Bayesian Theory

— Lecture Notes —

Laurenz Wiskott
Institut für Neuroinformatik
Ruhr-Universität Bochum, Germany, EU

14 December 2016

— Summary —

Bayesian theory deals with probabilities of discrete random variables and probability distributions of continuous random variables. Bayesian theory forms the basis of any probabilistic framework. Bayes' rule $P(A|B) = P(B|A)P(A)/P(B)$ is simple but powerful and allows to invert the direction of inference.

1 Bayesian inference introduces the mathematical concepts of Bayesian theoy, such as discrete random variables, conditional, joint and marginal probabilities, Bayes' rule, total probability, partial evidence, expectation values, continuous random variables, ...
→ Videos 1.1, 1.2-1.4, 1.5-1.6, Exercises, Solutions

2 Application: Visual attention + shows a nice application of Bayesian theory to the modeling of visual attention. What is interesting here is that attention is not modeled as a means to manage limited resources, as is often done, but to deal with uncertainty.

Core text and formulas are set in dark red, one can repeat the lecture notes quickly by just reading these; ♦ marks important formulas or items worth remembering and learning for an exam; ◊ marks less important formulas or items that I would usually also present in a lecture; + marks sections that I would usually skip in a lecture.
More teaching material is available at https://www.ini.rub.de/PEOPLE/wiskott/Teaching/Material/.

Contents

1 Bayesian inference

(This section is largely based on (Bishop, 2006; Cowell, 1999; Jensen and Lauritzen, 2001; Jordan and Weiss, 2002).)

1.1 Discrete random variables and basic Bayesian formalism

Random variables (D: Zufallsgrößen, Zufallsvariablen) **shall be indicated by capital letters** $A, B, C, ...$**, the values they assume by lower case letters** $a, b, c, ...$**, and the set of possible values by calligraphic letters** $\mathcal{A}, \mathcal{B}, \mathcal{C}, ...$ **If a variable** A **has assumed a concrete value** a**, it is called *instantiated*** (for convenience, however, I will often use the phrase "A is known"); **if certain values can be excluded** (but it it might still not be clear which value the variable assumes) **it is said to have received *evidence*.** The reduced set of possible values for A after having received evidence α is $\mathcal{A}_\alpha$. For simplicity we will consider discrete random variables, which can only assume a discrete and finite set of values.

A probability distribution (D: Wahrscheinlichkeitsverteilung) **for** a random variable A **is indicated by** $P_A(A)$ **or simply** $P(A)$**. The probability** (D: Wahrscheinlichkeit) **for** A **assuming a particular value then is** $P(A = a)$ **or** $P(a)$**.** If a variable assumes a concrete value, say 5, one can write $P_A(5)$ or $P(A = 5)$ but should avoid writing $P(5)$, since it is not clear which variable and therefore which probability distribution is referred to. **For convenience I will allow swapping the positions of variables, so that** $P(A, B) = P(B, A)$**.** Thus, the variables are identified by names, not by positions. The word 'distribution' is sometimes dropped for brevity, so that one speaks of a probability $P(A)$ even though it is actually a distribution.

There are some basic definitions and rules for probabilities which the reader should be familiar with. **Any probability distribution must be *normalized* to one**

$$\blacklozenge \quad \sum_a P(a) = 1\,. \quad \text{(normalization)} \tag{1.1}$$

This also holds for conditional probabilities, $\sum_a P(a|B) = 1$, but not for marginals of joint probabilities, of course, i.e. in general $\sum_a P(a, B) \neq 1$, see below for the definitions.

The *joint probability distribution* (D: Verbundwahrscheinlichkeitsverteilung) for A and B is $P(A, B)$. **If a joint probability distribution is given, the probability distribution of a single variable can be**

obtained by summing over the other variables, a process called *marginalization* (D: Marginalisieren), resulting in the marginal distribution (D: Randverteilung)

$$\blacklozenge \quad P(A) = \sum_b P(A,b) \quad \text{(marginal distribution)}. \tag{1.2}$$

The *conditional probability distribution* (D: bedingte Wahrscheinlichkeitsverteilung) for A given B and not knowing anything about other variables is $P(A|B)$. **Joint and conditional probabilities are related by**

$$\blacklozenge \quad P(A,B) = P(A|B)P(B) \tag{1.3}$$

$$\blacklozenge \quad \Longleftrightarrow P(A|B) = \frac{P(A,B)}{P(B)}. \tag{1.4}$$

($P(A|B)P(B)$ is often written from left to right, $P(B)P(A|B)$. Somehow, I prefer writing it from right to left, maybe because I am so used to matrix notation in linear algebra. I think it is more elegant.)

