Noisy Reasoners: Errors of Judgement in Humans and AIs

Fintan Costello

School of Computer Science and Informatics,
University College Dublin,
Belfield, Dublin 4, Ireland
fintan.costello@ucd.ie

Abstract. This paper examines reasoning under uncertainty in the case where the AI reasoning mechanism is itself subject to random error or noise in its own processes. The main result is a demonstration that systematic, directed biases naturally arise if there is random noise in a reasoning process that follows the normative rules of probability theory. A number of reliable errors in human reasoning under uncertainty can be explained as the consequence of these systematic biases due to noise. Since AI systems are subject to noise, we should expect to see the same biases and errors in AI reasoning systems based on probability theory.

1 Introduction

The ability to reason under uncertainty is fundamental to AI. In this paper I consider this type of reasoning in the case where the AI reasoning mechanism is itself subject to random error or noise in its own processes.

Many AI systems reason using the rules of probability theory, which are normatively correct and provably optimal in at least some situations. It may appear obvious that noise in the workings of a intelligent agent will result in nothing more than random variation around the correct response. This, however, is not the case. There are a number of ways in which random variation can produce systematic biases in reasoning, leading to reliable deviations from the normatively correct responses in particular situations; that is, to reliable errors in reasoning. My main aim in this paper is to present these systematic biases due to random variation.

In addition, I show that a number of reliable errors in human reasoning under uncertainty can be explained as the systematic effects of random variation or noise in a reasoning process that follows the normative rules of probability theory. I argue that, since AI systems (like everything else in the universe) are subject to noise, we should expect to see the same biases and errors in AI reasoning.

The organisation of the paper is as follows. In the first section I describe four well-established and systematic errors in human probabilistic reasoning: conservatism, subadditivity, the conjunction error, and the disjunction error. In the second section I describe how noise can cause systematic biases in a

Fig. 1. Scatterplot showing probability estimates SP(subjective probabilities) versus objective, true probabilities (OP), from [3]. Probability estimates which agree with objective probabilities fall on the 45° line. For low objective probablities estimates fall above that line, while for high objective probabilities estimates fall below that line, demonstrating conservatism.

reasoning process that follows the equations of probability theory, and show how these these systematic biases due to noise produce exactly the patterns of conservatism, subadditivity and the conjunction and disjunction errors seen in humans (as far as I am aware this is the first time a unified account has been given for these four distinct patterns of error). In the third section I present a modified expression for event probability can reduce some of these errors.

2 Biases and errors in human probabilistic reasoning

A very extensive literature exists demonstrating systematic biases and errors that people make in estimating probability. Here I review 4 of these: conservatism, subadditivity, the conjunction error, and the disjunction error. I take $P(A)$ to represent the objective, true probability of some event A, $P_E(A)$ to represent a reasoner's estimate of that probability as influenced by random noise in the reasoning process, and $\overline{P_E}(A)$ to represent the mean or expected value of $P_E(A)$ (the average estimate of the probability of event A).

2.1 Conservatism

Probabilities fall between 0 and 1 by definition. A large body of literature demonstrates that people tend to keep away from these extremes in their probability judgments, and so are 'conservative' in their probability assessments. These results show that the closer $P(A)$ is to 0, the more $\overline{P_E}(A)$ is greater than $P(A)$, while the closer $P(A)$ is to 1, the more $\overline{P_E}(A)$ is less than $P(A)$ [3]. Figure 2 shows this relationship for one set of data.

2.2 Subadditivity

A set of events is mutually exclusive if at most 1 member of that set can occur. A fundamental and obvious requirement of probability theory concerns mutually exclusive events. Let $A_1 \ldots A_n$ be a set of n mutually exclusive events, and let $A = A_1 \vee \ldots \vee A_n$ be the disjunction (the 'or') of those n events, so that A occurs if any of those n events occur. Then probability theory requires that

$$P(A_1) + \ldots + P(A_n) = P(A)$$

More specifically, if $A_1 \ldots A_n$ is a set of n mutually exclusive events that is *complete* - so that exactly 1 of those events is certain to occur - then probability theory requires that

$$P(A_1) + \ldots + P(A_n) = 1$$

Given the obvious nature of these requirements, it is surprising to find that people violate them reliably and systematically. However, experimental studies have shown that people do violate these requirements, and in a characteristic way. Results show that, for mutually exclusive events $A_1 \ldots A_n$

