If Subsequent Results Are Too Easy to Obtain, the Proof Most Probably Has Errors: Explanation of the Empirical Observation

Olga Kosheleva and Vladik Kreinovich

Abstract Many modern mathematical proofs are very complex, checking them is difficult; as a result, errors sneak into published proofs, even into proofs published in highly reputable journals. Sometimes, the errors are repairable, but sometimes, it turns out that the supposedly proven result is actually wrong. When the error is not noticed for some time, the published result is used to prove many other results – and when the error is eventually found, all these new results are invalidated. This happened several times. Since it is not realistic to more thoroughly check all the proofs, and we want to minimize the risk of errors, it is desirable to come up with some methods to select the most suspicious proofs – so that we can be more attentive when checking those. One such heuristic – developed by mathematicians – is that if subsequent results are too easy to obtain, the proof most probably has errors. This empirical heuristic works in many cases, which leads to a natural question: Why does it work? In this paper, we provide a possible explanation for this heuristic’s success.

1 Formulation of the problem

Errors in mathematical proofs are a serious problem. Modern mathematical proofs are very difficult to check – they deal with complex subjects in which few people are very knowledgeable, and often they combine results from several areas of mathematics, making it even harder to find people who can check them. As a result, sometimes, errors are found in the published proofs; see, e.g., [1]:

Olga Kosheleva
Department of Teacher Education, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: olgak@utep.edu

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: vladik@utep.edu
• mistakes were found in the original proof of the 4-color problem,
• gaps were found in the original proof of the Fermat’s Last Theorem, etc.

Sometimes, the errors are eventually corrected, and a new clean proof appears. However, sometimes, the supposedly proven result turns out to be wrong, and only a restricted version of the original result turns out to be true.

This problem is important because in many cases, once a paper is published (especially if it is published in a serious highly reputable journal), many other papers appear that use this paper’s result – and all these results may collapse if the original result turns out to be mistaken.

So what can we do to better detect errors?

Heuristic ideas can help. As we have mentioned, experience shows that it is not currently possible to more attentively check every part of every proof: people try their best, but errors still sneak through. So, to decrease the probability of an error, a natural idea is to find some heuristic criteria that would help find proofs (or parts of the proof) that are most suspicious and thus, most probable to contain errors. Then, instead of checking all proofs and all parts of proofs, we can spend more time on these suspicious parts – and thus, increase the possibility that we detect the corresponding errors.

Mathematicians have come up with several such heuristic criteria. One of these criteria – explicitly mentioned in [1] – is that if subsequent results are too easy to obtain, then the original proof most probably has errors. This heuristic works: it helped many mathematicians find errors in their proofs.

Natural question. The empirical success of this heuristic gives rise to the following natural question: Why does this heuristic work?

What we do in this paper. In this paper, we provide a possible explanation for the success of this heuristic.

2 Our explanation

Detecting errors in proofs is a particular case of a general problem. In this paper, we are interested in detecting errors in proofs. But let us consider a more general problem: how do we detect any malfunctioning in a system? For example, how do we detect that something is probably wrong with a person’s health?

How is a similar problem solved in medicine? The usual answer to the above medical question is that we measure some characteristics – body temperature, blood pressure, sugar level, etc., and if the values of some of these characteristics is outside the desired interval, this means that something may be wrong. It is not necessarily wrong: when one of us (VK) receives the results of the annual check-up blood test, there are usually several areas marked as red – meaning that the corresponding values are outside the desired intervals – but for most of them, his doctor just tells him to ignore them. These outside-interval values are not necessary indications of some
“error” in a body, but they indicate, to the doctor, what may be suspicious and what needs to be checked more thoroughly.

Where do these “desired intervals” come from? A naive person may think that these intervals are a result of complex research by Top Medical Specialists. Maybe some are, but most of these intervals are obtained in a very simple way:

- we get a sample of values \(x_1, \ldots, x_n\) of this characteristic from the population;
- we find the mean \(m\) and the standard deviation \(\sigma\) of this sample; and
- we select the \(k\)-sigma interval \([m - k \cdot \sigma, m + k \cdot \sigma]\), where usually \(k = 2\) (or \(k = 3\)), as the desired interval.

**This helps.** This simple procedure indeed helps to detect many diseases and abnormalities.

**Where does this procedure come from?** The above procedure comes from statistics. It is based on the fact that the value of each measured characteristic is affected by many different factors. One can read any wikipedia article and see that, e.g., there are many factors affecting blood pressure: unhealthy food, lack of physical activity, stress, etc.

