Validation of a Probabilistic Model of Language Acquisition in Children

A dissertation presented by

Kenneth Jerold Straus

to
The Department of Mathematics

In partial fulfillment of the requirements for the degree of Doctor of Philosophy in the field of Mathematics

Northeastern University
Boston, Massachusetts

April, 2008
Validation of a Probabilistic Model of Language Acquisition in Children

by

Kenneth Jerold Straus

ABSTRACT OF DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Mathematics in the Graduate School of Arts and Sciences of Northeastern University, April, 2008
Charles D. Yang has developed a mathematical model of how a child learns his first language. This model, based on Noam Chomsky’s *Principles and Parameters* theory of *Universal Grammar*, represents the actual process of a child learning the parameter settings for the grammar of his native language. The process is one of “natural selection” wherein the correct settings (for his native language) gradually “prevail” and the incorrect ones “die out”. Mathematically, this is expressed as probabilities associated with each of the possible settings; probabilities which give the likelihood that the child will use a particular parameter setting when he parses (makes grammatical sense of) the next sentence he encounters or constructs.

This dissertation proves that the probabilities for the native language settings all converge to 1, which means that, eventually, the child only uses his native grammar to parse sentences.

Put more precisely, for each parameter $\varphi_i$ there is a sequence of random variables \( \{X_{n,i}\} \); and, for each \( n \), \( X_{n,i} \) is the probability that the child chooses the native language setting of $\varphi_i$ for parsing the ‘next’ (the \( n^{th} \)) sentence. We prove that, for each \( i \), \( \{X_{n,i}\} \) converges to 1, almost surely.

Thus, this dissertation lends support to Yang’s mathematical model.
Dedication

I dedicate and offer this dissertation to my beloved Spiritual Master, Avatar Adi Da Samraj. He lifts my heart to His. I am His devotee forever.
... In the usual discussions of such matters, artificial intelligence is presumed to be something generated by computers. In actuality, however, language is the first form of artificial intelligence created by human beings.

... (But) Real Intelligence is tacit (or intrinsically wordless) living existence.

... (And) That Which Is to Be Realized is not in the realm of mind.

— Avatar Adi Da Samraj, from THE ANCIENT WALK-ABOUT WAY, pgs. 55, 56, 58
Acknowledgements

I thank my adviser, Prof. Samuel Gutmann, for his strong support and encouragement during the writing of this dissertation. He is my (mathematical) friend.

And thank Prof. Gordana Todorov for her kind support.

I thank my generous and good-hearted dear sister, Anita Margolis, for her financial and moral support.

I thank my dear wife, Elizabeth DeBrine, for her loving support. I love you, Elizabeth.

I send my love and blessings to my dear son, Jonah Straus.

And my love and blessings to my stepsons Andrew and William Engelbrecht.

And voice my love for my parents, William and Bonnie Straus – grateful for their love and sacrifice for me. They would have been pleased by this accomplishment.

I thank my friends William Lewis, Hugh O’Doherty, John Schorske, Randa Sununu – whose patience and support while I was consumed by this dissertation is much appreciated.

I thank my friends Anna Adams-Bain, Michael Costabile, Robert DeLollis, Mark Harris, Wally Muszynski, Frans van den Bossche, Donovan Train – for their friendship, and encouragement (and Donovan for financial help).

Thanks to Vero Siv and the Nuon family (including Pisey, Mara, and Sam), the owners of the Wellesley Bakery Cafe, where Prof. Gutmann and I had many
Friday-at-2pm meetings.

And I thank the other mathematics professors who taught me what mathematics was about, inspired and encouraged me, and befriended me: At UC-Berkeley – Prof. Charles C. Pugh, Prof. John R. Stallings, Prof. R. Sherman Lehman, Prof. Rainer K. Sachs, Prof. Abraham Seidenberg (who was my dear friend). At UW-Madison – Prof. Walter Rudin. All of them, and Prof. Gutmann, quickened my love of mathematics.
TABLE OF CONTENTS

Abstract 2

Dedication 4

Quotation 5

Acknowledgements 6

Table of Contents 8

Section 1: Introduction - Learning Language 9

Section 2: Mathematical Preliminaries 16

Section 3: Main Section 26

Section 4: Bounding the Expected Time until Convergence 45
  Subsection 4.1: Part I 45
  Subsection 4.2: Part II 52

Section 5: Submartingale Convergence Applies Only for $k \leq 2$ 56

Bibliography 59
1. Introduction - Learning Language

At around 8 months an infant starts babbling. At about 1 year he utters his first words. Sentences appear at a year and a half. By age three most children can form complex sentences and can carry on conversations with adults, “obeying” many “rules” of grammar to express and apprehend thoughts and feelings from simple to subtle. And by age four they have essentially mastered the language; for the rest of their lives they will only be making refinements or adding to their vocabulary.

Every one of us humans, whatever our native language, has gone through this remarkable process in a similar way. And no other animal does. How do we do it?

Imitation? If so, then it is difficult to explain the kinds of grammatical ‘mistakes’ all children tend to make – and the kind they tend not to make. They don’t hear the mistakes they do make from their parents, since adults almost never make the typical grammatical errors so universally made by children. And there are certain grammatical ‘mal-formations’ that children never make, even from the beginning of their language attempts. So it seems unlikely to be just imitation that is at work here.

According to the linguist Noam Chomsky\(^1\) the learning of a first language (or, more precisely, its grammar) is a matter of both biology: the child has an inborn template for grammar; and experience: he uses what he hears to ‘fill in the blanks’ of this ‘universal grammar’ template.

---

\(^1\)My discussion of linguistics in this first section is drawn largely from *The Infinite Gift – How Children Learn and Unlearn the Languages of the World*, by Charles D. Yang (2006) and *Knowledge and Learning in Natural Language*, by Charles D. Yang (2002). Prof. Yang was a student of Noam Chomsky, and Yang’s work, which is the basis for this mathematical paper, builds upon Chomsky’s.
For learning words, imitation is a plausible explanation (although even here our ‘neural hardware’ has ‘built-in assumptions’ that allow a child to filter the flood of information from his environment so as to discern words), but what about putting these words together into sentences - whose rules of structure we call ‘grammar’? This structure gives much of the meaning to language – the relations between the objects and actions denoted by words. Every language has a particular grammar that must be learned. Imitation isn’t up to the task.

The fundamental claim of Chomsky’s linguistic theory is that there is a Universal Grammar for all languages and that every child is born ‘knowing’ it. The child does not learn it from experience, does not imitate it, does not deduce it – it is a capability innate to him, like the capability to walk or see. Chomsky calls Universal Grammar a “language acquisition device” – it enables a child to learn the particular grammar of his native language. An example supporting this claim is in order.

The example\(^2\) deals with how children form questions out of declarative sentences:

Rabbits are nimble.

Are rabbits nimble?

What is the rule for forming a question? Is it ‘Move the second word to the front’? No, since then we could have things like:

Brown rabbits are nimble.

Rabbits brown are nimble?

or

A purple mushroom is growing near the rabbit hole.

Purple a mushroom is growing near the rabbit hole?

\(^2\)From Yang (2006)
In English, to form a question, we cannot move just any word - we need to move *auxiliary* verbs such as “am”, “is”, “are”, “do”, and “does”. So, perhaps the child settles on the hypothesis that one should move the *first* auxiliary verb to the front of the sentence. This seems to work – except for sentences with *two* auxiliaries:

The rabbit that *was* chasing the tortoise was losing the race.

**Was** the rabbit that chasing the tortoise was losing the race?

where it obviously fails.

The principle at work here, and true in all languages, is that the hierarchical relations (and notions like first, second, next, etc.) used in sentence formation are not those between words, but those between *phrases*. Building on this universal principle gives the rule for forming questions from declarative statements specific to English: ‘Move the auxiliary verb that follows the first noun phrase (i.e., the subject phrase) to the front of the sentence.’ So, one would think, hearing questions like the last example shows the child that his previous hypothesis is wrong and leads him toward the correct one. But, research has shown that there’s almost no chance that a child will ever hear a double-auxiliary question. And yet, when such questions are intentionally elicited by experimenters, even very young children (about 3 years old) do not make the mistake, as in the last example, of just moving the first auxiliary verb. Children who’ve learned to move auxiliaries form the question correctly:

The rabbit that was chasing the tortoise *was* losing the race.

**Was** the rabbit that was chasing the tortoise losing the race?

So, to reiterate, the child’s most likely experience does *not* support the right hypothesis, does *not* undermine the wrong (simpler) hypothesis (given above), and does not undermine the many other possible “reasonable” hypotheses (given his
experience) for forming questions. And still, the child who’s learned to move aux-
iliaries, when put into a situation where he must form a question from a double 
auxiliary statement, does so correctly - every time. This strongly suggests an in-
born knowledge of this Universal Grammar\(^3\) principle, called *structure dependence*
by linguists, that the units of hierarchical relations used in sentence formation are
phrases.\(^4\)

And note that it is **only structure dependence** that is shown to be innate here.
The rule in English for moving auxiliary verbs to the front of the sentence when
forming questions is demonstrated to the child in many questions that he hears
(in an English environment). His experience does lead him to that rule. And
adherence to it could not be inborn because it is not universal – children in China
do not adhere to it. So children **will** violate this rule – sometimes even English
learners, in the beginning. And learners of other languages will eventually always
violate it. But children, whatever first language they’re learning, will **never** violate
the rule (or principle) of *structure dependence*.