If $P(A|B)$ and $P(B)$ are given, **the *total probability*** (D: totale Wahrscheinlichkeit) **of A is**

$$\blacklozenge \quad P(A) = \sum_b P(A|b)P(b)\,. \tag{1.5}$$

Example: Imagine you have a red and a blue die. The red one is a normal die with six faces with the numbers 1 to 6; the blue die has six faces but only the numbers 1 to 3, each twice. In any event, you pick one die at random, the red die twice as often as the blue die, and then you role it. It is easy to see that the probabilities for color and number are overall like in Table 1.1.

$P(C,N)$	1	2	3	4	5	6	any number
red	1/9	1/9	1/9	1/9	1/9	1/9	6/9
blue	1/9	1/9	1/9	0	0	0	3/9
any color	2/9	2/9	2/9	1/9	1/9	1/9	**9/9**

Table 1.1: Example of a red and a blue die. Joint and marginal probabilities. Bold face indicates the normalization condition.

The joint probabilities $P(\text{color}, \text{number})$, or $P(C,N)$ for short, are shown in the center area, e.g. $P(\text{red}, 5) = 1/9$. The marginal probabilities $P(C)$ and $P(N)$ are shown to the right and bottom, respectively, and can be calculated by summing the joint probabilities over the other variable, e.g. $P(\text{red}) = \sum_{N=1}^{6} P(\text{red}, N) = 2/3$ as we would expect since we pick the red die twice as often as the blue one. The sum over all joint probabilities can be calculated in different ways and should always be 1, i.e. $\sum_{C,N} P(C,N) = \sum_C P(C) = \sum_N P(N) = 1$ as one can verify in the table.

To get the conditional probabilities one has to renormalize the joint probabilities according to (1.4) resulting in Table 1.2

The conditional probabilities tell us, for instance, that the probability for rolling 1 is greater if we use the blue die than if we use the red die. Or is we have rolled a 1, it comes equally likely from the red and the blue die, but if we role a 4, it must come from the red die.

Notice that the conditional probabilities summed over the variable conditioned on does not yield 1. □

Bayes' rule (D: Bayes-Formel) **tells us how to invert conditional probabilities,**

$$\lozenge \quad P(A,B) = P(A|B)P(B) \tag{1.6}$$

$$\lozenge \quad = P(B|A)P(A) \tag{1.7}$$

$$\blacklozenge \quad \Longrightarrow \quad P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad \text{(Bayes' rule)}. \tag{1.8}$$

$P(C\|N)$	1	2	3	4	5	6	$\sum_N P(C\|N)$
red	1/2	1/2	1/2	2/2	2/2	2/2	9/2
blue	1/2	1/2	1/2	0	0	0	3/2
any color	**2/2**	**2/2**	**2/2**	**2/2**	**2/2**	**2/2**	
$P(N\|C)$	1	2	3	4	5	6	any number
red	1/6	1/6	1/6	1/6	1/6	1/6	**6/6**
blue	2/6	2/6	2/6	0	0	0	**6/6**
$\sum_C P(N\|C)$	3/6	3/6	3/6	1/6	1/6	1/6	

Table 1.2: Example of a red and a blue die. Conditional probabilities. Bold face indicates normalization conditions.

In this equation $P(B)$ **is called the *a priory* probability, or simple the *prior*,** $P(A|B)$ **is the *likelihood* of** B **for a fixed** A**, and** $P(B|A)$ **is the *a posteriori* probability of** B **given** A**.**

Two variables are *statistically independent* iff (if and only if) $P(A|B) = P(A)$ or, equivalently, $P(B|A) = P(B)$, which also implies that their joint probability distribution factorizes, i.e.

$$\blacklozenge \quad P(A|B) = P(A) \quad \text{(statistical independence)} \tag{1.9}$$

$$\lozenge \overset{(1.8)}{\iff} \frac{P(B|A)P(A)}{P(B)} = P(A) \tag{1.10}$$

$$\blacklozenge \iff P(B|A) = P(B) \quad \text{(statistical independence)} \tag{1.11}$$

$$\lozenge \iff P(B|A)P(A) = P(A)P(B) \tag{1.12}$$

$$\blacklozenge \overset{(1.3)}{\iff} P(A,B) = P(A)P(B) \quad \text{(statistical independence).} \tag{1.13}$$

All the rules above generalize to more random variables, i.e. one could, for instance, write $P(A,B,C,D) = P(A,B|C,D)P(C,D)$ or $P(A,B|C,D) = P(B,C|A,D)P(A,D)/P(C,D)$. **One can also condition everything on another random variable**, e.g. Bayes rule could read $P(B|A,C) = \frac{P(A|B,C)P(B|C)}{P(A|C)}$.

