# Extension of the Number Game

**Ridhika Agrawal (rra10001@nyu.edu)**

**Hoa Duong (hhd2020@nyu.edu)**

**Allen Huang (sh7008@nyu.edu)**

**Kathleena Inchoco (ki2130@nyu.edu)**

## Abstract

In this paper, we extend the number game experiment from Tenenbaum's "A Bayesian Framework for Concept Learning" in order to model human priors from empirical data. We construct a survey to collect empirical data and develop our analysis using the Bayesian framework. We build out the prior using a variety of methods that include different interpretations of the hypothesis space and explore alternative computations of lambda. Our results show that the perceived complexity of the hypothesis space seems to influence human decision-making.

## Introduction

Human concept learning aims to model how humans understand concepts when given a few positive examples. Tenenbaum's thesis explores a computational framework to deconstruct concept learning using a Bayesian inference model, which he understands through the number game. In class, we replicated the number game along with several simplifying assumptions. In this paper, we hope to provide a more nuanced, empirical alternative to these simple assumptions. To this end, we designed a survey to understand how humans think of the possibilities of different concepts, i.e."priors".

## Survey Design

In this project, we hypothesized that priors are not uniformly distributed. We acknowledge that it is difficult to get people to think about priors without showing any data. In fact, asking participants to describe a probability distribution of priors without providing data is generally not tractable. In order to get around this problem, we reasoned that there are other ways we can approximate priors well after showing data, especially by being careful to ask questions that make people introspect on their reasoning.

We designed the survey to reflect the actual test cases we used in the Bayes homework and added follow-up questions to gain insight into people's approaches when solving the problems presented. The three cases we consider are Example Set 1: [16], Example Set 2: [16, 8, 2, 64], and Example Set 3: [16, 23, 19, 20].

Our survey incorporates some important differences from Tenenbaum's experiment. First, we allow people to choose from all numbers from 1 to 100, instead of only 30 probing numbers. We do this to allow people more freedom and flexibility. We only include 11 math hypotheses in the survey for the multiple choice section so that the survey would not be too confusing but still allow the scope for the participants to choose between relatively "simple" hypotheses and "complex" hypotheses. We intentionally ask open-ended questions, where participants type in their answers (for ex."What number sets did you consider when making your selection for [example]?"), before multiple-choice questions, where participants simply select from the preset answers (for ex."Of the following sets, which number sets did you consider when making your selection for [example]?"). The order of the questions avoids bias in ensuring that participants would not be led to certain conclusions. We also kept the survey as open-ended as possible to capture the true thinking of participants without having the multiple-choice questions influence their reasoning. We also asked participants to reveal whether they thought of mathematical rules or interval hypotheses first and to rank the hypotheses they considered when making their decisions.

Additionally, Tenenbaum asked questions about "probability-of-acceptance interpretations": 1, less than 5%"; 2, 10%"; 3, 30%"; 4, 50%"; 5, 70%"; 6, 90%; 7, greater than 95%" (Tenenbaum, 1999, p. 200). After that, the data were scaled linearly from the 1-7 rating scale to the interval [0:1] for comparison with the predicted probabilities (Tenenbaum, 1999, p. 221). However, we chose to ask binary questions, so as not to scale, since scales can be easily misinterpreted. Furthermore, we wanted our survey to be as accessible as possible without making it overly complicated.

## Survey Results

### Count of Membership

Results from 38 participants were aggregated into Figure 1. The height of each bar represents the aggregated number of participants who decided that the corresponding number would be accepted by the computer program, given the examples listed at the top of each plot. Unlike Tenembaum's paper, where only 30 probe numbers were tested, we tested the membership of 100 numbers from 1 to 100. Thus, a missing bar means no participants considered that the objective number would be accepted by the computer program.

We see that when presented with only 16 in the example set, most of the participants agreed that numbers such as 2, 4, 8, 16, 32, and 64 would be accepted by the computer program. This bias towards powers of two is not shown in Tenembaum's paper, shown in Figure 2, where powers of two do not have much higher average predicted probabilities of being accepted by the computer program than other numbers.