This is the pattern for all of children’s language learning. In Chomsky’s terms,
there are **principles**, such as structure dependence, and there are **parameters** –
moving the auxiliary verb (as opposed to the main verb) to the front for questions

\(^3\)See Yang (2006) and Yang (2002) for a much fuller discussion and justification of the in-
nateness of Universal Grammar.

\(^4\)A quote from the vision scientist Donald Hoffman that Yang (2006) gives is illuminating
(and follows Hoffman making an explicit connection of his work to Chomsky’s Universal Grammar
theory):

“Kids aren’t taught how to see. Parents don’t sit down with their kids and explain how to use
motion and stereo to construct depth, or how to carve the visual world into objects and actions.
Indeed, most parents don’t know how they do this themselves. And yet it appears that each
normal child comes to construct visual depth, shape, colors, objects, and actions pretty much
the same way as any other normal child. Each normal child, without being taught, reinvents the
visual world; and all do it much the same way. This is remarkable, because in so doing each child
overcomes the fundamental problem of vision. . . . The image at the eye has countless possible
interpretations.” Hoffman (1998)
is a ‘value’ for one of these parameters. The *principles* come from the Universal Grammar and are not learned. The *parameters* involve properties that are different from language to language – they can be thought of as switches that are set one way or another in each language. These settings are what the child learns when learning the grammar of a language. And the particular combination of settings for all these parameters (there are a finite number of parameters, 40 to 50 it is thought; and there are 2 settings for each parameter) is what completely characterizes a language’s grammar.

The parameters *themselves* are part of Universal Grammar. They are the ‘questions’, that can be answered Yes or No, which cover the points of variance between languages (grammars). The ‘questions’ come from Universal Grammar; the ‘answers’, Yes or No, vary from language to language for each question.

For example, there is a “verb-to-complementizer” parameter, which concerns the movement of the (main) verb in a sentence. The movement of verbs in forming questions from statements, discussed earlier, is an example of this property. There are those languages which move the verb – such as Welsh and Zapotec (a native American language of Mexico). And those that do not – such as French and English.\(^6\)

There is a “head-directionality” parameter concerning whether the “head” of a phrase comes at its beginning or end. The head is the key word in the phrase: the noun in a noun phrase, the verb in a verb phrase, the preposition in a prepositional phrase. In English and Edo (a Nigerian language) we have “head-initial”, in Japanese and Navajo, “head-final”. And note that a language has all phrases one way or it has all phrases the other. Here is an example for a verb phrase:

---

\(^5\)Linguists have found that these parameters are binary – they can take one of two settings, Yes or No, On or Off.

\(^6\)English just moves the auxiliary verb when forming questions, the main verb stays put. Earlier forms of English moved the main verb: “Saw you my master?”; “Know you not the cause?”. These are quotes from Shakespeare’s plays.
The boy [grabbed the book]. (English)

versus

Boy [book grabbed]. (Navajo)

“Topic drop” is another parameter - whether the subject can be absent when it can be inferred from context. Chinese is one of the camp that does it. English is in the other camp that does not. So:

Speaker A: Herbert loves Minnie.

Speaker B (in English): But he loved Shirley first.

versus


Topic drop is an “error” that all English-learning children often make early on. They say “Tickles me.” or “Drink tea.” or “But loved Shirley first.”. But, according to our theory, it’s not an error. The child is, in a sense, speaking a foreign language. Chinese, for instance.

And this is evidence that children are ‘trying out’ grammars from other languages. They are making mistakes (the same mistakes for all children) that they do not hear from others and that are perfectly grammatical in another language. Charles Yang’s theory, to which this paper gives mathematical support, is that the learning of the proper settings for all the parameters is a process of ‘natural selection’. Settings (most often those for the child’s native, or target, language) are reinforced by success – in the sense that the likelihood of trying that setting next time increases – and weakened by failure. The combination of settings for the
The reader may take *parse* to basically mean: understand a sentence correctly via its grammatical structure; also, construct a grammatically correct spoken sentence.
2. Mathematical Preliminaries

Our basic model for a child learning his first language is that he hears\textsuperscript{8} one sentence after another and tries to parse each sentence using one of all the ‘possible’ grammars (we explain possible grammars later). All these grammars are “built into” his mind. He either succeeds or fails in parsing a sentence. Which grammar he happens to choose for a particular sentence is a matter of chance, but there are probabilities associated with each grammar at a particular time, so certain grammars are more likely to be chosen than others. These probabilities change after each sentence, depending on his success or failure in parsing that sentence with the grammar chosen. So, for example, if the child tried Chinese grammar for a sentence and succeeded in parsing, then he will be more likely to choose Chinese grammar to parse the next sentence encountered. [Note that a ‘foreign’ grammar can sometimes successfully parse a sentence of the native or target language (we’ll assume the target language here is English, the language of the child’s environment, which he is trying to acquire).] On the other hand, if he fails to parse the sentence using Chinese grammar, then he will be less likely to choose it for the next sentence (A note of clarification: the child is using Chinese grammar applied to the words of the target language – English, in our case. Furthermore, what we precisely mean by ‘the child is using Chinese grammar’ is that he is setting the parameters the way they are set in Chinese; our more picturesque phrasing is for expository purposes.)

We need to make exact our model of grammar here. There are a certain number of parameters that characterize the grammars of all languages (or all possible languages). [See the discussion in Section 1.] These parameters have values ‘yes’\textsuperscript{8}\[\text{I will usually say hear a sentence, but our discussion also applies to speaking, where the child constructs a sentence using one of the grammars.}\]
or ‘no’ (or, 1 or 0), so they are like switches set to ‘on’ or ‘off’ and the combination of settings for all these switches is what determines a particular grammar. All the possible combinations of settings gives all the possible grammars (or languages).

To illustrate, suppose we had only 5 parameters: \( \varphi_1, \ldots, \varphi_5 \). And suppose the grammar of language A is given by

\[
\varphi_1 = 0, \; \varphi_2 = 1, \; \varphi_3 = 1, \; \varphi_4 = 1, \; \varphi_5 = 0,
\]

while the grammar of language B is

\[
\varphi_1 = 0, \; \varphi_2 = 1, \; \varphi_3 = 1, \; \varphi_4 = 0, \; \varphi_5 = 0.
\]

These two languages (grammars) could be called ‘similar’ since they have the same settings for all but one of their parameters. Notice that with 5 parameters, there are \( 2^5 = 32 \) possible languages.

Each of these parameters has a probability associated with it: a probability \( p \) of being ‘on’ and \( 1 - p \) of being ‘off’, at this time. And the probability, at this time, of the child choosing a particular grammar to parse a sentence is the product of these probabilities of the parameter settings for this grammar. So, again using the example of 5 parameters, suppose grammar A has, as before, the settings

\[
\varphi_1 = 0, \; \varphi_2 = 1, \; \varphi_3 = 1, \; \varphi_4 = 1, \; \varphi_5 = 0.
\]

Then, letting \( p_1, \ldots, p_5 \) be the probabilities associated with \( \varphi_1, \ldots, \varphi_5 \), we have the probability of grammar A being chosen for the next sentence as the product

\[
(1 - p_1) p_2 p_3 p_4 (1 - p_5).
\]

And, suppose again that grammar B has the settings

\[
\varphi_1 = 0, \; \varphi_2 = 1, \; \varphi_3 = 1, \; \varphi_4 = 0, \; \varphi_5 = 0,
\]

then the probability for choosing grammar B next is the product

\[
(1 - p_1) p_2 p_3 (1 - p_4) (1 - p_5).
\]

---

\(^9\)Which of the two modes of a parameter we label 1 and which 0 is an arbitrary choice; we will pick a convention later and stick with it.
Now, suppose we choose Grammar B and succeed in parsing the next sentence with it. Then we boost the probability of each parameter setting for Grammar B. In other words, we raise the probability that $\varphi_1$ will be chosen next time as having the setting 0 or ‘off’ [by decreasing the value of $p_1$, and hence increasing the value of $(1 - p_1)$]; we raise the probability that $\varphi_2$ will be chosen next time as having the setting 1 or ‘on’ [by increasing $p_2$]; we raise the probability that $\varphi_3$ will be chosen next time as being ‘on’ [by increasing $p_3$]; we raise the probability that $\varphi_4$ will be chosen next time as being ‘off’ [by decreasing $p_4$]; and we raise the probability that $\varphi_5$ will be chosen next time as being ‘off’ [by decreasing $p_5$]. Since each factor of the product giving the probability for Grammar B has increased, the whole probability for Grammar B has increased.

If, on the other hand, we fail to parse with Grammar B, then we lower the probability that each parameter has its Grammar B setting: we increase $p_1$, decrease $p_2$, decrease $p_3$, increase $p_4$, and increase $p_5$.

The exact adjustment we make to the $p_i$ is given by the following [Notation: $p_{n,i}$ $\equiv$ the current value for $p_i$; $p_{n+1,i} \equiv$ the next value for $p_i$. $\gamma$ is a constant for our model; $0 \leq \gamma \leq 1.$]:

$$p_{n+1,i} = \begin{cases} 
\gamma p_{n,i}, & \text{for decrease} \\
 p_{n,i} + (1 - \gamma)(1 - p_{n,i}), & \text{for increase.}
\end{cases}$$

For the mathematical model to be realistic, eventually the probability of choosing the grammar of the target language must converge to 1 – and the thesis of this paper is that it does. This convergence means that the child finally learns the target language and only uses the grammar of that language.