It is important to realize that Bayesian formalism is not about causality but about beliefs what knowing the value of some variables tells us about the unknown variables. Imagine you like to take a long walk (B = 'long walk') on Saturday mornings if and only if you woke up early in the morning (A = 'woke up early'). That is your general habit, but there is still some chance that you take a long walk if you got up late (A = 'woke up late') or that you do not take a walk (B = 'no walk') even though you got up early. Now, your walk causally depends on the time you wake up and not vice versa. Taking a walk does not make you wake up early. However, if you tell your brother at the phone that you have taken a walk in the morning, he can guess that probably you got up early in the morning. Thus, Bayesian inference is about knowing or believing things, not about causality, although causality might play an important role in defining the system and knowing the (conditional) probability distributions in the first place.

Further readings: (Cowell, 1999, secs. 2, 3).

1.2 Partial evidence

So far, the value of a random variable was either known, i.e. the variable was instantiated, or completely unknown. However, it may also happen that one has *partial evidence* (D: Teilevidenz(?)) $\mathcal{E}$ about a variable.

Imagine I role a die and you have to guess the result. Initially, you would assume that the numbers from 1 to 6 all have the probability 1/6. But then I give you partial evidence by telling you that I have gotten an even number. Then of course you would correct your expectation to a probability of 1/3 for 2, 4, and 6 and to zero for the odd numbers. Thus, partial evidence changes probabilities.

It is quite obvious, how joint probabilities change with partial evidence. You simply discard those cases that are not possible anymore and renormalize the probabilities to one. For the example above, one could make the following table, with A indicating the number I might have roled and $\mathcal{E}_A = \{2, 4, 6\}$ indicating the set of values that are still possible, thereby indicating the evidence about A that I have roled an even number.

A	1	2	3	4	5	6	possible values of A
$P(A)$	1/6	1/6	1/6	1/6	1/6	1/6	original probabilities of A
$P(A)\|\mathcal{E}_A$	-	1/6	-	1/6	-	1/6	some entries discarded by evidence $\mathcal{E}_A$
$P(A\|\mathcal{E}_A)$	-	1/3	-	1/3	-	1/3	renormalized probabilites of A given the evidence

I have invented the notation $P(A)|\mathcal{E}_A$ here to indicate the original probabilities masked by the evidence. Formally one could write

$$P(A|\mathcal{E}_A) \quad = \quad \begin{cases} \frac{P(A)}{\sum_{a' \in \mathcal{E}_A} P(a')} & \text{for } A = a \in \mathcal{E}_A \\ 0 & \text{otherwise} \end{cases} , \tag{1.14}$$

or, for short,

$$\lozenge \quad P(A|\mathcal{E}_A) \quad = \quad \frac{P(A)}{\sum_{a' \in \mathcal{E}_A} P(a')} , \tag{1.15}$$

with the understanding that the probability is zero for discarded values $a \notin \mathcal{E}_A$.

The same applies to joint probabilities, e.g.

$$\lozenge \quad P(A, B|\mathcal{E}_{AB}) \quad = \quad \frac{P(A, B)}{\sum_{(a',b') \in \mathcal{E}_{AB}} P(a', b')} , \tag{1.16}$$

where $\mathcal{E}_{AB}$ is a set of duplets (a, b).

To simplify matters a bit, we assume that evidence is independently given for the different variables at hand, so that we can write

$$\blacklozenge \quad P(A, B|\mathcal{E}_A, \mathcal{E}_B) \quad = \quad \frac{P(A, B)}{\sum_{a' \in \mathcal{E}_A, b' \in \mathcal{E}_B} P(a', b')} . \tag{1.17}$$

A marginal probability is then

$$\blacklozenge \quad P(A|\mathcal{E}_A, \mathcal{E}_B) \quad \overset{(1.2)}{=} \quad \sum_{b \in \mathcal{E}_B} P(A, b|\mathcal{E}_A, \mathcal{E}_B) \tag{1.18}$$

$$\overset{(1.17)}{=} \quad \frac{\sum_{b \in \mathcal{E}_B} P(A, b)}{\sum_{a' \in \mathcal{E}_A, b' \in \mathcal{E}_B} P(a', b')} . \tag{1.19}$$