Some participants agree that even numbers such as 6, 10, 12, 14, etc. would also be accepted by the computer program. Interestingly, every number from 1-100 was chosen by at least one participant to be accepted by the computer program.

When presented with the example of [16, 8, 2, 64], the majority of the participants in our survey picked numbers such as 2, 4, 8, 16, 32, and 64 to be accepted by the computer program. Some also considered number 1. However, not many participants considered other numbers, and some numbers were not chosen at all. This is slightly different from the results of Tenenbaum's paper. In Tenenbaum's paper, even though powers of two also dominate in terms of predicted membership probability, even numbers and some intervals containing 16 were also considered acceptable by the computer program. This once again shows a bias towards the power of two in our participants' behaviors. One possible explanation for this was that humans' judgments tend to gravitate towards slightly more complex and detailed hypotheses, such as powers of two, over more simple and broad hypotheses, such as even numbers, even when both hypotheses would produce the example of [16, 8, 2, 64]. According to feedback from some individuals in our study, if the computer program were to accept even numbers, even numbers that are not powers of two such as 6 or 12 would have appeared in the example.

Lastly, when presented with the example of [16, 23, 19, 20], most participants in our survey chose numbers in intervals of various sizes surrounding 16 to 23 as acceptable by the computer program, similar to the findings of Tenenbaum's paper. However, about a quarter of participants also considered that other numbers would be accepted by the computer program, unlike in Tenenbaum's paper, where the average predicted membership probability for numbers smaller than 12 or greater than 28 is rather small, much smaller than 25%. As explained in more detail in the next section, this is because participants in our survey have a strong bias towards mathematical rules and away from interval rules. Thus, many participants came up with very complicated mathematical rules to explain the numbers 16, 23, 19, and 20, such as "multiples of 4 and multiples of 4 minus 1", or "pairs of multiple of 4 and prime numbers", rather than resorting to intervals. In fact, most participants in our survey consistently considered mathematical rules before considering interval rules, as shown in Figure 3, and some participants do not even consider interval rules at all. This could be due to the fact that most of our participants are currently attending New York University, or other institutions, and have had or are pursuing education in a STEM-related field. Thus, participants in this study have a natural habit of using mathematical rules to explain various concepts. In fact, this aligns with an expectation from Tenenbaum's paper that as children, one tends to focus on interval rules exclusively, but as one ages, one tends to place more weight on the mathematical rules than on the interval rules, to the point where one favors mathematical rules over interval rules (Tenenbaum, 1999, p. 209).

### Count of Hypotheses

In our experiment, to avoid bias, we asked participants to list their hypotheses for each example set before presenting our hypotheses set. As such, there was no limit on what hypotheses could be considered when selecting the membership of each number for each example set.

Figure 4 shows the number of participants who considered the corresponding hypothesis when presented with the example [16]. It is worth noting that to avoid double-counting, if a participant put the same answer for both the open-ended question and the multiple-choice question, the answer will only be counted once. We see a trend where most participants considered "powers of 2" and "even numbers", with more people considering "power of 2" than "even numbers", despite the latter being more inclusive.

Figure 5 shows the number of participants who considered the corresponding hypothesis when presented with the example [16, 8, 2, 64]. Once again, "powers of 2" dominates all other hypotheses. Notably, 28 out of 38 participants typed in some form of "powers of 2", such as "$2^n$", in the open-ended question, before being exposed to our preset list of hypotheses, suggesting that participants in our survey have a tendency to favor this specific hypothesis.