If we let the parameter settings for the target language all be thought of as ‘on’ [this is the convention we establish], and let $k$ be the total number of parameters, then the probability of the child choosing the target language’s grammar for the
next sentence is written as the product

\[ p_1 p_2 p_3 \ldots p_{k-2} p_{k-1} p_k. \]

So all the factors are of the form \( p_i \).

Thus, for the probability of the target grammar to converge to 1 requires that every \( p_i, \ i = 1, 2, \ldots, k \), converge to 1. At this stage, our model can best be explained in terms of a diagram, but first we modify our notation:

Let \( X_{n,i} \) denote \( p_i \) at the \( n^{th} \) sentence (Note that the \( X_{n,i} \) are random variables whose \underline{values} are probabilities). And let \( \vec{X}_n \) be the vector whose \( k \) elements are \( X_{n,1}, X_{n,2}, \ldots, X_{n,k} \). So we actually have a sequence of random vectors \( \{ \vec{X}_n \} \), a stochastic process. We start the process with \( \vec{X}_0 \) assuming some non-random initial value \( \vec{x}_0 \). (This suffices to imply convergence for any distribution of \( \vec{X}_0 \))

Our stochastic process is a Markov chain:

\[
P\left( \vec{X}_n = \vec{x}_n \mid \vec{X}_{n-1} = \vec{x}_{n-1}, \ldots, \vec{X}_0 = \vec{x}_0 \right) = P\left( \vec{X}_n = \vec{x}_n \mid \vec{X}_{n-1} = \vec{x}_{n-1} \right),
\]

\[ \forall n \) and for all sequences \( \vec{x}_0, \ldots, \vec{x}_{n-1} \) for which

\[
P\left( \vec{X}_{n-1} = \vec{x}_{n-1}, \ldots, \vec{X}_0 = \vec{x}_0 \right) > 0,
\]

because for each value of \( \vec{X}_{n-1} = [X_{n-1,1}, \ldots, X_{n-1,k}] \), there are only two possible values each \( X_{n,i} \) can assume:

\[
\gamma X_{n-1,i} \quad \text{or} \quad X_{n-1,i} + (1 - \gamma)(1 - X_{n-1,i})
\]

And so these give the \( 2^k \) possible values \( \vec{X}_n \) can assume for each value of \( \vec{X}_{n-1} \).

And, furthermore, the probabilities that \( \vec{X}_n \) assumes one of these values are also determined by \( \vec{X}_{n-1} \); we explain this in more detail as we discuss the diagram.

In the diagram on the next page, the process of encountering sentences and choosing a grammar to parse them is pictured as a tree. For each new sentence, we descend through the tree by taking (at random) one of the branches (i.e., choosing one of the grammars to parse the sentence). The probability of taking a particular branch is written next to the branch and changes each time we go through the tree.
In our diagram, there are 3 parameters and hence 8 \((2^3)\) branches (grammars), numbered 1 to 8.

Branch 1 is the grammar of the target language. It has \(x_1 x_2 x_3\) as its probability (all its parameters are ‘on’, by convention, as mentioned before) and, since we cannot fail to parse sentences on it, it has no sub-branches. On all other branches there are two sub-branches (labeled \(a\) and \(b\)), and the probability of succeeding or failing to parse is written next to the two sub-branches — on branch \(j\), for instance, as \(1 - \lambda_j\) and \(\lambda_j\); these probabilities are fixed in time and can differ from branch to branch. At the tip of each sub-branch (or branch for the branch-1 case) is a reference to a box elsewhere on the page which gives the formulas for the adjustments to the random variables after going down that branch and sub-branch (i.e., the change in all the \(X_{n,i}\) after using a particular grammar to parse the sentence and then either succeeding or failing).

For instance, suppose we go down branch 2 and sub-branch \(b\) (i.e., we fail to parse with Grammar 2), then looking at Box 3 we see that the changes in the random variables are:

\[
\begin{align*}
    x_1 &\rightarrow \gamma x_1 \\
    x_2 &\rightarrow \gamma x_2 \\
    x_3 &\rightarrow x_3 + (1 - \gamma)(1 - x_3)
\end{align*}
\]

In order to save space on the diagram, we have changed the notation in these boxes: the arrows in each line above signify the current value of \(X_{n-1,i}\) on the left becoming the new value \(X_{n,i}\) on the right. So, for example, the third line is a shorthand way of saying: \(X_{n,3} = X_{n-1,3} + (1 - \gamma)(1 - X_{n-1,3})\).

After adjusting the probabilities of the parameters, we encounter the next sentence and repeat the process just described – but using these new probabilities – of descending through the grammar tree. Note again, however, that \(\gamma\) and the \(\lambda_j\), \(j = 2, 3, \ldots, 2^k\), (with \(k = 3\) in our diagram) always keep the same value.
Our assertion is that as we keep descending through this tree, again and again, all the $X_{n,i}$ converge to 1 and so the probability of choosing grammar 1, the target grammar, converges to 1. In terms of our diagram, this essentially means that eventually we only go down branch 1, forever after.

We recapitulate the description of our model and set out the notation (some of which has already been introduced above) for the core mathematical section of this paper, which follows.

There are $k$ (grammar) parameters $\varphi_1, \ldots, \varphi_k$ and so there are $2^k$ grammars. Each grammar is characterized by the combination of settings, ‘on’ or ‘off’, of the different parameters. Each parameter has a (current) probability of being on and a complementary probability of being off. Sentences are encountered, one after another, and the child attempts to parse them by randomly choosing one of the grammars. The (current) probability of choosing a grammar is given as the product of the (current) probabilities for this grammar’s settings of the parameters. The probabilities for the parameters, and hence for the grammars, change with each succeeding sentence, as the child parses it, or fails to parse it, with the chosen grammar. For each grammar (except the target grammar), there is a non-zero probability (fixed in time) of failing to parse sentences of the target language.

We picture the process – of choosing a grammar to parse each sentence heard – as descending through a tree with $2^k$ branches, each branch representing a grammar. The probability of choosing a particular branch (for a particular descent), then, is the (current) probability of choosing its grammar. These probabilities are shown next to each branch in the diagram for the tree. Each branch, except branch 1 which represents the target grammar, has two sub-branches (labeled $a$ and $b$), representing success or failure in parsing; the probabilities for these are shown in the diagram next to each sub-branch. At the tip of branch 1 and the tip of all the sub-branches, there is a reference to a box elsewhere in the diagram which gives the changes in the probabilities for the parameters after reaching that tip (i.e.,
after traversing that branch and sub-branch).

The attempt to parse the $n^{th}$ sentence encountered (using one of the grammars) is often referred to as the $n^{th}$ pass (through the grammar probability tree).

We will interchange the use of the terms branch and grammar, usually using the former. ‘Going down [or taking] branch $j$ on the $n^{th}$ pass’ is the same as ‘Using grammar $j$ for the $n^{th}$ sentence’.

$X_{n,i} \equiv$ the random variable whose value is the probability, at the $n^{th}$ pass, of parameter $\varphi_i$ being on.

$X_{n,i}$ is raised or lowered after each pass, depending on whether there is success or failure in parsing the sentence.

$\vec{X}_n \equiv (X_{n,1}, X_{n,2}, \ldots, X_{n,k})$ is the random $k$-vector of all the $X_{n,i}$.

Note that each $X_{0,i}$ has an initial constant value of $x_{0,i}$ and thus $\vec{X}_n$ has as its initial value the constant vector $\vec{x}_0 \equiv (x_{0,1}, x_{0,2}, \ldots, x_{0,k})$. In other words, $\vec{X}_0 = \vec{x}_0$. All these initial constants are between 0 and 1, but otherwise can be chosen arbitrarily.

The $X_{n,i}$, $i = 1 \ldots k$, are discrete random variables, each of them having $2^n$ possible values.

And, thus, the random vector $\vec{X}_n$ is also discrete. It has $2^{nk}$ possible values.
To see that the $X_{n,i}$ are discrete random variables, and hence $\vec{X}_n$ is a discrete random vector, note that there is an initial value $x_{0,i}$ and then:

$X_{1,i}$ can have 2 possible values: $x_{0,i} + (1 - \gamma)(1 - x_{0,i})$ or $\gamma x_{0,i}$.

So $X_{2,i}$ can have $2^2$ possible values at most.

So $X_{3,i}$ can have $2^3$ possible values at most.

\vdots

$X_{n,i}$ can have $2^n$ possible values at most.

Thus $\vec{X}_n$ has $2^{nk}$ possible values at most.

Therefore the $X_{n,i}$ and $\vec{X}_n$ are discrete, for all $n$.

\{ $\vec{X}_n$ \} is a Markov chain.

$A_n \equiv$ the event of going down branch 1 from the $n^{th}$ pass on, i.e. forever.

$B_n[j] \equiv$ the event of going down branch $j$ on the $n^{th}$ pass.

$B_n^c[j] \equiv$ the event of not going down branch $j$ on the $n^{th}$ pass.

$M_n \equiv$ the random variable whose value is the number of the branch with maximum probability for the $n^{th}$ pass.