A conditional probability is

$$\blacklozenge \quad P(B|A, \mathcal{E}_A, \mathcal{E}_B) \quad \overset{(1.4)}{=} \quad \frac{P(A, B|\mathcal{E}_A, \mathcal{E}_B)}{P(A|\mathcal{E}_A, \mathcal{E}_B)} \tag{1.20}$$

$$\overset{(1.17,1.19)}{=} \quad \frac{\frac{P(A,B)}{\sum_{a' \in \mathcal{E}_A, b' \in \mathcal{E}_B} P(a',b')}}{\frac{\sum_{b \in \mathcal{E}_B} P(A,b)}{\sum_{a'' \in \mathcal{E}_A, b'' \in \mathcal{E}_B} P(a'',b'')}} \tag{1.21}$$

$$= \quad \frac{P(A, B)}{\sum_{b \in \mathcal{E}_B} P(A, b)} \tag{1.22}$$

1.3 Expectation values

One is often not interested in the probabilities of individual outcomes but in the average outcome. This can be determined by averaging over all outcomes weighted with the respective probabilities, which is **the *expectation value*** (D: Erwartungswert) or *mean (value)* (D: Mittelwert)

$$\blacklozenge \quad E\{A\} := \bar{A} := \langle a \rangle_a := \sum_a a P(a) \,. \tag{1.23}$$

You see, there are several different ways to write the expectation value. I will usually use the angle brackets notation. The average over a random variable, of course, only exists, if it is a numerical value or any other value for which a weighted sum is defined.

Equation (1.23) generalizes to **the expectation value of functions** of random variables,

$$\blacklozenge \quad \langle f(a) \rangle_a := \sum_a f(a) P(a) \,. \tag{1.24}$$

which is particularly useful if one cannot take the average over a directly, because it is a qualitative variable rather than a quantitative variable. The range of $f(a)$ may be real, complex, vectorial, or whatever can be averaged over. (1.24) includes (1.23) as a special case, if the function is the identity.

When taking **the expectation value of a function in two** or more random **variables** one can play around with the order of summation and with conditional probabilities.

$$\lozenge \quad \langle f(a,b) \rangle_{a,b} = \sum_{a,b} f(a,b) P(a,b) \tag{1.25}$$

$$\lozenge \quad = \langle \langle f(a,b) \rangle_{a|b} \rangle_b = \sum_b \left(\sum_a f(a,b) P(a|b) \right) P(b) \tag{1.26}$$

$$= \langle \langle f(a,b) \rangle_{b|a} \rangle_a = \sum_a \left(\sum_b f(a,b) P(b|a) \right) P(a) \,. \tag{1.27}$$

Taking the expectation value is a linear operation, meaning that

$$\lozenge \quad \langle \gamma f(a) \rangle_a = \gamma \langle f(a) \rangle_a \,, \tag{1.28}$$

$$\lozenge \quad \langle f(a) + g(a) \rangle_a = \langle f(a) \rangle_a + \langle g(a) \rangle_a \,, \tag{1.29}$$

for scalars γ, wich follows directly from the linearity of the sum in the definition (1.24) of the expectation value.

Besides the mean $\bar{a} = \langle a \rangle_a$, another particularly important expectation value is the squared difference between a and its mean, which is the variance $\langle (a - \bar{a})^2 \rangle_a$. In this context I often use the bar notation besides the angle brackets notation.

1.4 Continuous random variables

The formalism above can be readily generalized to continuous random variables. Basically, one simply has to replace sums by integrals and think in terms of probability densities instead of probabilities. The notation also changes a bit. One uses a lower case p instead of upper case P for the densities, and the random variable is usually lower case as well, e.g. x instead of A.

The normalization rule (1.1) becomes

$$\lozenge \quad \int_x p(x) \, \mathrm{d}x = 1 \,. \tag{1.30}$$

The marginal distribution (1.2) becomes

$$\lozenge \quad p(x) = \int_y p(x,y)\,\mathrm{d}y\,. \tag{1.31}$$

Bayes rule (1.8) is still the same

$$\lozenge \quad p(y|x) = \frac{p(x|y)p(y)}{p(x)}\,. \tag{1.32}$$

The expectation value (1.24) becomes

$$\lozenge \quad \langle f(x)\rangle_x := \int_x f(x)p(x)\,\mathrm{d}x\,. \tag{1.33}$$

Thinking in terms of probability densities means that one should not ask "What is the probability that x assumes a particular value?" because that probability is always zero (except if the probability density function is a δ-functions). One can only ask "What is the probability that x lies within a certain interval?", which would be

$$\blacklozenge \quad P(x_1 \le x \le x_2) = \int_{x_1}^{x_2} p(x)\,\mathrm{d}x\,. \tag{1.34}$$

Asking for a particular value would mean that $x_1 = x_2$ and the integral would vanish.