$$\overline{P_E}(A_1) + \ldots + \overline{P_E}(A_n) > \overline{P_E}(A)$$

holds, so that on average the sum of people's estimates for the probability of the constituent events of A is reliably greater than their estimate for the probablity of A) and that the difference

$$\overline{P_E}(A_1) + \ldots + \overline{P_E}(A_n) - \overline{P_E}(A)$$

increases reliably as n increases. Result also show that for mutually exclusive and complete events $A_1 \ldots A_n$

$$\overline{P_E}(A_1) + \ldots + \overline{P_E}(A_n) > 1$$

so that on average the sum of people's estimates for the probability of events $A_1 \ldots A_n$ is reliably greater than 1 with the difference increasing reliably as n increases. There is one reliable exception to this last pattern, which occurs for mutually exclusive and complete events in the specific case where $n = 2$. In this specific case we find

$$\overline{P_E}(A_1) + \overline{P_E}(A_2) = 1$$

holds, so that on average people's estimates for the probability of events A_1 and A_2 will sum to 1 as required by probability theory (see [7] for a review).

3 Conjunction error

The above two biases concern averages of estimated probability values. The next two errors concern differences between people's probability estimates. Let A_1 and A_2 be any two events ordered so that $P(A_1) \leq P(A_2)$. Also let $A_1 \wedge A_2$ represent the conjunction of those two events, so that $A_1 \wedge A_2$ is true only when A_1 and A_2 both occur. Then

$$P(A_1 \wedge A_2) \leq P(A_1)$$

must always hold. This is an obvious and transparent requirement, following from the fact that $A_1 \wedge A_2$ can only occur if A_1 itself occurs. In most cases people follow this requirement when assessing conjunctive probability. People reliably violate this requirement for some events, giving estimates where

$$P_E(A_1 \wedge A_2) > P_E(A_1)$$

This 'conjunction error' does not occur for all or even most conjunctions (people correctly follow the rules of probability theory for most conjunctions). Numerous experimental studies have shown that the occurence of this error depends on the average estimated probability for A_1 and A_2. In particular, the greater the difference between $\overline{P_E}(A_1)$ and $\overline{P_E}(A_2)$, the more frequent the conjunction error is, and the greater the estimated conditional probability $\overline{P_E}(A_1|A_2)$, the more frequent the conjunction error is. The frequency of the error can be high when these two conditions hold (see [1] for a detailed review).

3.1 Disjunction error

Again let A_1 and A_2 be two events ordered by increasing probability, and let $A_1 \vee A_2$ represent the disjunction of those two events (so that $A_1 \vee A_2$ is true if either A_1 or A_2 occurs). Then

$$P(A_1 \vee A_2) \geq P(A_2)$$

must always hold. This follows from the fact that $A_1 \vee A_2$ necessarily occurs if A_2 itself occurs. While in most cases people follow this requirement, they reliably violate this requirement for some events, giving estimates where

$$P_E(A_1 \vee A_2) < P_E(A_2)$$

Just as for the conjunction error, the greater the difference between $\overline{P_E}(A_1)$ and $\overline{P_E}(A_2)$, and the higher the estimated conditional probability $\overline{P_E}(A_1|A_2)$, the higher the rate of occurence of the disjunction error. Studies which examine the rate of both errors show a strong correlation between the frequency of the conjunction error for a given pair of events and the frequency of the disjunction error for that same pair(see [2] for a review).

3.2 The reality of these errors

Given the obvious nature of the requirements violated in conjunction and disjunction errors, it is natural to question the reality of these patterns in people's probabilistic judgment. Researchers have considered this issue carefully, and have have attempted to explain away the conjunction error by arguing that it arises only because participants understand the word 'probability' in a way different from that assumed by experimenters, or by asserting that the conjunction error occurs because participants, correctly following the pragmatics of communication, interpret the single statement A_1 as meaning 'A_1 and not A_2'. Very extensive experimental studies (over 100 published papers) have undermined these objections, and confirmed these errors as a reliable aspect of people's probability judgments [1]. In the next section I show how we can explain these errors as a consequnce of random variation in a reasoner using the equations of probability theory.

4 The Systematic influence of Random Variation

In discussing the influence of random variation on probability estimates I assume a rational reasoner with a long-term episodic memory. I assume a 'perfect' reasoner: if the reasoner were not subject to random variation then each estimate $P_E(A)$ would be equal to $P(A)$. I assume a long-term memory containing n episodes where each episode i contains a flag $f_i(A)$, set to 1 if i contains event A and to 0 otherwise. I assume a minimal form of transient error, in which there is some small probability d that when the state of some flag $f_i(A)$ is read, the value obtained is not the correct value for that flag. I take $C(A)$ to be number of flags that were read as 1 and T_A be the number of flags whose correct value is actually 1.