Lastly, figure 6 shows the number of participants who considered the corresponding hypothesis when presented with the example [16, 23, 19, 20]. As expected, many participants considered "Consecutive numbers in an interval" as a possible number set generated by the computer program. However, it is worth noting that only 19 out of 38 participants actually typed in some form of "intervals" in the open-ended question, suggesting that half of our participants did not think about interval rules before being presented with one. Additionally, more than one participant actually invented more complicated mathematical rules to explain the example set, such as "4n and 4n-1" or "n +/- 3", rather than thinking about interval rules. Many participants also tried to use two rules to explain this example set, using one rule to explain [16, 20], and another rule to explain [19, 23], such as "multiples of four and prime numbers". This explains why, unlike Tenenbaum's paper, in our survey, many participants considered numbers that were outside of the range of 12 to 28 to be acceptable by the computer program. Once again, this highlights the bias toward mathematical rules in our survey participants.

## Empirically Modeling Bayesian Framework

Tenenbaum's simplistic Bayesian frameworks had the following assumptions (Tenenbaum, 1999): Lambda is a free parameter, with $\frac{2}{3}$ being a good estimate since humans tend to prefer mathematical hypotheses over interval hypotheses. All hypotheses in the mathematical hypothesis space are equally likely i.e. $P(h)$ follows a uniform distribution over all the hypotheses. The space itself consists of 30 different hypotheses. For interval hypothesis space, intervals of intermediate size are favored (rather than very small or large hypotheses) by reweighting according to an Erlang distribution,

$P(h) \propto \frac{|h|}{\sigma^2} \times \exp(\frac{|h|}{\sigma})$ where $\sigma = 10$. The space itself consists of a rolling window from 1-100. In order to model human concept learning more effectively, we made some alterations to Tenenbaum's simplistic modeling assumptions.

## Lambda

We decided to use an empirical lambda as a better estimate of the degree to which humans prefer mathematical hypotheses over interval hypotheses. To do this, we employ two methods:

- Pure Lambda: We counted the number of times "Consecutive numbers in an interval" was selected and divided by the number of math and interval hypotheses in the first example, [16]. When counting for the denominator, we treated it as a binary value for the math hypothesis. This means that if a participant selected "powers of 2" and "multiples of 3", the denominator would just increase by 1. We did this to avoid over-inflating the lambda in favor of math hypotheses. Finally, this pure lambda is unaffected by seeing a previous example, which might bias participants' priors. We get a value of 0.8.

- General Lambda: We counted the number of times "Consecutive numbers in an interval" was selected and divided by the number of math and interval hypotheses selected across the three examples. The denominator was calculated in the same way as it was for pure lambda. The argument in favor of the general lambda lies in the fact that all three examples were of different natures and thus the aggregate lambda captures the true, underlying prior beliefs of the participants. This way we get rid of any example-specific bias, as is the case with pure lambda. We get a general lambda value of 0.6.

## Mathematical Hypotheses

Tenenbaum initially assumed that humans believe that all hypotheses in the math hypothesis space are equally likely to occur. However, this is not true from our understanding of the world. If we ask you to think of sets of number patterns, you are very likely to reply with "even numbers" and "odd numbers" as opposed to something like "multiples of 7". This is because as humans, when we think of math hypotheses, we already have unequal priors even before we see a number. Take, for instance, a mathematician who regularly deals with complex hypotheses spaces compared to a child who is in middle school. A mathematician might have a higher prior probability associated with math hypotheses such as "squares" and "cubes" as compared to a middle-schooler who does not even know of the existence of such a pattern. Thus, there could be many distributions of priors in a mathematical hypothesis space. To that end, in our paper, we use empirical priors by counting the number of times a participant selected a particular hypothesis across all three studies. We then reweight the counts to probabilities such that they sum to 1. The reason we can interpret counts from our survey to mean priors is discussed in Section: Case for Empirical Priors.

We modeled the Bayesian framework over two hypothesis spaces:

- Limited Hypothesis Space: We included hypotheses just from the provided list of sets in the question "Of the following sets, which number sets did you consider when making your selection for [example]?". The reason to only include a limited set is because participants are unequally motivated to write their own response, and most would probably consider the provided list as a complete hypothesis space. This means that including the written hypothesis is not a completely accurate measure of the prior beliefs about that hypothesis; if we had provided the written hypothesis in the list of hypotheses, we postulate that the prior might be higher for that hypothesis.