1.5 A joint as a product of conditionals

Assume a probability distribution for five random variables is given by $P(A,B,C,D,E)$. Every statistical expression involving these five variables can be inferred from this joint probability distribution by the rules given above. However, computationally it is rather expensive, because if each variable can assume ten values then 100,000 probabilities have to be stored to define $P(A,B,C,D,E)$. Furthermore, computing the marginal distribution $P(A)$, for instance, would require summation over all other variables, requiring 9,999 additions. Also from an analytical point of view is the expression $P(A,B,C,D,E)$ not very helpful, because it does not reflect any of the structure you might know about the variables. **Thus, it would be nice to get a more efficient and informative representation that takes advantage of the structure of the problem at hand.**

In many cases the **structure can be inferred from the causal relationships** between the variables. For instance, if you consider a sequence of events and you know that event A has a direct causal influence on B, B and D together influence E, B and E together influence C, and that these are all causal relationships there are, then you can simply define

$$\blacklozenge \quad \underbrace{P(A,B,C,D,E)}_{100{,}000\,\odot} := \underbrace{\underbrace{P(C|B,E)}_{1{,}000\,\odot}\underbrace{P(E|D,B)}_{1{,}000\,\odot}\underbrace{P(D)}_{10\,\odot}\underbrace{P(B|A)}_{100\,\odot}\underbrace{P(A)}_{10\,\odot}}_{2{,}120\,\odot}\,. \tag{1.35}$$

and go from there.

The figures under the braces indicate the numbers of original probabilities to be stored with $\odot$ indicating the cost of storage. **They show that writing the joint probability distribution as a product of conditional probability distributions is much more efficient in terms of memory consumption.**

But keep in mind that **Bayesian formalism is not about causality**, e.g. you cannot say that $P(A|B) = P(A)$ because B does not have any causal influence on A, because B still tells you something about A.

You must also be careful with feedback loops. If A influences B, B influences C, and C influences A, you might be tempted to define $P(A,B,C) := P(A|C)P(C|B)P(B|A)$. However, that can cause all kinds of

trouble. For instance, since probabilities have to be normalized to one, we have $\sum_a P(a|C) = \sum_c P(c|B) = \sum_b P(b|A) = 1$, but from that does generally not follow $\sum_{a,b,c} P(a,b,c) = 1$ as you would expect. So, **do not use feedback loops in your product-of-conditionals expression.**

Further readings: (Cowell, 1999, sec. 6).

1.6 Marginalization

The next thing we might be interested in are the marginal distributions of subsets of variables. Knowing the marginals, for instance, would allow us to compute conditional probabilities with (1.4), e.g. $P(C|B) = P(B,C)/P(B)$.

Formally, **to compute the marginal of one variable one simply has to sum the joint probability over all other variables.** The computational costs for doing that, however, can vary dramatically depending on the way one does that. Consider the example of the previous section and three different ways of computing the marginal distribution $P(E)$ from $P(A,B,C,D,E)$.

$$\blacklozenge \quad P(E) = \underbrace{\sum_{a,b,c,d} P(a,b,c,d,E)}_{9{,}999\,\oplus} \tag{1.36}$$

$$\lozenge \quad \overset{(1.35)}{=} \underbrace{\sum_{a,b,c,d} P(c|b,E)P(E|d,b)P(d)P(b|a)P(a)}_{40{,}000\,\otimes + 9{,}999\,\oplus} \tag{1.37}$$

$$\blacklozenge \quad = \underbrace{\sum_b \underbrace{\left(\sum_c P(c|b,E)\right)}_{=1} \underbrace{\left(\sum_d P(E|d,b)P(d)\right)}_{10\,\otimes + 9\,\oplus} \underbrace{\left(\sum_a P(b|a)P(a)\right)}_{10\,\otimes + 9\,\oplus}}_{210\,\otimes + 189\,\oplus} \tag{1.38}$$

where $\otimes$ and $\oplus$ indicate the expenses of a multiplication and a summation operation, respectively. We see here that not only the memory consumption but also **the number of operations needed reduces dramatically if we use the product of conditionals and compute the sums efficiently,** i.e. if we apply the sums only to those factors that depend on the variable summed over. Computing the sum efficiently in this way is **referred to as *variable elimination*.** Summing over a in the example above, for instance, eliminates A and leaves the term $P(b) = \sum_a P(b|a)P(a)$, which only depends on b. (Variable elimination is usually explained as rearranging the factors and shifting the sums as much to the right side as possible. I find it more intuitive to say that one rearranges the factors (which was not necessary in the example above) and focuses the summations by (possibly nested) parentheses.)