Each branch, other than branch 1, has sub-branches $a$ and $b$, corresponding to success or failure in parsing sentence $n$ with grammar $j$.

There is probability $1 - \lambda_j$ of success [taking sub-branch $a$] and $\lambda_j$ of failure [taking sub-branch $b$]. These are constant probabilities, i.e., they don’t depend on $n$. We have $B_n[j, a]$ as the event of taking branch $j$ and succeeding on the $n^{th}$ pass; $B_n[j, b]$ as taking branch $j$ and failing on the $n^{th}$ pass.
On branch 1 there are no sub-branches since it represents the grammar of the target language (the language of the child’s environment) and thus can only succeed in parsing.
3. Main Section

BASIC IDEA OF PROOF. Actually, we prove something stronger than just convergence of $\vec{X}_n$ to $\vec{1}$. We prove that from some $N$ onward, we only go down branch 1. In other words, $P(\bigcup_{n=1}^{\infty} A_n) = 1$.

But, on closer scrutiny of the adjustments made to parameters after traversing a branch/sub-branch (see Boxes 1 - 8 on diagram), one can see that $\vec{X}_n$ could not converge to $\vec{1}$ unless, for some $N$, $A_N$ occurred. Otherwise, infinitely often, $\vec{X}_n$ would jump outside any $\epsilon$-neighborhood of $\vec{1}$. So, in our situation, convergence of $\vec{X}_n$ to $\vec{1}$ and $\bigcup_{n=1}^{\infty} A_n$ occurring are equivalent.

The strategy of the proof, then, is to show that eventually we only go down branch 1. We start with a disjoint infinite union of events, each event consisting of going down branch 1 forever, from the $n^{th}$ step on, after not going down branch 1 on the $(n-1)^{th}$ step. The probability of this union is, of course, between 0 and 1. After changing this into a summation, each term of which is a product of probabilities, we spend most of the rest of the proof finding a non-zero lower bound for the first factor of all the terms: $P( A_n \mid B_{n-1}^c[1] )$. Establishing this non-zero lower bound quickly leads, by means of the Borel-Cantelli Lemma, to the result that there must be a last time that we don’t go down branch 1, i.e., a last time that $B_{n-1}^c[1]$ occurs. A rough intuitive way to think about it is: there are endless chances to shoot down branch 1 forever, and they all are bounded away from 0. Thus we are assured that at some point we will go down branch 1 forever after.

A key part of the proof is in showing that the adjustments to the $X_{n,i}$ after going down a branch/sub-branch, regardless of which branch/sub-branch, keep the $X_{n,i}$ above certain lower bounds (See (2.2) and (8.5)). Having this lower bound on $X_n$, in (2.2) for example, allows the infinite product in (3.1) - (3.3).
to converge to a non-zero value. And this infinite product equals the probability that we go down branch 1 forever. An intuitive way to think about this is that every time we go down branch 1, the probability of going down it again increases — and increases by “enough” (enough in the same sense that allows the infinite product to converge to a positive number — see NOTE 8) so that the probability of going down it forever is bounded away from 0. It is important to note that this proof suggests a picture of language learning that is unrealistic. It appears that the child will wander about trying different grammars, making no particular progress, and then suddenly stick to the target grammar forever. This unrealistic scenario is not our intended one. Rather, we see the child gradually using branch 1 more and more, with the use of other grammars appearing less and less, until finally grammar 1 is used exclusively from a certain point onward. Our proof does not preclude this latter scenario, and under stronger assumptions (about the relation between the fail-to-parse probabilities for a grammar and the number of parameters set ‘incorrectly’ in it), we might be able to prove stronger results about the speed of convergence.

We’ve established a baseline here: using a model with only mild conditions, we’ve shown convergence. And this leads to a second point: our method of proof works for more models than the one used here.

Now we move to the theorem that is our focus.

**Theorem 1.** For arbitrary initial distribution of $\vec{X}_0$,

$$\vec{X}_n \to \vec{1}, \text{ a.s., as } n \to \infty.$$  

Equivalently,

$$P \left( \bigcup_{n=1}^{\infty} A_n \right) = 1.$$
Proof. We start with

$$0 \leq P \left( \bigcup_{n=1}^{\infty} \left[ A_n \cap B_{n-1}^c[1] \right] \right) \leq 1$$

Then

$$P \left( \bigcup_{n=1}^{\infty} \left[ A_n \cap B_{n-1}^c[1] \right] \right) = \sum_{n=1}^{\infty} P \left( A_n \mid B_{n-1}^c[1] \right) P \left( B_{n-1}^c[1] \right),$$

since the union is disjoint.

Now, focusing on the first factor in each term of the summation:

$$P \left( A_n \mid B_{n-1}^c[1] \right)$$

$$= P \left( A_n \cap \{ M_{n-1} \neq 1 \} \mid B_{n-1}^c[1] \right) + P \left( A_n \cap \{ M_{n-1} = 1 \} \mid B_{n-1}^c[1] \right)$$

$$= P \left( A_n \mid M_{n-1} \neq 1, B_{n-1}^c[1] \right) P \left( M_{n-1} \neq 1 \mid B_{n-1}^c[1] \right) +$$

$$P \left( A_n \mid M_{n-1} = 1, B_{n-1}^c[1] \right) P \left( M_{n-1} = 1 \mid B_{n-1}^c[1] \right)$$

$$= r P \left( A_n \mid M_{n-1} \neq 1, B_{n-1}^c[1] \right) +$$

$$(1 - r) P \left( A_n \mid M_{n-1} = 1, B_{n-1}^c[1] \right)$$

// where $r \equiv P \left( M_{n-1} \neq 1 \mid B_{n-1}^c[1] \right) ; r$ is used to simplify notation //

$$\geq \min \left[ P \left[ A_n \mid M_{n-1} \neq 1, B_{n-1}^c[1] \right], P \left[ A_n \mid M_{n-1} = 1, B_{n-1}^c[1] \right] \right] \quad (0)$$
3. MAIN SECTION

NOTE 1
In the first equality above, we divide $A_n$ into two events by intersecting it with the event where branch 1 did not have the highest probability (on the $(n - 1)^{th}$ step) and the event where it did. Our goal is to get an estimate (a lower bound) for $P(A_n \mid B_{n-1}^c[1])$; we keep “breaking it down into pieces” (by sub-events) to accomplish this.

NOTE 2 - FACT 1
For the second equality above, we use the fact that $\forall$ events $U, V, W$: $P(U, V \mid W) = P(U \mid V, W) P(V \mid W)$.

NOTE 3 - FACT 2
The inequality in (0) follows from the fact that $\forall u, v \in \mathbb{R}$ and with $0 \leq r \leq 1$: $ru + (1 - r)v \geq \min \left[ u, v \right]$.

First we consider, from (0), $P \left( A_n \mid M_{n-1} \neq 1, B_{n-1}^c[1] \right)$, which is

\[
\geq P \left( A_n \cap B_{n-1}[M_{n-1}, b] \mid M_{n-1} \neq 1, B_{n-1}^c[1] \right) \tag{0.5}
\]

\[= P \left( A_n \mid B_{n-1}[M_{n-1}, b], M_{n-1} \neq 1, B_{n-1}^c[1] \right) \cdot P \left( B_{n-1}[M_{n-1}, b] \mid M_{n-1} \neq 1, B_{n-1}^c[1] \right) \tag{1} \]

// Using Fact 1 again

NOTE 4
In (0.5), we intersect $A_n$ with $B_{n-1}[M_{n-1}, b]$. We define $B_{n-1}[M_{n-1}, b]$ as the event that, on the previous pass, we went down the branch that had maximum probability, and failed to parse. This intersecting creates a smaller event, with lower probability.
NOTE 5
Explaining (1): If we take the max branch and we don’t take branch 1, then the max branch is not branch 1: \( B_{n-1}[M_{n-1}] \cap B_{n-1}^c[1] \implies M_{n-1} \neq 1 \).
So, \( \{B_{n-1}[M_{n-1}] \cap B_{n-1}^c[1]\} \subset \{M_{n-1} \neq 1\} \),
which \( \implies \{B_{n-1}[M_{n-1}] \cap B_{n-1}^c[1] \cap M_{n-1} \neq 1\} = \{B_{n-1}[M_{n-1}] \cap B_{n-1}^c[1]\} \)
The added condition of failing to parse, ‘b’ in the square brackets, doesn’t change anything, since: if \( U \subset W \), then \( U \cap V \subset W \), \( \forall U, V, W \).

Now, consider the first factor in (1): \( P(A_n \mid B_{n-1}[M_{n-1}, b], B_{n-1}^c[1]) \). We build toward an estimate of this quantity, i.e., obtain a lower bound.

If we go down the \([M_{n-1}, b]\) branch and sub-branch [i.e., the maximum branch and then fail] on the \((n-1)^{th}\) pass, then (recalling the definition of \( \vec{X}_n \) from Section 2)

\[
\vec{X}_n \geq \vec{\alpha}_1, \quad \text{where } \alpha_1 \equiv \min \left[ \frac{\gamma}{2}, 1 - \gamma \right] \text{ and } \vec{\alpha}_1 \equiv \left( \alpha_1, \alpha_1, \ldots, \alpha_1 \right).
\]

NOTATION: \( \vec{X}_n \geq \vec{\alpha}_1 \) means that every element of the first vector is \( \geq \) than the corresponding element of the second. And, as shown above, we’ve created a vector of \( k \) identical elements out of the value \( \alpha_1 \), labeling it \( \vec{\alpha}_1 \).