- Extended Hypothesis Space: We included hypotheses from the provided list of sets in the question "Of the following sets, which number sets did you consider when making your selection for [example]?" as well as any participant-written hypothesis. In the number game, a hypothesis space inherently means that those hypotheses excluded from the space have a prior probability of zero. Thus, including the written hypothesis avoids this incorrect implicit assumption. In order to avoid double counting the written-in hypothesis space, we scanned participants' answers across all questions where they could have typed in a response. This means that if a participant wrote down the same hypothesis while answering "Of the following sets, which number sets did you consider when making your selection for [example]?" and "What number sets did you consider when making your selection for [example]?", we just counted this as one.

For simplicity, we only include the extended hypothesis space when reporting the models. Next, we used two methods to calculate the math priors:

- Empirical Priors: We replaced the initially uniform prior distribution with empirical probability. We counted how many times a math hypothesis was selected, and then normalized it such that the numbers sum to 1 and it can be a valid probability measure.

- Weighted Empirical Priors: We postulate that there is a tendency for humans to prefer more complex hypotheses. This pull is supported by theory: Tenenbaum (1999) states that learners have a generative model of observations in their mind "Rather, a preference for powers of two that emerges only after several examples have been observed must be intrinsically statistical, something like a drive to detect and avoid unexplained coincidences in the relation between concepts and their examples". To account for complexity, we make a simplifying assumption that more complex hypotheses would typically have lower cardinality. Thus, we aimed to reweight the empirical probabilities such that they prefer smaller hypotheses. To that end, our approach to characterizing model complexity was through the use of the Erlang distribution, which provided a measure of the variability in the number of independent events required for a certain level of probability. We computed the Erlang distribution score for each hypothesis space size in order to compare their model complexity, with larger scores indicating greater complexity. By utilizing this proxy measure of complexity, we could gain insights into the implicit preference of more complex models, and ultimately identify the sigma that strikes the optimal approximation to human decision. The prior probability is determined as follows:

$$WeightedP(h) \propto \frac{|h|}{\sigma^2} \times \exp(\frac{|h|}{\sigma}) \times P(h)$$

**Case for Empirical Priors** We want to calculate the prior, $P(h)$, but it is usually easier to observe $P(h|x)$, where $x$ are some data, since we cannot really understand how humans think about $h$, without giving some data. We posit that we can approximate $P(h)$ by taking the probabilities across all types of data in our universe. Defining the set of questions in our survey as the scope of our problem, we reason that finding $P(h)$ given no data is best approximated by finding $P(h)$ over all possibilities. We believe that combining information from the three examples generalizes enough to give us a prior which is free of example-specific biases. This is because the examples differ in terms of cardinality - the first example only has one number, while the other two examples have four numbers each. Next, the second and third examples have the same cardinality but differ in type - the second example seems to be drawn from a mathematical hypothesis, while the third example seems to be drawn from an interval hypothesis. This universe includes all of the possible priors we expect a participant to consider (mathematical and interval priors). Thus, we can interpret this as $P(h)$.

Another way we approximate $P(h)$ is by taking the second-best case. Ideally, what we would like is to have people reveal their priors without any given data. However, we reasoned that if that is not feasible then the second-best case should be a good approximation. Consider the first case where we only give one data point: [16]. When we ask participants to give their hypotheses here, this is the closest case we have to model a true prior without giving any data. When the participants only have one number to reason with, they still need to make up for the sparse context by having to think of hypotheses that they believe have the greatest probability of being true in this situation.

### Interval Hypotheses

We largely agreed with Tenenbaum's assumptions when calculating priors for interval hypotheses. Even in the simple case, we think it is a pretty good model of human concept learning. We thus decided to calculate priors as described in Tenenbaum's approach, which consist of using a rolling window from 1-100, and using Erlang distribution to reweight such that intervals of intermediate size are favored as opposed to very small or large hypotheses.

### Model Evaluation and Comparison

We have four models. The average predicted probabilities using each model are shown in Figure 7 to Figure 10.