If one wants to compute the marginal distribution of all different random variables, one should store and reuse some of the intermediate sums for greater efficiency. We see in **the lecture on graphical models** that bookkeeping of this reuse of sums can be conveniently done in a graph structure reflecting the dependencies between the variables.

Further readings: (Jensen and Lauritzen, 2001, sec. 3.1; Jordan and Weiss, 2002, sec. 3.1).

2 Application: Visual attention +

Imagine you are a subject in a psychophysical experiment and have to fixate the center of a blank computer screen. In each trial a little stimulus in the shape of a light Gaussian spot is presented to you with equal probability at a particular location on the left or on the right of your fixation point or,

in half of the cases, there is no stimulus presented at all. **Your task is to respond if a stimulus is being presented. As a hint you get a precue, which tells you with a reliability of 80% whether the stimulus is going to be on the left or on the right, if it comes up at all.** See Figure 2.1 for an illustration of the trial conditions. The task is difficult because the stimulus is barely visible and its perception corrupted by noise, so you have to concentrate and probably you will tend to attend to the cued location.

Figure 2.1: Protocol for the different types of trials in the attention experiment. Figure: (Shimozaki et al., 2003, Fig. 6, URL)[2.1]

Attention in this kind of experiment is usually conceptualized as a way to optimally use limited resources, the idea being that your visual system is not able to dedicate full computational power to both locations simultaneously. So it has to make a decision where to concentrate the computational resources in order to optimize the performance. **Shimozaki et al. (2003) have used Bayesian theory to model visual attention in a rather different way. They** do not assume limited computational resources but **assume that you trust your results differently because of prior information,** the precue in this case.

The experiment sketched above can be formalized as follows. For symmetry reasons we do not distinguish left and right but only cued and uncued location, which spares us one variable. **Let V be the discrete random variable for the visual input** with subscripts c **and** u **indicating the cued and uncued location,** respectively. V can assume the values $v \in \{s, n\}$ where s **indicates 'stimulus present' and** n **indicates 'no stimulus present'.** We can distinguish four cases and assign the following probabilities to them.

$$\lozenge \quad \text{double stimulus:} \quad P(V_c = s, V_u = s) := 0\,, \tag{2.1}$$

$$\lozenge \quad s \text{ at cued location:} \quad P(V_c = s, V_u = n) := 0.4\,, \tag{2.2}$$

$$\lozenge \quad s \text{ at uncued location:} \quad P(V_c = n, V_u = s) := 0.1\,, \tag{2.3}$$

$$\lozenge \quad \text{no stimulus:} \quad P(V_c = n, V_u = n) := 0.5\,. \tag{2.4}$$

The first case does not occur, i.e. has probability zero, but has been added for the sake completeness to make later summations easier.

The perceptual accuracy at the cued and uncued location are assumed to be equal and independent of each other. Thus, there are no perceptual resources that could be distributed more to the cued location. However, **perception is assumed to be corrupted by Gaussian noise.** If A **indicates** the continuous random

variable for **the internal response** or activity to the visual input, we have

$$\blacklozenge \quad p(A=a|V=v) \;:=\; \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(a-\mu_v)^2}{2}\right), \tag{2.5}$$

$$\lozenge \quad \text{with} \quad \mu_n \;:=\; 0 \tag{2.6}$$

$$\lozenge \quad \text{and} \quad \mu_s \;\geq\; 0\,. \tag{2.7}$$

Add a c or u as a subscript if you want to use this for the cued or uncued location, respectively. The variance of the Gaussians is assumed to be one for simplicity. If no stimulus is present, the Gaussian is centered at zero; if a stimulus is present, the Gaussian is centered at $\mu_s \geq 0$ and a is greater on average than if the stimulus is not present. However, if μ_s is small, the distributions overlap so much that the value of A does not reliably indicate whether the stimulus is present or not.