To see this, note that

\[
P \left( B_{n-1}[M_{n-1}] \mid \vec{X}_{n-1} = \vec{x}_{n-1} \right) = \prod_{i=1}^{k} f_{n-1,i},
\]
each factor \( f_{n-1,i} \) being of the form \( x_{n-1,i} \) or \( (1 - x_{n-1,i}) \) [see the tree diagram].
Now, since we’re taking the maximum probability branch, each \( f_{n-1,i} \) would have to be \( \geq \frac{1}{2} \). Hence,

\[
f_{n-1,i} = \begin{cases} 
\vec{x}_{n-1,i}, & \text{if } x_{n-1,i} \geq \frac{1}{2}; \\
1 - \vec{x}_{n-1,i}, & \text{if } x_{n-1,i} < \frac{1}{2}.
\end{cases}
\]
And then after going down the branch and failing — $B_{n-1}[M_{n-1}, b]$ — we must have [again, see the tree diagram]

$$
x_{n,i} = \begin{cases} 
\gamma x_{n-1,i}, & \text{if } f_{n-1,i} = x_{n-1,i}; \\
x_{n-1,i} + (1 - \gamma)(1 - x_{n-1,i}), & \text{if } f_{n-1,i} = 1 - x_{n-1,i}.
\end{cases}
$$

But in the first case, $f_{n-1,i} = x_{n-1,i}$ means that $x_{n-1,i} \geq \frac{1}{2}$, and that makes $x_{n,i} \geq \frac{1}{2}\gamma$.

And in the second case, since $x_{n-1,i} + (1 - \gamma)(1 - x_{n-1,i}) = (1 - \gamma) + \gamma x_{n-1,i}$, we have $x_{n,i} \geq 1 - \gamma$.

Thus, we have

$$x_{n,i} \geq \alpha_1, \text{ for } i = 1, 2, \ldots, k, \text{ where, again, } \alpha_1 = \min \left[ \frac{\gamma}{2}, 1 - \gamma \right]$$

Or, $\vec{X}_n \geq \vec{\alpha}_1$. \hspace{1cm} (2)

So, to sum up, $B_{n-1}[M_{n-1}, b] \implies \vec{X}_n \geq \vec{\alpha}_1$.

Therefore, $B_{n-1}[M_{n-1}, b] \cap B_{n-1}^c[1] \subseteq \{\vec{X}_n \geq \vec{\alpha}_1\}$, \hspace{1cm} (2.1)

implying $P \left( \vec{X}_n \geq \vec{\alpha}_1 \mid B_{n-1}[M_{n-1}, b], B_{n-1}^c[1] \right) = 1$. \hspace{1cm} (2.2)
Now, \( P(A_n \mid \vec{X}_n) = \)
\[
\prod_{j=0}^{\infty} \left[ \prod_{i=1}^{k} (1 - \gamma^j (1 - X_{n,i})) \right] \geq \prod_{j=0}^{\infty} \left[ \prod_{i=1}^{k} (1 - \gamma^j (1 - \alpha_1)) \right] \quad (3.1)
\]
// on \( \{ \vec{X}_n \geq \vec{\alpha}_1 \} \)
\[
= \prod_{j=0}^{\infty} [1 - \gamma^j (1 - \alpha_1)]^k \quad (3.2)
\]
\[
= \eta_1 > 0 \quad (3.3)
\]
// \( \eta_1 \) is some positive number < 1

And then \( P\left( A_n \mid \vec{X}_n \right) \geq \eta_1 \) when \( \vec{X}_n \geq \vec{\alpha}_1 \), implies
\[
P\left( A_n \mid \vec{X}_n \geq \vec{\alpha}_1 \right) \geq \eta_1. \quad \text{// See NOTE 8.1} \quad (3.4)
\]

NOTE 6

In regard to the expression inside the first double product in (3.1):
The event \( A_n \) consists of going down branch 1 forever after pass \( n \), so the successive values of \( X_{n,i} \) \((i = 1, \ldots, k)\) after \( n \) (the product of which, \( \prod_{i=1}^{k} X_{n+j,i} \), for each \( j \), is the probability of taking branch 1 on the next pass, \( j + 1 \)) are
\[
X_{n+1,i} = (1 - \gamma) + \gamma X_{n,i}
= 1 - \gamma (1 - X_{n,i})
\]
\[
X_{n+2,i} = (1 - \gamma) + \gamma X_{n+1,i}
= 1 - \gamma^2 (1 - X_{n,i})
\]
\[
\vdots
\]
\[
X_{n+j,i} = 1 - \gamma^j (1 - X_{n,i})
\]
NOTE 7

In regard to the inequality in (3.1), we have from (2), \( \forall i: \)

\[
X_{n,i} \geq \alpha_1,
\]

which leads to

\[
1 - X_{n,i} \leq 1 - \alpha_1
\]

\[
-\gamma^j(1 - X_{n,i}) \geq -\gamma^j(1 - \alpha_1)
\]

\[
1 - \gamma^j(1 - X_{n,i}) \geq 1 - \gamma^j(1 - \alpha_1)
\]

NOTE 8

With regard to the infinite product in (3.2) having a positive value \( \eta_1 \) in (3.3):

An infinite product \( \prod_{j=0}^{\infty} (1 - a_j) \), \( 0 \leq a_j < 1 \), converges to a non-zero value \( \eta \) iff \( \sum_{j=0}^{\infty} a_j \) converges. In our case, \( a_j = k(1 - \alpha_1)\gamma^j \) and

\[
\sum_{j=0}^{\infty} k(1 - \alpha_1)\gamma^j = k(1 - \alpha_1) \sum_{j=0}^{\infty} \gamma^j = \frac{k(1 - \alpha_1)}{1 - \gamma},
\]

since \( \sum \gamma^j \) is a geometric series.
NOTE 8.1

Proving the assertion in (3.4):

\[ P \left( A_n \mid \bar{X}_n \geq \bar{\alpha}_1 \right) \]

\[ = \frac{P \left( A_n \cap \{ \bar{X}_n \geq \bar{\alpha}_1 \} \right)}{P \left( \bar{X}_n \geq \bar{\alpha}_1 \right)} \]

\[ = \frac{\sum_{\bar{x} \geq \bar{\alpha}_1} P \left( A_n \cap \{ \bar{X}_n = \bar{x} \} \right)}{\sum_{\bar{x} \geq \bar{\alpha}_1} P \left( \bar{X}_n = \bar{x} \right)} \]

\[ = \frac{\sum_{\bar{x} \geq \bar{\alpha}_1} P \left( \bar{X}_n = \bar{x} \right) P \left( A_n \mid \bar{X}_n = \bar{x} \right)}{\sum_{\bar{x} \geq \bar{\alpha}_1} P \left( \bar{X}_n = \bar{x} \right)} \]

\[ \geq \frac{\sum_{\bar{x} \geq \bar{\alpha}_1} P \left( \bar{X}_n = \bar{x} \right) \eta_1}{\sum_{\bar{x} \geq \bar{\alpha}_1} P \left( \bar{X}_n = \bar{x} \right)} \] // \( P(A_n \mid \bar{X}_n) \geq \eta_1 \) on \( \bar{X}_n \geq \bar{\alpha}_1 \)

\[ = \eta_1 \]

So, summarizing, we have

\[ P \left( \bar{X}_n \geq \bar{\alpha}_1 \left| B_{n-1}[M_{n-1}, b], B_{n-1}^c[1] \right. \right) = 1 \]

and

\[ P(A_n \mid \bar{X}_n \geq \bar{\alpha}_1) \geq \eta_1 > 0. \]

But, we want:

\[ P \left( A_n \left| B_{n-1}[M_{n-1}, b], B_{n-1}^c[1] \right. \right) \geq \eta_1. \]

The key to this is that \( \{ \bar{X}_n \} \) is Markov (see NOTE 10). We proceed as follows:

First, for simpler notation, let \( C_{n-1} \equiv B_{n-1}[M_{n-1}, b] \cap B_{n-1}^c[1] \).
Then,

\[ P \left( A_n \mid C_{n-1} \right) = \sum_{i=1}^{(2^n)^k} P \left( A_n, \bar{X}_n = \bar{x}_i \mid C_{n-1} \right) \]  

(4.1)

\[ = \sum_{i=1}^{2^n k} P \left( A_n \mid \bar{X}_n = \bar{x}_i, C_{n-1} \right) P \left( \bar{X}_n = \bar{x}_i \mid C_{n-1} \right) \]  

(4.2)

\[ = \sum_{i=1}^{2^n k} P \left( A_n \mid \bar{X}_n = \bar{x}_i \right) P \left( \bar{X}_n = \bar{x}_i \mid C_{n-1} \right) \]  

(4.3)

\[ \geq \sum_{\bar{x}_i \geq \bar{\alpha}_1} P \left( A_n \mid \bar{X}_n = \bar{x}_i \right) P \left( \bar{X}_n = \bar{x}_i \mid C_{n-1} \right) \]  

(4.4)

\[ = \sum_{\bar{x}_i \geq \bar{\alpha}_1} g(\bar{x}_i) P \left( \bar{X}_n = \bar{x}_i \mid C_{n-1} \right) \]  

(4.5)

\[ \geq \sum_{\bar{x}_i \geq \bar{\alpha}_1} g(\bar{\alpha}_1) P \left( \bar{X}_n = \bar{x}_i \mid C_{n-1} \right) \]  

(4.6)

\[ = \eta_1 \sum_{\bar{x}_i \geq \bar{\alpha}_1} P \left( \bar{X}_n = \bar{x}_i \mid C_{n-1} \right) \]  

(4.7)

\[ = \eta_1. \]  

(4.8)

**NOTE 9**

Regarding the summation’s upper limit in (4.1):

For \( i = 1, \ldots, k \), \( X_{n,i} \) has \( 2^n \) possible values. So, \( \bar{X}_n \) has \( (2^n)^k \) possible values.
NOTE 10
In going from (4.2) to (4.3) we drop the second condition, $C_{n-1}$, from the first probability. This is allowed since the Markov property of our process makes the second condition superfluous.