- Model I. General Lambda, Empirical Prior for Math Hypothesis
- Model II. General Lambda, Erlang Weighted Empirical Prior Probability for Math Hypothesis
- Model III. Pure Lambda, Empirical Prior for Math Hypothesis
- Model IV. Pure Lambda, Erlang Weighted Empirical Prior Probability for Math Hypothesis

To evaluate the performance of the four models, we implement the KL divergence score with respect to human data distribution. $P(x)$ is the distribution of the proposed model, $Q(x)$ is the human data distribution, and X number set from 1 to 100.

$$KL(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$$

We compute the uniform distributed prior which was applied in the homework as the benchmark with a score of 8.5643. The results of the four proposed models are as follows:

Table 1: Model Results

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| KL divergence | 8.6480 | 8.0953 | 8.9908 | 8.2577 |

The results indicate that the general lambda outperforms the pure lambda in the given context, while the weighted empirical priors demonstrate superior performance compared to the empirical priors. However, it is crucial to note that among the various models analyzed, only the weighted empirical priors models were able to surpass the benchmark value established by the uniformly distributed hypotheses model.

Furthermore, the investigation revealed that the optimal sigma value for the Erlang distributions is 14. It is important to acknowledge that relatively complex mathematical hypotheses, such as powers of 2, possess a size of 6, while more straightforward hypotheses, like even numbers, exhibit a size of 50.

## Conclusion

In this study we evaluated the performance of four postulated prior probability models to approximate human decision. While the empirical priors did not surpass the benchmark, the incorporation of hypothesis complexity yielded promising results, suggesting that the complexity of hypotheses indeed influences participants' selection of number sets.

Furthermore, our hypothesis that participants tend to favor more complex models was corroborated by the sigma value in the Erlang distribution. Although the maximum Erlang probability was not generated by the smallest hypothesis space, the observed right-skewed distribution indicates a preference for smaller space sizes. This finding emphasizes the importance of considering hypothesis complexity in modeling human decision-making.

## Limitations and Future Work

One of the limitations we faced was not having enough responses from the survey. We had 38 people respond to our survey but it would have been better if we could have gathered more responses if we had more time. The limited data reduces the power of the results we have. Further, the survey was very long because we asked all participants to choose out of all 100 numbers for each example set. On the other hand, Tenenbaum only asked people to select numbers out of a smaller subset. Even though we gained additional insights from having a larger list of questions, this meant we had to sacrifice overall accuracy as participants might have gotten tired. We also simplified our experiment and asked people to choose out of 11 mathematical hypotheses rather than out of 30 mathematical hypotheses. We did this to balance out the length of the survey, however, doing so means we lost information about priors. Next, we included 18 questions that involved open-ended short answers in addition to multiple-choice questions. This added a lot of instability to our results, not only because it was a free-form response, but because people often chose not to answer since they had to write it in themselves.

Tenenbaum also randomized the example sets in his experiment whereas the order in which we ask the questions in our survey may have implicated some bias in our survey (Tenenbaum, 1999). We expect participants to cite mathematical rules in the second case before the last sequence where we expect participants to cite interval hypotheses. The word choice in our questions may have prompted participants to consider mathematical rules before interval hypotheses. One example is that because people were asked to think of "math" when given the prompt this implied a formalism which leads people to think of strict rules.

Overall, while there are some pitfalls in our experiment design, we believe that using weighted empirical priors is a better, more nuanced model of human concept learning. The number game thus successfully provides us relevant insights, and is a good generalization of the way humans learn categories.