By simple Bayesian calculus we can derive

$$\lozenge \quad p(A_c, A_u|V_c, V_u) \overset{(2.5)}{=} p(A_c|V_c)p(A_u|V_u)\,, \tag{2.8}$$

$$\lozenge \quad P(V_c, V_u|A_c, A_u) \overset{(1.8)}{=} \frac{p(A_c, A_u|V_c, V_u)P(V_c, V_u)}{p(A_c, A_u)}\,, \tag{2.9}$$

$$\lozenge \quad \text{with} \quad p(A_c, A_u) \overset{(1.2,1.3)}{=} \sum_{v_c, v_u \in \{s,n\}} p(A_c, A_u|v_c, v_u)P(v_c, v_u)\,. \tag{2.10}$$

The conditional probability for a stimulus being present or absent given the responses then is

$$\lozenge \quad P(s|A_c, A_u) \;=\; P(V_c = s, V_u = n|A_c, A_u) + P(V_c = n, V_u = s|A_c, A_u)\,, \tag{2.11}$$

$$\lozenge \quad P(n|A_c, A_u) \;=\; P(V_c = n, V_u = n|A_c, A_u)\,, \tag{2.12}$$

respectively, **where s stands for $(V_c = s, V_u = n) \vee (V_c = n, V_u = s)$ and n stands for $(V_c = n, V_u = n)$. Optimal performance is** obviously **achieved if the subject responds with 'stimulus present' when**

$$\blacklozenge \quad P(s|A_c, A_u) \;>\; P(n|A_c, A_u) \tag{2.13}$$

$$\overset{(1.1)}{\Longleftrightarrow} \quad P(s|A_c, A_u) \;>\; 0.5 \tag{2.14}$$

$$\overset{(1.1)}{\Longleftrightarrow} \quad P(n|A_c, A_u) \;\leq\; 0.5\,. \tag{2.15}$$

It should be emphasized that this Bayesian formalism gives the optimal performance under the assumptions given above. **No better performance can be achieved by any other model.** All information that is available at the cued as well as at the uncued location is being used. There is no particular distribution of computational capacity in favor of the cued location.

Now, if R indicates the response of the subject, which like V can assume the values $r \in \{s, n\}$, and if we assume that the subject responds deterministically according to (2.13) we have

$$P(R = s|A_c, A_u) = \Theta[P(s|A_c, A_u) - P(n|A_c, A_u)]\,, \tag{2.16}$$

where $\Theta(\cdot)$ is the Heaviside step function (with $\Theta(x) := 0$ if $x \leq 0$ and $\Theta(x) := 1$ otherwise).

In (Shimozaki et al., 2003) the performance is quantified with the correct hit and false alarm rate, the former being separately computed for valid cues and invalid cues.

$$\text{correct hit rate given valid cue:} \quad P(R = s|V_c = s, V_u = n)\,, \tag{2.17}$$

$$\text{correct hit rate given invalid cue:} \quad P(R = s|V_c = n, V_u = s)\,, \tag{2.18}$$

$$\text{false alarm rate:} \quad P(R = s|V_c = n, V_u = n)\,. \tag{2.19}$$

These expressions are not easily computed analytically, because they involve difficult to solve integrals, namely

$$P(R = s|V_c, V_u) \quad = \quad \int\int P(R = s|a_c, a_u)\, p(a_c, a_u|V_c, V_u)\, \mathrm{d}a_c\, \mathrm{d}a_u\,. \tag{2.20}$$

which can be further resolved with equations (2.16; 2.11; 2.9; 2.2, 2.3, 2.4; 2.8; 2.5), in that order, resulting in a quite complicated integral. **The authors have** therefore **determined the correct hit and false alarm rates numerically** with Monte Carlo methods, i.e. by random sampling. Figure 2.2-left shows these rates as a function of the signal-to-noise ratio (SNR), which is μ_s in this case. **They also performed psychophysical experiments for comparison, see Figure 2.2**-right. For these experiments the SNR has been estimated from the image contrast. **Figure 2.3 illustrates the effect the cueing has on the correct hit rate.**

Figure 2.2: Correct hit rates given a valid cue (pHv) or an invalid cue (pHi) and false alarm rates (pFA) for different signal-to-noise ratios (SNR). Simulation results for the Bayesian model are shown on the left, experimental results for the first of three subject are shown on the right.
Figure: (Shimozaki et al., 2003, Figs. 4, 7, URL)[2.2]

Figure 2.3: Cueing effect defined as the difference between pHv and pHi for three models (left) and three subjects (right). The sum-of-lilelihoods model is the Bayesian model discussed here (bottom, most wiggly line). The authors also considered a more traditional attentional-switching model (top line) and a linear-weighting model (middle line), both of which did not explain the experimental data well.
Figure: (Shimozaki et al., 2003, Figs. 5, 8, URL)[2.3]

Apart from a certain misfit between the scaling of the abscissas, **the fit between the Bayesian model and human performance is rather good.** The scaling of the abscissas might be due to the fact that human sensitivity to the stimuli is worse than the theoretical optimum by about a factor of two, which is reasonable.