NOTE 11
In (4.5), $g(\bar{x}) \equiv \prod_{j=0}^{\infty} \left[ \prod_{s=1}^{k} \left( 1 - \gamma^j (1 - x_s) \right) \right]$.

NOTE 12
In (4.6), we have the inequality because $g$ is an increasing function and $\bar{x}_i \geq \bar{\alpha}_1$.

NOTE 13
Explaining the transition from (4.6) to (4.7):
$$g(\alpha_1) = \prod_{j=0}^{\infty} \left[ 1 - \gamma^j (1 - \alpha_1) \right]^k = \eta_1$$

NOTE 14
Justifying going from (4.7) to (4.8):
Recalling (2.1) and the def'n of $C_{n-1}$, we have $C_{n-1} \subset \{ \bar{X}_n \geq \bar{\alpha}_1 \}$.
So,
$$\bigcup_{\bar{x}_i \geq \bar{\alpha}_1} \left[ \{ \bar{X}_n = \bar{x}_i \} \cap C_{n-1} \right] = C_{n-1}.$$ 

Hence,
$$\sum_{\bar{x}_i \geq \bar{\alpha}_1} P(\bar{X}_n = \bar{x}_i \mid C_{n-1}) = \sum_{\bar{x}_i \geq \bar{\alpha}_1} \frac{P(\bar{X}_n = \bar{x}_i \mid C_{n-1})}{P(C_{n-1})} = \frac{P(C_{n-1})}{P(C_{n-1})} = 1.$$ 

So, as desired, we have
$$P \left( A_n \mid B_{n-1}[M_{n-1}, b], B_{n-1}^c[1] \right) \geq \eta_1.$$ (4.9)
3. MAIN SECTION

Now for the second factor in (1):

We claim that

**Lemma 1.**

\[
P \left( B_{n-1} [M_{n-1}, b] \mid M_{n-1} \neq 1, B_{n-1}^c[1] \right) \geq \frac{\lambda_0}{2^k},
\]

where \( \lambda_0 = \min \{ \lambda_2, \lambda_3, \ldots, \lambda_{2k} \} \)

**Proof.**

\[
P \left( B_{n-1} [M_{n-1}] \mid M_{n-1} \neq 1, B_{n-1}^c[1] \right)
\]

\[
= \frac{P \left( B_{n-1} [M_{n-1}], M_{n-1} \neq 1, B_{n-1}^c[1] \right)}{P \left( M_{n-1} \neq 1, B_{n-1}^c[1] \right)} \tag{6.1}
\]

\[
= \frac{P \left( B_{n-1} [M_{n-1}], M_{n-1} \neq 1 \right)}{P \left( M_{n-1} \neq 1, B_{n-1}^c[1] \right)} \tag{6.2}
\]

\[
\geq \frac{P \left( B_{n-1} [M_{n-1}], M_{n-1} \neq 1 \right)}{P \left( M_{n-1} \neq 1 \right)} \tag{6.3}
\]

**Note 15**

Explanation of dropping \( B_{n-1}^c[1] \) from numerator in going from line (6.1) to (6.2):

If we take the max. branch and the max. branch is not branch 1, then we don’t take branch 1. I.e.,

\[ B_{n-1} [M_{n-1}] \cap M_{n-1} \neq 1 \Rightarrow B_{n-1}^c[1] \]

So, \( \{ B_{n-1} [M_{n-1}] \cap M_{n-1} \} \subset B_{n-1}^c[1] \)

So, \( B_{n-1} [M_{n-1}] \cap M_{n-1} \cap B_{n-1}^c[1] = B_{n-1} [M_{n-1}] \cap M_{n-1} \).
NOTE 16

Regarding the inequality in (6.3):
For any \( U, V \), \( P(U \cap V) \leq P(U) \).
And then for any \( W \),
\[
\frac{P(W)}{P(U \cap V)} \geq \frac{P(W)}{P(U)}.
\]

Now, since the event \( \{ M_{n-1} = 1 \} \) occurs iff each \( X_{n-1,i} \geq \frac{1}{2} \), the event \( \{ M_{n-1} \neq 1 \} \) occurs iff at least one of the \( X_{n-1,i} < \frac{1}{2} \), i.e., \( \bar{X}_{n-1} \not\geq \frac{1}{2} \). So, the event \( \{ M_{n-1} \neq 1 \} \) consists of the disjoint union of the events \( \{ \bar{X}_{n-1} = \bar{x}_s \} \), where \( s \) runs through all vectors \( \bar{x}_s \) that have \( \bar{x}_s \not\geq \frac{1}{2} \) [but are, of course, between 0 and 1].

Thus, continuing with the numerator from (6.3):
\[
P(B_{n-1}[M_{n-1}], M_{n-1} \neq 1) = P \left( B_{n-1}[M_{n-1}] \cap \bigcup_{\bar{x}_s \not\geq \frac{1}{2}} \{ \bar{X}_{n-1} = \bar{x}_s \} \right)
\]
\[
= P \left( \bigcup_{\bar{x}_s \not\geq \frac{1}{2}} \{ B_{n-1}[M_{n-1}], \bar{X}_{n-1} = \bar{x}_s \} \right)
\]
\[
= \sum_{\bar{x}_s \not\geq \frac{1}{2}} P \left( B_{n-1}[M_{n-1}], \bar{X}_{n-1} = \bar{x}_s \right)
\]
\[
= \sum_{\bar{x}_s \not\geq \frac{1}{2}} \left[ P \left( B_{n-1}[M_{n-1}] \mid \bar{X}_{n-1} = \bar{x}_s \right) P \left( \bar{X}_{n-1} = \bar{x}_s \right) \right]
\]
\[
\geq \frac{1}{2k} \sum_{\bar{x}_s \not\geq \frac{1}{2}} P \left( \bar{X}_{n-1} = \bar{x}_s \right) \quad // \text{See NOTE 17.} \]
\[= \frac{1}{2^k} P(M_{n-1} \neq 1)\]

NOTE 17

\[P\left( B_{n-1}[M_{n-1}] \mid \bar{X}_{n-1} = \bar{x}_s \right) = \prod_{j=1}^{k} f_{s,j}\]

where \(f_{s,j} = \begin{cases} x_{s,j}, & \text{if } x_{s,j} \geq \frac{1}{2}; \\ 1 - x_{s,j}, & \text{if } x_{s,j} < \frac{1}{2}. \end{cases}\)

Since we’re taking the max. probability branch (given \(\bar{X}_{n-1} = \bar{x}_s\)),
each factor \(f_{s,j}\) in the probability expression for that branch will be \(\geq \frac{1}{2}\).
Thus, the whole probability will be \(\geq \frac{1}{2^k}\).

Now replacing the numerator in (6.3) with the last expression and then simplifying the fraction, we get (from (6.1) - (6.3)):

\[P\left( B_{n-1}[M_{n-1}] \mid M_{n-1} \neq 1, B_{n-1}^c[1] \right) \geq \frac{1}{2^k}.\]

Moving on to the third part of the lemma’s proof,

\[P\left( B_{n-1}[M_{n-1}, b] \mid M_{n-1} \neq 1, B_{n-1}^c[1] \right)\]

\[= P\left( B_{n-1}[M_{n-1}] \cap D_{n-1} \mid M_{n-1} \neq 1, B_{n-1}^c[1] \right)\] \hspace{1cm} (8.1)

// \(D_{n-1}\): see NOTE 18

\[= P\left( D_{n-1} \mid B_{n-1}[M_{n-1}], M_{n-1} \neq 1, B_{n-1}^c[1] \right).\] \hspace{1cm} (8.2)
\[ P \left( B_{n-1}[M_{n-1}] \, | \, M_{n-1} \neq 1, B_{n-1}[1] \right) \]

\[ = P \left( D_{n-1} \, | \, B_{n-1}[M_{n-1}], B_{n-1}[1] \right) \cdot P \left( B_{n-1}[M_{n-1}] \, | \, M_{n-1} \neq 1, B_{n-1}[1] \right) \quad (8.3) \]

\[ \geq \lambda_0 \cdot \frac{1}{2^k} \quad (8.4) \]

So the lemma, (5), is proved. \(\square\)

**NOTE 18**
(8.1): We label 'failure on the \((n - 1)^{th}\) pass' as \(D_{n-1}\). This is the same as going down the \(b\) sub-branch, i.e.,

\[ D_{n-1} \equiv \bigcup_{j=2}^{2^k} B_{n-1}[j, b] \]