In statistics, there is a well-known Central Limit Theorem (see, e.g., [2]), according to which, under some reasonable conditions, the probability distribution of a quantity that is affected by many approximately-equal-size independent factors – this distribution is close to Gaussian (normal). And for the Gaussian distribution, with high probability, all the values are located within a \(k\)-sigma interval:

- these values are located within a 2-sigma interval \([m - 2 \cdot \sigma, m + 2 \cdot \sigma]\) with probability 95%,
- these values are located within a 3-sigma interval \([m - 3 \cdot \sigma, m + 3 \cdot \sigma]\) with probability 99.9%, etc.

**Comment.** This, by the way, explains why a doctor recommends to ignore some read marks on a blood test. Usually, 2-sigma intervals are used in medicine, since in medicine, it is much more critical not to miss a dangerous disease than to undertake one more unnecessary test. In this case, there is a 100% – 95% = 5% probability that an absolutely healthy value will lead to a red mark. So, for a test with 20 or more numerical values, it is practically certain that one of the values will be outside the healthy interval. A red mark is not necessarily a sign of a disease, it is an indication what to pay attention to – and if other symptoms of the corresponding disease are missing, the doctor naturally recommends to ignore the red mark.

**Is this the best way?** Many medical professionals argue that this is not the best way to describe a healthy desired intervals. For example, since a significant proportion of the US population is overweight, the mean \(m\) – and thus, the resulting interval

\[
[m - k \cdot \sigma, m + k \cdot \sigma]
\]

– is skewed in the unhealthy direction (as can be easily seen if we compare current intervals with similar intervals that the same procedure produced several decades ago).
They also argue that while these statistics-based interval seem to indicate that it is normal for many characteristics like muscle strength to drastically decrease with age, the observed decrease – based on which such recommendations are made – is largely caused by the lack of physical activity in many folks.

These professionals are right: it is better to have more medically justified intervals. However, for many characteristics, we do not have such justified intervals, so the above statistical approach is the best we can do.

**Let us go back to proofs, this way, we will get our explanation.** In contrast to some medical cases, for proofs, we do not have any justified ideas. Thus, we have to use the statistical heuristics. To apply these heuristics, we need to select an appropriate numerical characteristics of the proof – something similar to blood pressure and blood sugar level in medicine. Which characteristics should we select?

To decide on this, let us recall why the problem of finding errors in proofs is so important. If an erroneous theorem is published and never used for a long time, this is not good but not that catastrophic. The big problem is when a supposed-to-be-theorem is used to proof many other results. Usually, this starts with the author(s) of the proof: they usually find some consequences, and then other mathematicians catch up. For such an actively-used theorem, we get quite a few new results based on the original theorem. So, when designing heuristics for error detection, it makes sense to focus of theorems that have many subsequent results.

What numerical characteristics can we produce based on these results? We can gauge the length of the proof of each new result, we can gauge the time that it took us to prove these results. If there are many such subsequent results, their lengths and times probably deviate a lot, so to decrease the effect of this deviation, it makes sense to take the average – just like when you have several measurements with the same Gaussian measurement error, the best way to estimate the actual value is to take the arithmetic average (see, e.g., [2]).

So, the resulting natural criterion is to make sure that the average length and average time-to-prove for these subsequent results lie within the desired intervals. If one of these averages is much smaller or much larger than usual, this is a sign that something may be wrong. So, from this viewpoint, we seem to have two cases that indicate a possible error:

- when subsequent results have unusually short proofs or at least proofs that require unusually small time to prove – i.e., when subsequent results are too easy to obtain, or
- when subsequent results have unusually long proofs and/or proofs that require unusually large time to prove – i.e., when subsequent results are too difficult to obtain.

However, in the second case, when subsequent results are too difficult to prove, most probably we will not get too many subsequent results – maybe not any, so this is exactly the case when we should not worry that much about the errors.

So, our conclusion is that we should focus special attention on proofs for which subsequent results are too easy to obtain – which is exactly the empirical heuristic that we are trying to justify.
Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), HRD-1834620 and HRD-2034030 (CAHSI Includes), EAR-2225395 (Center for Collective Impact in Earthquake Science C-CIES), and by the AT&T Fellowship in Information Technology.

It was also supported by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

References