So we know now that the Bayesian model performs optimally and that human subjects perform similarly to the Bayesian model except for the scaling of the SNR-axis. The pHv- and pFA-curves in Figure 2.2-left are intuitively clear. But why is the pHi-curve always below the pHv-curve even though all information is used at both locations? And why does the pHi-curve first drop and only then rise to approach one as the SNR gets large?

To get a formal understanding for why the pHi-curve is below the pHv-curve consider Equation (2.13) for a concrete pair of values a_c and a_u.

$$\overset{(2.13)}{0} < P(s|a_c,a_u) - P(n|a_c,a_u) \tag{2.21}$$

$$\overset{(2.11)}{=} P(V_c=s,V_u=n|a_c,a_u) + P(V_c=n,V_u=s|a_c,a_u) - P(V_c=n,V_u=n|a_c,a_u) \tag{2.22}$$

$$\overset{(2.9;2.2,2.3,2.4)}{\Longleftrightarrow} 0 < 0.4\,p(a_c,a_u|V_c=s,V_u=n) + 0.1\,p(a_c,a_u|V_c=n,V_u=s) - 0.5\,p(a_c,a_u|V_c=n,V_u=n) \tag{2.23}$$

$$\overset{(2.8;2.5)}{=} 0.4\exp\left(-\frac{(a_c-\mu_s)^2+a_u^2}{2}\right) + 0.1\exp\left(-\frac{a_c^2+(a_u-\mu_s)^2}{2}\right) - 0.5\exp\left(-\frac{a_c^2+a_u^2}{2}\right). \tag{2.24}$$

If the stimulus is at the cued location, a_c is typically greater than a_u and therefore also $\exp\left(-\frac{(a_c-\mu_s)^2+a_u^2}{2}\right)$ is greater than $\exp\left(-\frac{a_c^2+(a_u-\mu_s)^2}{2}\right)$. If we take the same values a_c and a_u but exchange them, which would be the case if the cued and uncued location swapped roles, then (2.24) becomes smaller, because of the different weighting factors 0.4 and 0.1, and the system is less likely to report 'stimulus present'. **Intuitively one might phrase it as follows: If a_c is relatively large and greater than a_u then this is an indication for the stimulus being present at the cued location and it is consistent with the general expectation that the stimulus appears at the cued location; if a_u is relatively large and greater than a_c then this is an indication for the stimulus being present at the uncued location, but it is in conflict with our general expectation that the stimulus appears at the cued location. Thus, we trust the evidence less and tend to believe that it is just noise taking large values that we see.**

The second question why the pHi-curve first drops is more difficult to answer and I did not come up with a nice intuitive explanation. Suggestions are welcome.

I like this model for two reasons. Firstly, it models attention in a very principled way based on Bayesian theory and the notion of optimality. No *ad-hoc* assumptions are necessary. **Secondly, the model shows that attentional effects can arise without assuming limited resources.** This opens a new way of thinking about attention. I am sure that such a Bayesian approach does not work for all aspects of attention. Limited resources probably play an important role in some attentional tasks. But this model shows that not everything has to be due to limited resources.

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. Neural Networks/Kernel Methods/Graphical Models.

Cowell, R. (1999). Introduction to inference for Bayesian networks. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 9–26. MIT Press.

Jensen, F. and Lauritzen, S. (2001). Probabilistic networks. In *Handbook of Defeasible and Uncertainty Management Systems, Vol. 5: Algorithms for Uncertainty and Defeasible Reasoning*, pages 289–320. Kluwer Academic Publishers.

Jordan, M. I. and Weiss, Y. (2002). Graphical models: Probabilistic inference. In Arbib, M., editor, *Handbook of Neural Networks and Brain Theory.* MIT Press, 2nd edition.

Shimozaki, S. S., Eckstein, M. P., and Abbey, C. K. (2003). Comparison of two weighted integration models for the cueing task: linear and likelihood. *Journal of Vision*, 3(3):209–229.

Notes

[2.1] Shimozaki, Eckstein, Abbey, 2003, Journal of Vision, 3(3):209–229, Fig. 6, http://jov.arvojournals.org/article.aspx?articleid=2192557

[2.2] Shimozaki, Eckstein, Abbey, 2003, Journal of Vision, 3(3):209–229, Figs. 4, 7, http://jov.arvojournals.org/article.aspx?articleid=2192557

[2.3] Shimozaki, Eckstein, Abbey, 2003, Journal of Vision, 3(3):209–229, Figs. 5, 8, http://jov.arvojournals.org/article.aspx?articleid=2192557