4.1 Explaining conservatism

Our reasoner can compute $P_E(A)$ by querying episodic memory to find count all episodes containing A and dividing by the total number of episodes, giving

$$P_E(A) = \frac{C(A)}{n}$$

Random variation afffects $P_E(A)$ when it causes some flag $f_i(A)$ be read incorrectly. The expected value of $P_E(A)$ is given

$$\overline{P_E}(A) = \frac{T_A(1-d) + (n-T_A)d}{n}$$

(since on average $1-d$ of the T_A flags whose value is 1 will be read as 1, and d of the $n - T_A$ flags whose value is 0 will be read as 1). Since by definition

$$P(A) = \frac{T_A}{n}$$

we get

$$\overline{P_E}(A) = d + (1-2d)P(A) \tag{1}$$

or equivalently

$$\overline{P_E}(A) = P(A) + d(1-2P(A)) \tag{2}$$

and so the average value of $P_E(A)$ deviates from $P(A)$ in a way that systematically depends on both d and $P(A)$. If $P(A) = 0.5$ this difference will be 0, if $P(A) < 0.5$ then since d cannot be negative we have $\overline{P_E}(A) > P(A)$ with the difference approaching $+d$ as $P(A)$ approaches 0, and if $P(A) > 0.5$ then $\overline{P_E}(A) < P(A)$ with the difference approaching $-d$ as $P(A)$ approaches 1. Thus random error or noise in eposodic memory produces conservatism just as seen in people's probability judgments.

As a sanity check on Equation 2 we can measure the degree of fit between Equation 2 and the data in Figure 2 for a range of values of d. Because we expect the degree of random error in episodic memory to be low but not negligible, we would expect the best fit to occur for a low, but not too low, value of d. Figure 4.1 shows that the best fit occurs for values of d around 0.2, consistent with this expectation.

5 Explaining subadditivity

Recall that subadditivity occurs when, for mutually exclusive events $A_1 \ldots A_n$ with A being the disjunction of all those events, people's probability estimates show the pattern

$$\overline{P_E}(A_1) + \ldots + \overline{P_E}(A_n) > \overline{P_E}(A)$$

with the value of the difference rising as n increases.

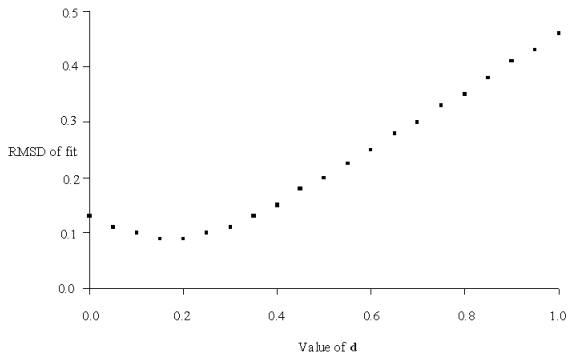

Fig. 2. Scatterplot showing Root Mean Squared Difference (RMSD) between subjective probabilities from Figure 2 and estimates computed by Equation 2 using the corresponding objective probabilities, for a range of values of d.

From Equation 1 we have

$$\overline{P_E}(A_1)+\ldots+\overline{P_E}(A_n) = (P(A_1)+\ldots+P(A_n))+d(n-2\,(P(A_1)+\ldots+P(A_n)))$$

since by assumption

$$(P(A_1)+\ldots+P(A_n)) = P(A)$$

we can rewrite this as

$$\overline{P_E}(A_1)+\ldots+\overline{P_E}(A_n) = P(A)+d(n-2P(A))$$

Also from From Equation 1 we have

$$\overline{P_E}(A) = P(A)+d(1-2P(A))$$

and combining these two expressions we see that Equation 1 implies

$$\overline{P_E}(A_1)+\ldots+\overline{P_E}(A_n) > \overline{P_E}(A)$$

with the value of this expression rising as n increases, just as required.

