Basics of Probability Theory

Contents

Probability as a representation of our state of knowledge

Let's call I all the *information* you have a this given moment. And I am talking about absolutely everything: what your parents taught you, what you learned in school, what you learned in college, what your eyes see right now on some scientific instruments. Now consider a, well-defined, sentence A that says something about the world. For example, A could be "The result of the next coin toss John performs will be heads." Or anything really. We want a technical machinery that can turn all the information I we have into a real number that tells us how plausible it is that A is true. This is what probability theory does. It gives us such a number. Call it $p(A|I)$ and read it as "the probability that A is true given that we know I." So, probability theory is an attempt to represent our state of knowledge about the world.

But what about frequencies?

In introductory courses to probability or statistics, we usually learn that the probability of an event is the frequency with each it occurs in nature. This is absolutely fine if the event is something that indeed occurs repeatedly. However, this intrepretation is quite restrictive. In particular, what can we say about an event that can happen only once? This interpretation forbids the quantification of epistemic uncertainties. We will expand the interpretation of probability. It can be shown, see [Jaynes, 2003] for the proof, that this interpretation is compatible with the frequency interpretation. That is, when events occur repeatedly then the probabilities do become frequencies.

The common sense assumptions that give rise to the basic probability rules.

> Probability theory is nothing but common sense reduced to calculation. Pierre-Simon Laplace, Théorie analytique des probabilités (1814)

Consider the following three ingedients:

- A: a logical sentence
- B: another logical sentence
- I: all the information we know

No other restriction apart that A and B are not contradictions.

We need a bit of notation so that we write less math:

- not $A \equiv \neg A$
- A and $B \equiv A, B \equiv AB$
- A or $B \equiv A + B$

Now, let's try to make a robot that can argue under uncertainty. It should be able to take logical sentences (such as A and B above) and argue about them using all the information it has. What sort of system should govern this robot. The following desiderata, see [Jaynes, 2003], seem reasonable:

- Degrees of plausibility are represented by real numbers.
- The system should have a qualitative correspondence to common sense.
- The system should be consistence in the sense that:
  - If a conclusion can be reached in two ways, each way must lead to the same result.
  - All evidence relevant to a question should be taken into account.
  - Equivalent states of knowledge must be represented by equivalent plausibility assignments.

Cox's theorem shows that:

> The desiderata are enough derive the rules of probability theory.

Talking about probabilities

We read $p(A|BI)$ as:

- the probability of A being true given that we know that B and I are true; or
- the probability of A being true given that we know that B is true; or
- the probability of A given B.

Interpratation of probabilities

The probability $p(A|BI)$ is a number between 0 and 1 quantifying the degree of plausibility that A is true given B and I. Specifically:

- $p(A|B,I) = 1$ when we are certain that A is true if B is true (and I).
- $p(A|B,I) = 0$ when we are certain that A is false if B is true (and I).
- $0 < p(A|B,I) < 1$ when we are uncertain about A if B is true (and I).
- $p(A|B,I) = \frac{1}{2}$ when we are completely ignorant about A if B is true (and I).

The rules of probability

There are two rules of probability from which everything else can be derived. These are direct consequencies of the desiderate and Cox's theorem. They are:

- The **obvious rule**:

$$p(A|I) + p(\neg A|I) = 1.$$

The sum rule is obvious. It states that either A or its negation $\neg A$ must be true. (It is vitally important that you do not try to apply probability in a system that includes contradictions.)

- The **product rule** (or Bayes' rule or Bayes' theorem): More like the definition of conditional probability, and solving for p(AB)

$$p(A,B|I) = p(A|B,I)p(B|I).$$

The product rule is not obvious. Understanding it requires a bit of meditation. It states that the probability of A and B is the probability of A given that B is true times the probability that B is true. Even though the correspondance is not one to one, visualizing events using the Venn diagrams helps in understanding the product rule:

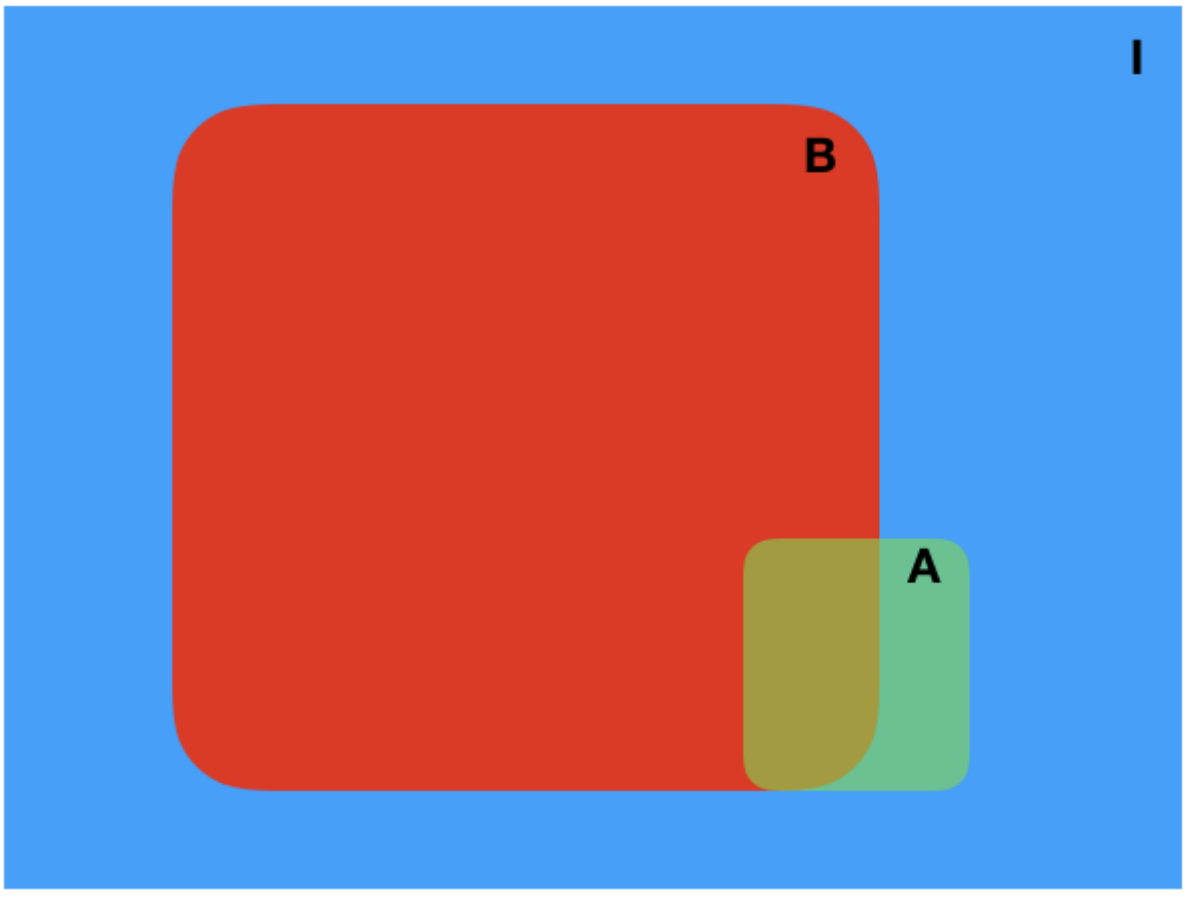


Fig. 2 Venn diagram.

In this diagram:

- $p(A, B|I)$ corresponds to the brown area (normalized by the area of I).
- $p(B|I)$ is the area of B (normalized by the area of I).
- $p(A|BI)$ is the brown area (normalized by the area of B).

I is like the event space

Example: Drawing balls from a box without replacement (1/3)

Consider the following information I:

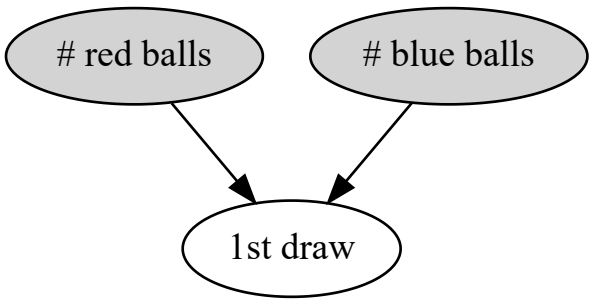
> We are given a box with 10 balls 6 of which are red and 4 of which are blue. The box is sufficiently mixed so that when we get a ball from it, we don't know which one we pick. When we take a ball out of the box, we do not put it back.

Fig. 3 A box with balls.