## References

Tenenbaum, J. B. (1999). *A bayesian framework for concept learning* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

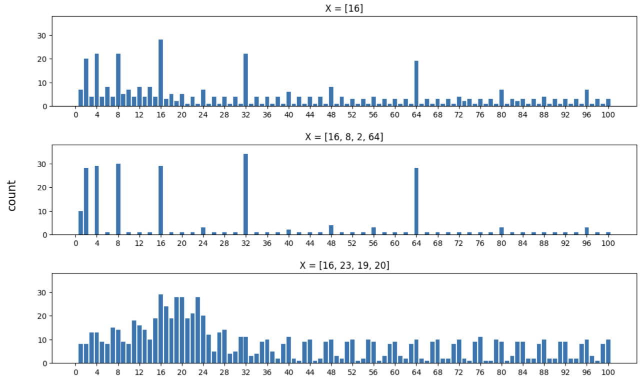

Figure 1: Membership Count

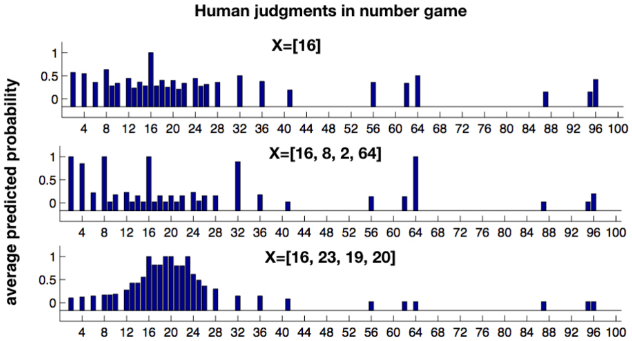

Figure 2: Tenenbaum's Membership Probability

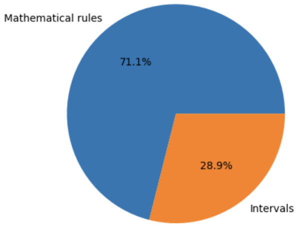

Figure 3: Hypotheses Participants First Considered

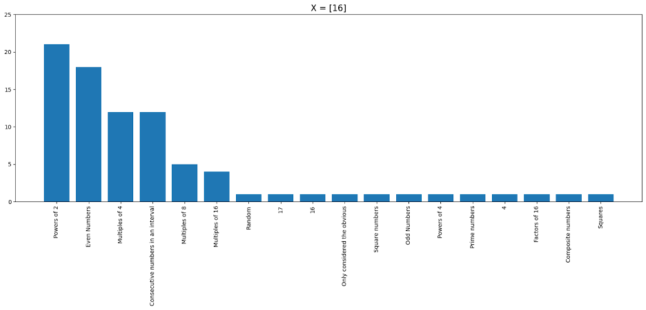

Figure 4: Count of Hypotheses for X = [16]

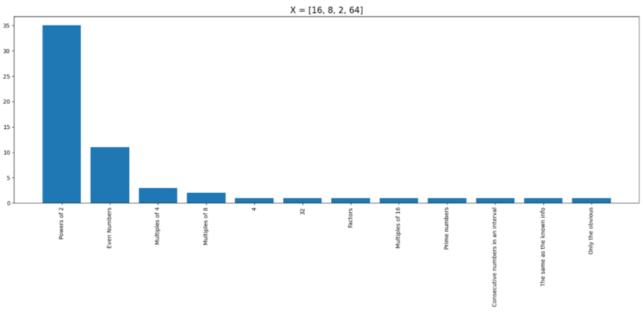

Figure 5: Count of Hypotheses for X = [16, 8, 2, 64]

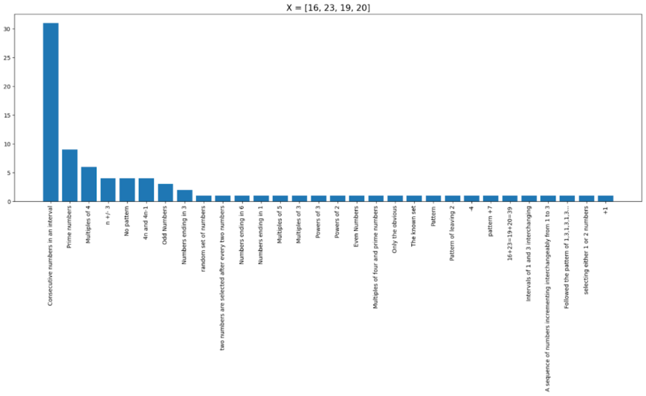

Figure 6: Count of Hypotheses for X = [16, 23, 19, 20]