Recall also that for mutually exclusive and complete events people's probability estimates show the pattern

$$\overline{P_E}(A_1)+\ldots+\overline{P_E}(A_n) > 1$$

except for $n = 2$ when

$$\overline{P_E}(A_1)++\overline{P_E}(A_2) = 1$$

Since for mutually exclusive and complete events we have $P(A) = 1$, from Equation 5 in this situation we get

$$\overline{P_E}(A_1)+\ldots+\overline{P_E}(A_n) = 1+d(n-2)$$

and so $\overline{P_E}(A_1)+\ldots+\overline{P_E}(A_n) > 1$ holds except when $n = 2$ in which case equality holds, just as in people's probaility estimates.

5.1 Explaining conjunction and disjunction errors

The previous two biases concerned the average of people's probability estimates. The conjunction and disjunction errors concern differences between 'samples' from people's probability estimates. Let A_1 and A_2 be any two events ordered by increasing probability so that $P(A_1)$ and $P(A_2)$. For a reasoner following the rules of probability theory we have

$$P_E(A_1 \wedge A_2) = P_E(A_2) \times P_E(A_1|A_2)$$

and so that reasoner's estimate of $P(A_1 \wedge A_2)$ at some time is equal to the product of their estimate for $P(A_2)$ at that time and their estimate for the conditional probability $P(A_1|A_2)$ at that time. Since the reasoner is subject to random variation, these estimates $P_E(A_2)$ and $P_E(A_1|A_2)$ may have some random (positive or negative) difference from the means $\overline{P_E}(A_2)$ and $\overline{P_E}(A_1|A_2)$, and so the equation for conjunction can be rewritten as

$$P_E(A_1 \wedge A_2) = (\overline{P_E}(A_2) + d_{A_2}) \times (\overline{P_E}(A_1|A_2) + d_{A_1|A_2}) \tag{3}$$

where d_{A_2} and $d_{A_1|A_2}$ represent these (positive or negative) deviations from the means. If we assume that A_1 is the less-probable constituent of the conjunction, the conjunction error will occur when

$$\begin{aligned} &\overline{P_E}(A_1) + d_{A_1} < P_E(A_1 \wedge A_2) \\ &\overline{P_E}(A_1) + d_{A_1} < (\overline{P_E}(A_2) + d_{A_2}) \times (\overline{P_E}(A_1|A_2) + d_{A_1|A_2}) \end{aligned} \tag{4}$$

(that is, when the probability of the conjunction from Equation 3 is greater than the probability of its least probable constituent A_1). Equation 4 is most likely to be true when $\overline{P_E}(A_1)$ is low and $\overline{P_E}(A_2)$ and $\overline{P_E}(A_1|A_2)$ are high (because in that situation the left side of Equation 4 is most likely to be low and the right side to be high). We thus expect the conjunction error to be most frequent when $\overline{P_E}(A_1)$ is low and $\overline{P_E}(A_2)$ and $\overline{P_E}(A_1|A_2)$ are both high. This is just the pattern seen when the conjunction error occurs in people's probability estimates.

We can give a similar account of the disjunction error. The probability theory equation for the disjunction $P(A_1 \vee A_2)$ is

$$P(A_1 \vee A_2) = P(A_2) + P(A_1) - P(A_1 \wedge A_2)$$

Just as above this disjunction can be expressed as

$$P_E(A_1 \vee A_2) = (\overline{P_E}(A_2) + d_{A_2}) + (\overline{P_E}(A_1) + d_{A_1}) - P_E(A_1 \wedge A_2)$$

The disjunction error occurs whenever this disjunctive probability $P_E(A_1 \vee A_2)$ is less than its greater constituent probability; that is, whenever

$$\begin{aligned} &P_E(A_1 \vee A_2) < (\overline{P_E}(A_2) + d_{A_2}) \\ &(\overline{P_E}(A_2) + d_{A_2}) + (\overline{P_E}(A_1) + d_{A_1}) - P_E(A_1 \wedge A_2) < (\overline{P_E}(A_2) + d_{A_2}) \end{aligned} \tag{5}$$

is true. Cancelling common terms and rearranging transforms Equation 5 to

$$\overline{P_E}(A_1) + d_{A_1} < P_E(A_1 \wedge A_2) \tag{6}$$

Whenever the inequality in Equation 6 is true, the disjunction error will occur. Equation 6 is identical to Equation 4, which predicts the occurrence of the conjunction error. In other words, Equation 6 predicts that the occurrence of the disjunction error for a given set of items should follow the occurrence of the conjunction error. Again, this is just the pattern seen when the disjunction error occurs in people's probability estimates.