**NOTE 19**
(8.3):

\[ B_{n-1}[M_{n-1}] \cap B_{n-1}^c[1] \subseteq \{M_{n-1} \neq 1\} \]

**NOTE 20**
(8.4): We’ve already established

\[ P \left( B_{n-1}[M_{n-1}] \, | \, M_{n-1} \neq 1, B_{n-1}^c[1] \right) \geq \frac{1}{2^k}. \]

And, recalling that \(\lambda_0 = \min \{\lambda_2, \lambda_3, \ldots, \lambda_{2^k}\}\), the probability of failure after going down any branch will be greater than \(\lambda_0\). So,

\[ P \left( D_{n-1} \, | \, B_{n-1}[M_{n-1}], B_{n-1}^c[1] \right) \geq \lambda_0. \]
We combine (5), (4.9), and (1), to get

\[
P \left( A_n \mid M_{n-1} \neq 1, B_{n-1}^c[1] \right) \geq \frac{\eta_1 \lambda_0}{2^k}.
\]

This takes care of the first probability in (0). Now, to estimate the second probability in (0):

\[
P \left( A_n \mid \{M_{n-1} = 1\}, B_{n-1}^c[1] \right)
\]

First,

\[
\{M_{n-1} = 1\} = \{\vec{X}_{n-1} \geq \frac{1}{2}\}.
\]

Then, note that \( M_{n-1} = 1 \) implies

\[
\vec{X}_n \geq \frac{1}{2} \gamma,
\]

since the least possible value for \( \vec{X}_n \) occurs after going down branch\([2^k, a]\) (on the \((n - 1)^{th}\) pass), and this yields

\[
X_{n,i} = \gamma X_{n-1,i} \quad i = 1, \ldots, k
\]

\[
\geq \gamma \cdot \frac{1}{2}
\]

Now, let \( \alpha_2 \equiv \frac{\gamma}{2} \). Then we have

\[
\{M_{n-1} = 1\} = \left\{\vec{X}_{n-1} \geq \frac{1}{2}\right\} 
\subseteq \left\{\vec{X}_n \geq \alpha_2 \right\}
\]

And so then we also have,

\[
\{M_{n-1} = 1\} \cap B_{n-1}^c[1] \subseteq \left\{\vec{X}_n \geq \alpha_2 \right\}.
\]

Let \( C_{n-1} \equiv \{M_{n-1} = 1\} \cap B_{n-1}^c[1] \), to simplify notation.
Then,

\[ P \left( A_n \mid C_{n-1} \right) = \sum_{i=1}^{2^{nk}} P \left( A_n, \bar{X}_n = \bar{x}_i \mid C_{n-1} \right) \] (9.1)

\[ = \sum_{i=1}^{2^{nk}} P \left( A_n \mid \bar{X}_n = \bar{x}_i, C_{n-1} \right) P \left( \bar{X}_n = \bar{x}_i \mid C_{n-1} \right) \] (9.2)

\[ = \sum_{i=1}^{2^{nk}} P \left( A_n \mid \bar{X}_n = \bar{x}_i \right) P \left( \bar{X}_n = \bar{x}_i \mid C_{n-1} \right) \] (9.3)

\[ \geq \sum_{\bar{x}_i \geq \bar{a}_2} P \left( A_n \mid \bar{X}_n = \bar{x}_i \right) P \left( \bar{X}_n = \bar{x}_i \mid C_{n-1} \right) \] (9.4)

\[ = \sum_{\bar{x}_i \geq \bar{a}_2} g(\bar{x}_i) P \left( \bar{X}_n = \bar{x}_i \mid C_{n-1} \right) \] (9.5)

\[ \geq \sum_{\bar{x}_i \geq \bar{a}_2} g(\alpha_2) P \left( \bar{X}_n = \bar{x}_i \mid C_{n-1} \right) \] (9.6)

\[ = \eta_2 \sum_{\bar{x}_i \geq \bar{a}_2} P \left( \bar{X}_n = \bar{x}_i \mid C_{n-1} \right) \] (9.7)

\[ = \eta_2 \] (9.8)

**NOTE 20**

(9.3): Here we’ve used the Markov property of our stochastic process, \( \{X_n\} \), to drop the second condition, \( C_{n-1} \), as superfluous.
NOTE 21

(9.7):
\[ \eta_2 \equiv g(\alpha_2) = \prod_{j=0}^{\infty} \left[ 1 - \gamma^j (1 - \alpha_2) \right]^k \]
and \( 0 < \eta_2 < \infty \).

See NOTE 8 for more about this infinite product. Also, see NOTES 11 and 12 regarding (9.5) and (9.6).

Overall, (9.1) - (9.8) follows the same pattern as (4.1) - (4.8).

NOTE 22

On the transition from (9.7) to (9.8):

Note that \( C_{n-1} \subset \{ \bar{X}_n \geq \bar{\alpha}_2 \} \). Now see NOTE 14.

So, we have the estimate for the second probability in (0):
\[ P \left( A_n \left| M_{n-1} = 1, B_{n-1}[1] \right. \right) \geq \eta_2 > 0. \]

Since \( \frac{\eta_1 \lambda_0}{2^k} < \eta_1 \leq \eta_2 \) (See NOTE 23), let
\[ \eta \equiv \frac{\eta_1 \lambda_0}{2^k} \]

Then,
\[ \forall n, \ P \left( A_n \left| B_{n-1}[1] \right. \right) \geq \eta > 0. \]

Therefore
\[ \sum_{n=1}^{\infty} \left[ P \left( A_n \left| B_{n-1}[1] \right. \right) P \left( B_{n-1}[1] \right) \right] \geq \eta \sum_{n=1}^{\infty} P \left( B_{n-1}[1] \right). \]
This implies
\[ \sum_{n=1}^{\infty} P \left( B_{n-1}^c[1] \right) < \infty, \]
which, by the Borel-Cantelli Lemma, implies
\[ P \left( B_{n-1}^c[1], \text{ i.o.} \right) = 0 \]
and, thus,
\[ P \left( \exists N \text{ s.t., a.s., } B_{N-1}^c[1] \text{ is the last time } B_{n-1}^c[1] \text{ occurs} \right) = 1. \]

In other words, \( P \left( \bigcup_{n=1}^{\infty} A_n \right) = 1 \), and so \( \vec{X}_n \to \vec{1} \), a.s., as \( n \to \infty \), and we converge to the target language. Or, put another way, from a certain point on, the child never uses any grammar but the right one.

\[ \square \]

**NOTE 23**

Showing that \( \eta_1 \leq \eta_2 \):

\[ \alpha_1 = \min \left[ \frac{\gamma}{2}, 1 - \gamma \right] \quad \text{and} \quad \alpha_2 = \gamma/2 \]

So, \( \alpha_1 \leq \alpha_2 \)

\[ 1 - \alpha_1 \geq 1 - \alpha_2 \]

\[ 1 - \gamma \left( 1 - \alpha_1 \right) \leq 1 - \gamma \left( 1 - \alpha_2 \right) \]

\[ \eta_1 = \prod_{j=0}^{\infty} \left[ 1 - \gamma^j (1 - \alpha_1) \right]^k \leq \prod_{j=0}^{\infty} \left[ 1 - \gamma^j (1 - \alpha_2) \right]^k = \eta_2 \]
4. Bounding the Expected Time until Convergence

4.1. PART I. The expected value for the number of sentences (or "steps") until a child learns his native language – i.e., until he begins to use only the target grammar to parse sentences – can be upper-bounded:

\[ E[\# \text{ of steps until learn target grammar}] \leq \frac{2^k}{\eta_1 \lambda_0} \left( 1 + \frac{k \gamma}{(1 - \gamma)^2} \right). \]

We derive this inequality as follows.

Each time that we do not take branch 1 can be thought of as marking the beginning of a trial that may lead to the "success" of taking branch 1 forever from that point on. The probability of success on such a trial is given by \( P \left( A_n \mid X_n \right) \), which we’ve shown to be \( \geq \frac{\eta_1 \lambda_0}{2^k} \) on \( B_{n-1} \), i.e. if we didn’t take branch 1 on the previous step. If we let \( L \) be the number of these trials before the language is learned (i.e., the number of unsuccessful trials), then

\[ EL \leq \frac{2^k}{\eta_1 \lambda_0} - 1, \]

since \( L + 1 \) is stochastically smaller than a geometric distribution with success probability \( \frac{\eta_1 \lambda_0}{2^k} \), and the expected number of trials until the first success for such a distribution is the reciprocal of the probability of success on any trial — so,

\[ E[L + 1] \leq \frac{1}{P(\text{success})} = \frac{2^k}{\eta_1 \lambda_0}. \]

Hereafter, we leave out the ‘−1’ in the estimate for \( EL \), since it is negligible.