Now, let's draw the first ball. Here is the graphical causal model up to this point:

```
from graphviz import Digraph
gu1 = Digraph('Urn1')
gu1.node('reds', label='# red balls', style='filled')
gu1.node('blues', label='# blue balls', style='filled')
gu1.node('first', label='1st draw')
gu1.edge('reds', 'first')
gu1.edge('blues', 'first')
gu1.render('urn1_graph', format='png')
```

```
gu1
```



Now, let's say that we draw the first ball. Let B_1 be the sentence:

> The first ball we draw is blue.

What is the probability of B_1? Our intuition tells us to set:

$$p(B_1|I) = \frac{4}{10} = \frac{2}{5}.$$

This is known as the *principle of insufficient reason*. We can now use the **obvious rule** to find the probability of drawing a red ball, i.e., of $\neg B_1$. Of course, $\neg B_1$ is just the sentence:

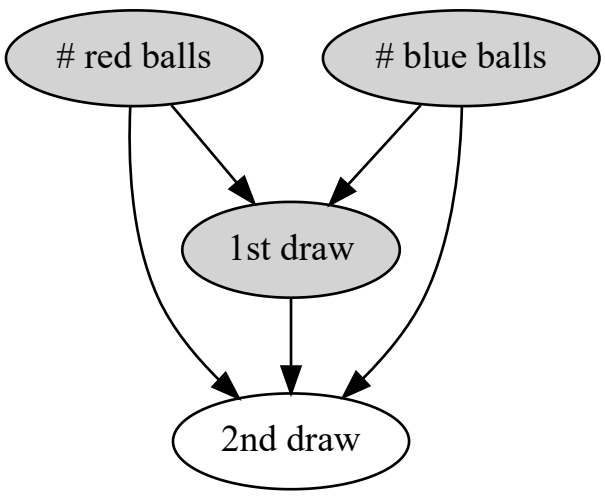
> The first ball we draw is red.

So, let's call it also R_1. It is:

$$p(R_1|I) = p(\neg B_1|I) = 1 - p(B_1|I) = 1 - \frac{2}{5} = \frac{3}{5}.$$

Consider the graphical model representation after we observe the first draw? We need to fill the node corresponding to the first draw with color:

```
gu3 = Digraph('Urn3')
gu3.node('reds', label='# red balls', style='filled')
gu3.node('blues', label='# blue balls', style='filled')
gu3.node('first', label='1st draw', style='filled')
gu3.node('second', label='2nd draw')
gu3.edge('reds', 'first')
gu3.edge('blues', 'first')
gu3.edge('first', 'second')
gu3.edge('reds', 'second')
gu3.edge('blues', 'second')
gu3.render('urn3_graph', format='png')
gu3
```



Consider the sentence R_2:

> The second ball we draw is red.

What is the probability of R_2 given that B_1 is true? We just need to use common sense to find this probability:

- We had 10 balls, 6 red and 4 blue.
- Since B_1 is true (the first ball was blue), we now have 6 red and 3 blue balls.
- Therefore, the probability that we draw a red ball next is:

$$p(R_2|B_1, I) = \frac{6}{9} = \frac{2}{3}.$$

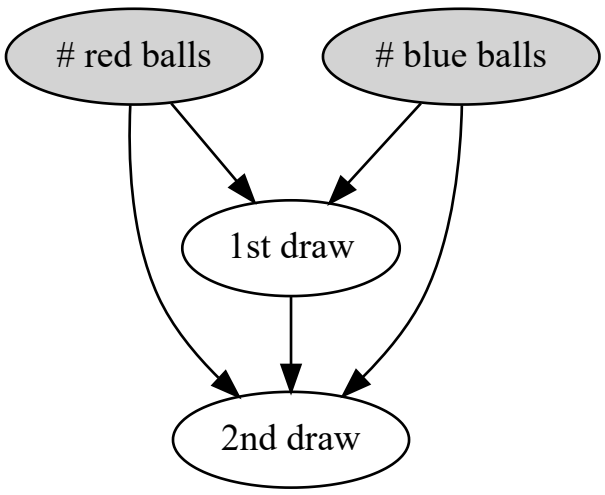
Similarly, we can find the probability that we draw a red ball in the second draw given that we drew a red ball in the first draw:

- We had 10 balls, 6 red and 4 blue.
- Since R_1 is true (the first ball is red), we now have 5 red and 4 blue balls.
- Therefore, the probability that we draw a red ball next is:

$$p(R_2|R_1, I) = \frac{5}{9}.$$

Let's consider a second draw without observing the result of the first draw. What is the graphical causal model now?

```
gu2 = Digraph('Urn2')
gu2.node('reds', label='# red balls', style='filled')
gu2.node('blues', label='# blue balls', style='filled')
gu2.node('first', label='1st draw')
gu2.node('second', label='2nd draw')
gu2.edge('reds', 'first')
gu2.edge('blues', 'first')
gu2.edge('first', 'second')
gu2.edge('reds', 'second')
gu2.edge('blues', 'second')
gu2.render('urn2_graph', format='png')
gu2
```



Recall: Gray means that we observe the value

Let's find the probability that we draw a blue ball in the first draw (A) and a red ball in the second draw (B). We have to use the **product rule**:

$$p(B_1, R_2|I) = p(R_2|B_1, I)p(B_1|I) = \frac{2}{3}\frac{2}{5} = \frac{4}{15}.$$

Other rules of probability theory

All other rules of probability theory can be derived from the two basic rules. Here are some examples.

Extention of the obvious rule

For any two logical sentences A and B we have:

$$p(A + B|I) = p(A|I) + p(B|I) - p(AB|I).$$

In words: the probability of A or B is the probability that A is true plus that probability that B is true minus the probability that both A and B are true. This is very easy to understand intuitively by looking at the Venn diagram.

The probability $p(A + B|I)$ is the area of the uninion of A with B (normalized by I). This area is indeed the area of A (normalized by I) plus the area of B (normalized by I) minus the area of A and B (normalized by I) which was doublecounted.

Let's see a formal proof of this.

$$
\begin{aligned}
p(A+B|I) &= 1 - p(\neg(A+B)|I) \\
&= 1 - p(\neg A, \neg B|I) \text{ (obvious rule)} \\
&= 1 - p(\neg A|\neg B, I)p(\neg B|I) \text{ (product rule)} \\
&= 1 - [1 - p(A|\neg B, I)]p(\neg B|I) \text{ (obvious rule)} \\
&= 1 - p(\neg B|I) + p(A|\neg B, I)p(\neg B|I) \\
&= 1 - p(\neg B|I) + p(A\neg B|I) \text{ (product rule)} \\
&= 1 - p(\neg B|I) + p(\neg B|A, I)p(A|I) \text{ (product rule)} \\
&= 1 - p(\neg B|I) + [1 - p(B|A, I)]p(A|I) \text{ (obvious rule)} \\
&= 1 - p(\neg B|I) + p(A|I) - p(B|A, I)p(A|I) \\
&= 1 - [1 - p(B|I)] + p(A|I) - p(B|A, I)p(A|I) \text{ obvious rule)} \\
&= p(A|I) + p(B|I) - p(B|A, I)p(A|I) \\
&= p(A|I) + p(B|I) - p(AB|I) \text{ (product rule)}.
\end{aligned}
$$

The sum rule

This is the final rule we are going to consider in this lecture. It is one of the most important rules. **You absolutely have to memorize it.** It goes as follows.

Consider the sequence of logical sentences $B_1, \ldots, B_n$ such that:

- One of them is definitely true:

$$p(B_1 + \cdots + B_n|I) = 1.$$

- They are mutually exclusive:

$$p(B_iB_j|I) = \delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}$$

i.e., they fill up the entire probability space and are disjoint

Then, for any logical sentence A we have:

p(A) = sum(p(AB_i))

$$p(A|I) = \sum_{i=1}^{n} p(AB_i|I) = \sum_{i=1}^{n} p(A|B_i, I)p(B_i|I).$$

Again, this requires a bit of meditation. You take any logical sentence A and set of exclusive but exhaustive possibilities $B_1, \ldots, B_n$ and you break down the probability of A in terms of the probabilities of the B_i's. The Venn diagrams helps to understand the situation:

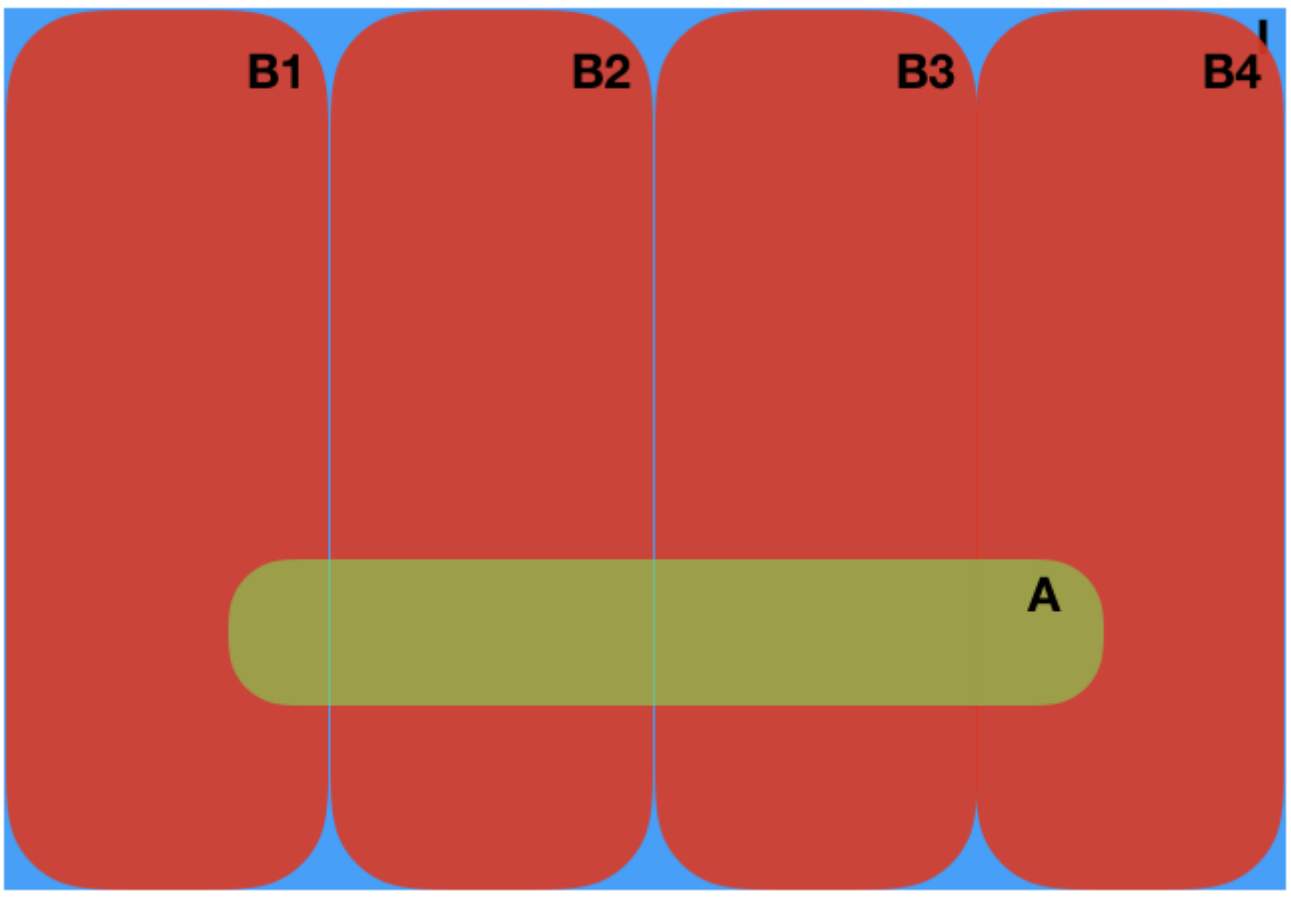


Fig. 4 Venn diagram demonstration of the sum rule.

The sum rule can be trivially proved by induction using only the obvious rule and the product rule. It is instructive to go through the proof. For $n = 2$ we have:

$$\begin{aligned} p(A|I) = & p(A \text{ and } (B_1 \text{ or } B_2)|I) \\ = & p\left((A \text{ and } B_1) \text{ or } (A \text{ and } B_2)|I\right) \\ = & p(A \text{ and } B_1|I) + p(A \text{ and } B_2|I) - p\left((A \text{ and } B_1) \text{ and } (A \text{ and } B_2)|I\right) \\ = & p(AB_1|I) + p(AB_2|I) - p(AB_1B_2|I) \\ = & p(AB_1|I) + p(AB_2|I), \end{aligned}$$

because

$$p(AB_1B_2|I) = p(B_1B_2|I)p(A|I) \leq p(B_1B_2|I) = 0.$$

And then, assume that it holds for n, you can easily show that it also holds for $n+1$ completing the proof.