6 Dealing with noise in AI reasoning systems

Many current approaches to reasoning under uncertainty take as their starting point the standard theory of probability; that is, the theory describing the probability of occurrence of repeatable events. These 'Bayesian' approaches to AI apply probability theory in many different areas such as learning, deduction, inference, decision-making, and so on; see Pearl's 1988 book [6], which in some ways founded this line of research (and currently has over 16,000 citations). It is clear from Pearl's work that probability theory provides normatively correct rules which an AI system must use to reason optimally about uncertain events. It is equally clear that AI systems (like all other physical systems) are unavoidably subject to a certain degree of random variation and noise in their internal workings. As we have seen, this random variation does not produce a pattern of reasoning in which probability estimates vary randomly around the correct value; instead,it produces systematic biases that push probability estimates in certain directions and so will produce conservatism, subadditivity, and the conjunction and disjunction errors in AI reasoning.

How can we minimise these biases? We can minimise noise in hardware and software; perhaps more importantly, we can design our AI reasoning systems to take account of internal noise.

6.1 Minimising noise

The previous discussions assumed a single simple form of random variation: an instantaneous random variation which at some particular time, caused some bit in memory to be read incorrectly. In chip design this type 'soft error' can occur due to changes in data being stored in memory or to changes in data being transferred during processing. This type of noise can be produced by cosmic ray impact, by particle decay in the hardware environment, or by random thermodynamic fluctuation. Logic circuits with higher capacitance and logic voltages are less likely to suffer such errors. Unfortunately, such "radiation hardened" designs result in a slower logic gate and a higher power dissipation. Reduction in chip size and voltage, desirable for many reasons, increase the soft error rate. The literature suggests that currently these errors occur at a rate of 1 error per Gbyte per day [5].

As well as using hardened chip design to minimise errors due to noise, designers can make use of error-correcting codes to recover from soft errors. These codes involve adding additional redundant information to data, allowing reconstruction in the event of random error. In general, the reconstructed data is the most likely original data: perfect reconstruction is not guaranteed. Just as with radiation hardened designs, these error correcting codes impose a significant processing cost in terms of time and chip area. Further, these codes cannot eliminate all error: there is an upper bound (the Shannon limit) on the amount of error these codes can remove from data[4].

6.2 Probabilities for noisy reasoners

Designing systems to minimise noise is costly, both in computational time and computational power. A better approach may be to design probabilistic reasoning systems to include an expectation of random error. To do this we can use the equations described previously, but with corrective estimates of the amount of random variation to which the reasoner is suceptible. Suppose the reasoner is suceptible to a known rate of noise d: that is, the reasoner knows that in the long run every X bits read from memory will contain dX bits whose read value is incorrect. For event A define a corrected probability estimate $P_C(A)$ as

$$P_C(A) = \frac{C(A)}{n(1-2d)} - \frac{d}{1-2d} = \frac{P_E(A) - d}{1-2d} \tag{7}$$

On average computed probability estimates $P_E(A)$ will tend to their mean, given by

$$\overline{P_E}(A) = d + (1-2d)P(A)$$

(see Equation 2), and so corrected probability estimates will tend to

$$\overline{P_C(A)} = \frac{\overline{P_E}(A) - d}{1-2d}$$

or substituting

$$\overline{P_C}(A) = \frac{d + (1-2d)P(A) - d}{1-2d} = P(A)$$

and we see that a reasoner that computes its estimate of $P(A)$ as in Equation 7 will in the long run compute estimates that are equal to the true probability of A. Such a reasoner will not suffer from the conservatism and subadditivity biases described earlier. Note, however, that values of $P_C(A)$ will still vary randomly around their mean, and so will still produce conjunction and disjunction errors due to that variation. Discovering ways of eliminating these errors in noisy reasoners is an aim for future work.

References

1. F. Costello. How probability theory explains the conjunction fallacy. *Journal of Behavioral Decision Making*, 22(3):213–234, 2009.

2. Fintan J. Costello. Fallacies in probability judgments for conjunctions and disjunctions of everyday events. *Journal of Behavioral Decision Making*, 22(3):235–251, 2009.
3. Ido Erev, Thomas S. Wallsten, and David V. Budescu. Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, 101(3):519–527, 1994.
4. Cary W. Huffman and Vera Pless. *Fundamentals of Error-Correcting Codes*. Cambridge University Press, 2003.
5. Shubu Mukherjee. *Architecture Design for Soft Errors*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.
6. J Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc. San Francisco, 1988.
7. A. Tversky and D.J. Koehler. Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101(4):547–566, 1994.