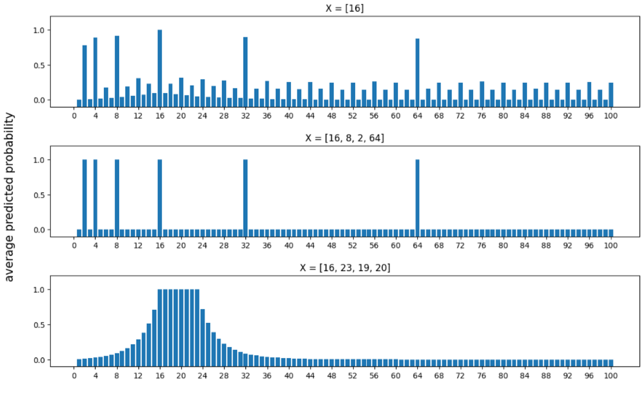

Figure 7: General Lambda, Empirical Prior for Math Hypothesis

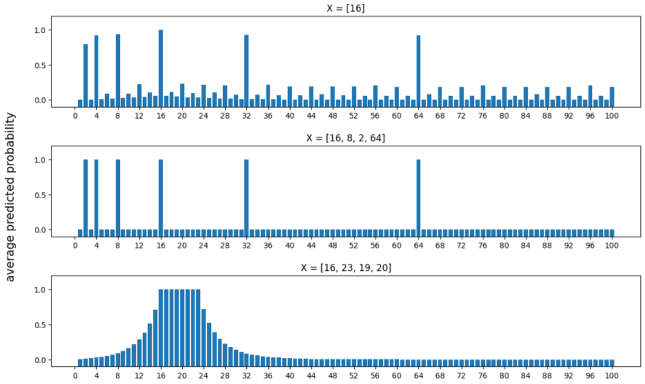

Figure 8: General Lambda, Erlang Weighted Empirical Prior Probability for Math Hypothesis

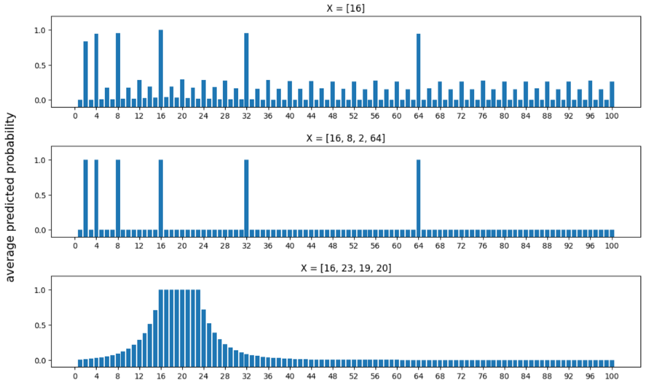

Figure 9: Pure Lambda, Empirical Prior for Math Hypothesis

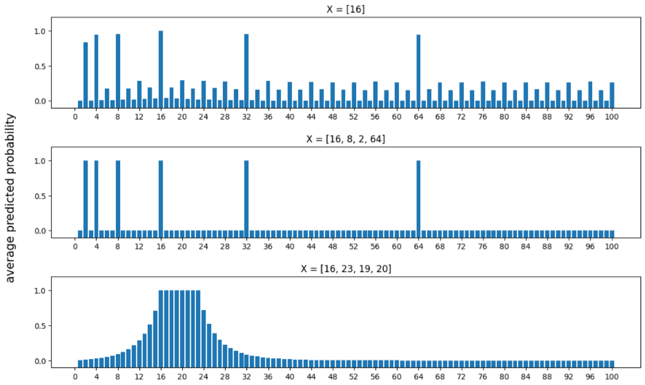

Figure 10: Pure Lambda, Erlang Weighted Empirical Prior Probability for Math Hypothesis