **probability** function of x, but a **likelihood** function of H. The posterior

probability is obtained by conditioning on x:

$$p[H \mid x] = \frac{p[H,\ x]}{p[x]}$$

Syntax here is a bit more complicated, because the number of columns of px don't match the number of rows of p. We use Transpose[] to exchange the columns and rows of p before dividing, and then use Transpose again to get back the 2x3 form:

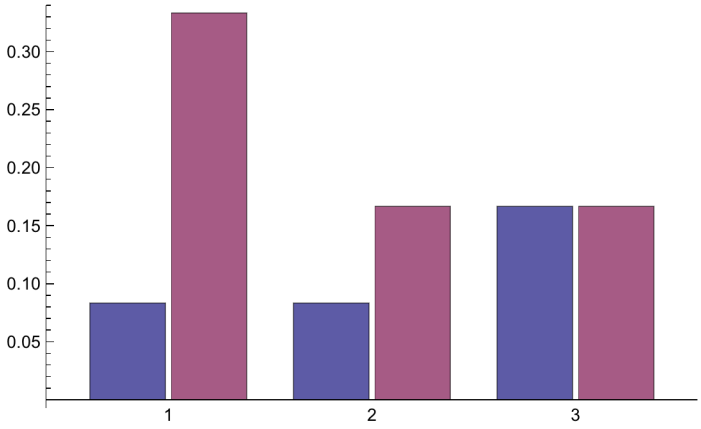
```
pHx = Transpose[Transpose[p] / px]
```

$$\begin{pmatrix} \frac{1}{5} & \frac{1}{3} & \frac{1}{2} \\ \frac{4}{5} & \frac{2}{3} & \frac{1}{2} \end{pmatrix}$$

Plotting the joint

The following BarChart[] graphics function requires in add-in package (**<< Graphics`Graphics`**), which is specified at the top of the notebook. You could also use **ListDensityPlot[]**.

```
BarChart[p[[1]], p[[2]]]
```



References

Links

Regarding conjugate priors, Joaquin Quiñonero-Candela and Carl Edward Rasmussen provide an excellent introduction to the problem as a background to general Bayesian inference:

http://mlg.eng.cam.ac.uk/teaching/4f13/1112/lect06.pdf

See too:
https://en.wikipedia.org/wiki/Bernoulli_process
https://en.wikipedia.org/wiki/Checking_whether_a _coin _is _fair

David MacKay (2002) Belgian euro coins: 140 heads in 250 tosses - suspicious?
http://www.inference.phy.cam.ac.uk/mackay/euro.pdf
http://www.inference.phy.cam.ac.uk/mackay/abstracts/euro.html

For python users, the book:
Probabilistic Programming and Bayesian Methods for Hackers. Provides python notebook introduc-

tions to Bayes methods using the PyMC toolbox, including the coin flipping problem.
http://nbviewer.jupyter.org/github/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/blob/master/Chapter1_Introduction/Ch1_Introduction_PyMC3.ipynb

For an excellent video introduction to Graphical models, see Christopher Bishop's lecture: https://youtu.be/ju1Grt2hdko?list=FLYfcFZYkzFbD3SoeOFvLatQ

Books and papers

Applebaum, D. (1996). Probability and Information . Cambridge, UK: Cambridge University Press.
Cover, T. M., & Joy, A. T. (1991). Elements of Information Theory. New York: John Wiley & Sons, Inc.
Duda, R. O., & Hart, P. E. (1973). Pattern classification and scene analysis . New York.: John Wiley & Sons.
Golden, R. (1988). A unified framework for connectionist systems. Biological Cybernetics, 59, 109-120.
Kersten, D. and P.W. Schrater (2000), Pattern Inference Theory: A Probabilistic Approach to Vision, in Perception and the Physical World, R. Mausfeld and D. Heyer, Editors. , John Wiley & Sons, Ltd.: Chichester. (pdf)
Kersten, D., Mamassian P & Yuille A (2004) Object perception as Bayesian inference. Annual Review of Psychology. (pdf, http://arjournals.annualreviews.org/doi/pdf/10.1146/annurev.psych.55.090902.142005)
Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. Current Opinion in Neurobiology, 13(2) http://gandalf.psych.umn.edu/~kersten/kersten-lab/papers/KerstenYuilleApr2003.pdf
Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal Inference in Multisensory Perception. PLoS ONE, 2(9), e943. http://doi.org/10.1371/journal.pone.0000943.t001
Murphy, K. P. (2012). Machine Learning: a Probabilistic Perspective. MIT Press.
Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (1st ed.). Morgan Kaufmann.
Pearl, J. (1996). Structural and probabilistic causality. Psychology of Learning and Motivation, 34, 393–435.
Ripley, B. D. (1996). Pattern Recognition and Neural Networks. Cambridge, UK: Cambridge University Press.
Yuille, A., Coughlan J., Kersten D.(1998) (pdf)
Zoran, D., & Weiss, Y. (2011). From learning models of natural image patches to whole image restoration, 479–486.
http://www.cs.huji.ac.il/~daniez/EPLLICCVCameraReady.pdf