Example: Drawing balls from a box without replacement (2/3)

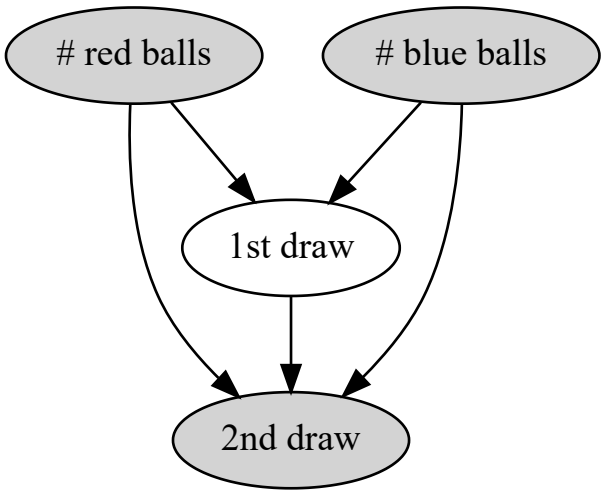
Let us consider the probability of getting a red ball in the second draw without observing in the first draw $p(B_1|I)$. We have two possibilities for the first draw. We either got a blue ball (B_1 is true) or we got a red ball (R_1 is true). In other words B_1 and R_1 cover all possibilities and are mutually exclusive. We can use the sum rule:

$$\begin{aligned} p(R_2|I) = & p(R_2|B_1, I)p(B_1|I) + p(R_2|R_1, I)p(R_1|I) \\ = & \frac{2}{3}\frac{2}{5} + \frac{5}{9}\frac{3}{5} \\ = & 0.6. \end{aligned}$$

Example: Drawing balls from a box without replacement (3/3)

If you paid close attention, in all our examples the conditioning we did followed the causal links. For instance, in the urn example we where writing $p(R_2|B_1, I)$ for the probability of getting a red ball in the second draw after having observed the blue ball in the first draw. This is the uncertainty propagation problem. However, conditioning on stuff **does not have to follow the causal links**. It is completely legitimate to ask what is the probability of a blue ball in the first draw given that you have observed that the result of the second draw is a red ball. The situation is visualized in the following graph:

```
gu4 = Digraph('Urn4')
gu4.node('reds', label='# red balls', style='filled')
gu4.node('blues', label='# blue balls', style='filled')
gu4.node('first', label='1st draw')
gu4.node('second', label='2nd draw', style='filled')
gu4.edge('reds', 'first')
gu4.edge('blues', 'first')
gu4.edge('first', 'second')
gu4.edge('reds', 'second')
gu4.edge('blues', 'second')
gu4.render('urn4_graph', format='png')
gu4
```



That is, you can write down the mathematical expression $p(B_1|R_2, I)$. This does not mean that R_2 is causing B_1. What happens here is that observing R_2 changes your state of knowledge about B_1. This is an example of information flowing in the reverse order of a causal link and a quintessential example of the inverse problem. Let's solve it analytically:

$$\begin{aligned} p(B_1|R_2, I) = & \frac{p(B_1, R_2|I)}{p(R_2|I)} \\ = & \frac{\frac{4}{15}}{0.6} \\ \approx & 0.44. \end{aligned}$$

This is greater than the probability of drawing a blue ball in the first place, $p(B_1|I) = 0.4$. Does this make sense? Yes it does! Here is how you should think:

- You draw a ball without seeing it and you put in a box.
- You draw the second ball and you see that it is a red one.
- This means that this particular red ball was not picked in the first draw.
- So, it is as if in the first draw you had one less red to worry about which increases the probability of a blue.
- So, it is as if you had 5 red balls and 4 blue balls giving you a probability of blue $\frac{4}{9} \approx 0.44$.

This is amazing! It agrees perfectly with the prediction of the product rule. This was one of our desiderata (if you compute something in two different ways you should get the same result). You can rest assured that as soon as you use the product rule and the sum rule and logic, it is impossible to get the wrong answer. That is, if you can actually carry out the computations.

By Ilias Bilionis (ibilion[at]purdue.edu)