[[ Also, from now on we re-label \( \eta_1 \) as \( \eta_0^k \). Recall from (3.3) that \( \eta_1 \) is a positive number equal to \( \prod_{j=0}^{\infty} [1 - \gamma^j(1 - \alpha_1)]^k \). So, \( \eta_0 \) is obtained when we set \( k = 1 \) in this infinite product. ]]}
Note that the \((L + 1)^{th}\) trial is successful; and it is the last trial. A picture of
the situation, where ‘+’ = ‘took branch 1’ and ‘−’ = ‘didn’t take branch 1’
and \(T_j\) is the number of steps in the \(j^{th}\) trial, is

\[
\begin{array}{cccccccccccc}
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
- & - & - & + & + & + & + & + & + & + & + & \cdots \\
\text{trial 3} & \text{trial L+1, success} \\
\end{array}
\]

Let \(S_L\) be the number of steps until the target grammar is learned, i.e., until
success. Then

\[
S_L = \sum_{j=1}^{L} (1 + T_j).
\]

Note that \(T_j\) can equal 0.

First we estimate the expectation of the number of steps in any trial \(T\), assum-
ing it’s unsuccessful:

\[
E \left( T \middle| T < \infty \right) \quad \text{// Note that we calculate this conditional on}
\]

\[
\text{// } X_{n,i}, \text{ but get an estimate independent of } X_{n,i}.
\]

\[
= \sum_{t=1}^{\infty} P \left( T \geq t \middle| T < \infty \right) \quad \text{// } E Y = \sum_{y=1}^{\infty} P(Y \geq y) \text{ when } Y \text{ has values 0, 1, 2,..}
\]

\[
= \frac{\sum_{t=1}^{\infty} P \left( t \leq T < \infty \right)}{P \left( T < \infty \right)}
\]

\[
= \frac{\sum_{t=1}^{\infty} \left[ P \left( T \geq t \right) - P \left( T = \infty \right) \right]}{P \left( T < \infty \right)}
\]

\[
= \frac{\sum_{t=1}^{\infty} \left[ \prod_{j=0}^{k} \prod_{i=1}^{t} \left[ 1 - \gamma^j \left( 1 - X_{n,i} \right) \right] - \prod_{j=0}^{k} \prod_{i=1}^{\infty} \left[ 1 - \gamma^j \left( 1 - X_{n,i} \right) \right] \right]}{1 - \prod_{j=0}^{\infty} \prod_{i=1}^{k} \left[ 1 - \gamma^j \left( 1 - X_{n,i} \right) \right]}
\]
4. BOUNDING THE EXPECTED TIME UNTIL CONVERGENCE

\[
\sum_{t=1}^{\infty} \left\{ \prod_{j=0}^{t} \prod_{i=1}^{k} \left[ 1 - \gamma^j(1 - X_{n,i}) \right] \right\} \times \left[ 1 - \prod_{j=t+1}^{\infty} \prod_{i=1}^{k} \left[ 1 - \gamma^j(1 - X_{n,i}) \right] \right] \\
1 - \prod_{j=0}^{\infty} \prod_{i=1}^{k} \left[ 1 - \gamma^j(1 - X_{n,i}) \right]
\]

\[
\leq \sum_{t=1}^{\infty} \left[ 1 - \prod_{j=t+1}^{\infty} \prod_{i=1}^{k} \left[ 1 - \gamma^j(1 - X_{n,i}) \right] \right] \\
1 - \prod_{j=0}^{\infty} \prod_{i=1}^{k} \left[ 1 - \gamma^j(1 - X_{n,i}) \right] \\
\]

// first factor in each term ≤ 1

Now, we invoke the following

**Lemma 2.** \( 1 - \prod_{j=1}^{\infty} (1 - \epsilon_i) \leq \sum_{j=1}^{\infty} \epsilon_j \), when \( 0 \leq \epsilon_j \leq 1, j = 1, 2, \ldots \)

*(See NOTE 24)*

which we apply to the numerator in the last line to get

\[
\leq \sum_{t=1}^{\infty} \left[ \prod_{j=t+1}^{\infty} \sum_{i=1}^{k} \gamma^j(1 - X_{n,i}) \right] \\
1 - \prod_{j=0}^{\infty} \prod_{i=1}^{k} \left[ 1 - \gamma^j(1 - X_{n,i}) \right] \\
\]

\[
\leq \sum_{i=1}^{k} \left( 1 - X_{n,i} \right) \sum_{t=1}^{\infty} \sum_{j=t+1}^{\infty} \gamma^j \\
1 - \prod_{j=0}^{\infty} \prod_{i=1}^{k} \left[ 1 - \gamma^j(1 - X_{n,i}) \right] \\
\]

\[
= \sum_{i=1}^{k} \left( 1 - X_{n,i} \right) \sum_{t=1}^{\infty} \sum_{j=t+1}^{\infty} \gamma^j \\
1 - \prod_{j=0}^{\infty} \prod_{i=1}^{k} \left[ 1 - \gamma^j(1 - X_{n,i}) \right] \\
\]
4. BOUNDER THE EXPECTED TIME UNTIL CONVERGENCE

\[
\begin{align*}
&= \frac{\left[ \sum_{i=1}^{k} (1 - X_{n,i}) \right] \sum_{t=1}^{\infty} \frac{\gamma^t}{1 - \gamma}}{1 - \prod_{j=0}^{\infty} \prod_{i=1}^{k} \left[ 1 - \gamma^j (1 - X_{n,i}) \right]} \\
&= \frac{\left[ \sum_{i=1}^{k} (1 - X_{n,i}) \right] \frac{\gamma}{(1 - \gamma)^2}}{1 - \prod_{j=0}^{\infty} \prod_{i=1}^{k} \left[ 1 - \gamma^j (1 - X_{n,i}) \right]} \\
&\leq \frac{\left[ k(1 - \hat{X}_n) \right] \frac{\gamma}{(1 - \gamma)^2}}{1 - \prod_{j=0}^{\infty} \prod_{i=1}^{k} \left[ 1 - \gamma^j (1 - X_{n,i}) \right]} \quad \text{// where } \hat{X}_n = \min_{i=1,\ldots,k} X_{n,i} \quad (\text{10})
\end{align*}
\]

Now, regarding the denominator:

Observe that

\[
\prod_{j=0}^{\infty} \prod_{i=1}^{k} \left[ 1 - \gamma^j (1 - X_{n,i}) \right]
\]

\[
= \prod_{i=1}^{k} [1 - (1 - X_{n,i})] \cdot \prod_{i=1}^{k} [1 - \gamma (1 - X_{n,i})] \cdot \prod_{i=1}^{k} [1 - \gamma^2 (1 - X_{n,i})] \cdot \ldots
\]

\[
\leq \prod_{i=1}^{k} [1 - (1 - X_{n,i})] \\
= \prod_{i=1}^{k} X_{n,i}
\]
4. BOUNDING THE EXPECTED TIME UNTIL CONVERGENCE

And so

\[
\frac{1}{1 - \prod_{j=0}^{\infty} \prod_{i=1}^{k} \left[ 1 - \gamma^j (1 - X_{n,i}) \right]} \leq \frac{1}{1 - \prod_{i=1}^{k} X_{n,i}}
\]

Thus, continuing from (10), we have

\[
\leq \frac{\gamma^k}{(1-\gamma)^2} \frac{1 - \tilde{X}_n}{1 - \prod_{i=1}^{k} X_{n,i}}
\]

\[
\leq \frac{\gamma^k}{(1-\gamma)^2} \quad \text{// since the second fraction is } \leq 1
\]

And so we have the estimate sought:

\[
E \left( T \mid T < \infty \right) \leq \frac{k \gamma}{(1-\gamma)^2} .
\]

NOTE 24

Regarding Lemma 2: To prove the finite version is a straightforward induction; and then we take the limit.

Now, returning to \( S_L = \sum_{l=1}^{L} (1 + T_j) \), we estimate

\[
E S_L = \sum_{l=1}^{\infty} E \left[ S_L \mid L = l \right] P \left( L = l \right)
\]

\[
= \sum_{l=1}^{\infty} E S_l \cdot P \left( L = l \right)
\]
= \sum_{l=1}^{\infty} lP(L = l) + \\
\sum_{l=1}^{\infty} \left(E[T_1 \mid T_1 < \infty] + \cdots + E[T_l \mid T_l < \infty]\right) \cdot P(L = l)
\leq EL + \sum_{l=1}^{\infty} l \cdot \frac{k\gamma}{(1-\gamma)^2} \cdot P(L = l)
= EL + \frac{k\gamma}{(1-\gamma)^2} \sum_{l=1}^{\infty} lP(L = l)
= EL + \frac{k\gamma}{(1-\gamma)^2} \cdot EL
\leq \frac{2^k}{\eta_0^k \lambda_0} \left(1 + \frac{k\gamma}{(1-\gamma)^2}\right)

\frac{2^k}{\eta_0^k} \text{ is the dominant part of the estimate for } ES_L \text{ given in the last expression and so we can use it as a rough measure. } \eta_0^k \text{ is the value of the function } g(\gamma) = \prod_{j=0}^{\infty} \left[1 - \gamma^j(1 - \alpha_1)\right]^k, \text{ for } \gamma \in [0, 1]. \text{ Recall that this function is used in our estimate for the probability of taking branch 1 forever from the } n^{th} \text{ sentence on, after not taking branch 1 for sentence } n - 1 \text{ (i.e., } P(A_n \mid B_{n-1}^c[1]) \geq \frac{\eta_0^k \lambda_0}{2^k}).

As we’ve just seen, this probability is used in estimating the expected value for the number of steps (sentences) until ‘success’ (taking branch 1 forever). When we graph } g(\gamma) \text{ on } [0, 1] \text{ (see figure on next page), we find that it reaches its (only) maximum at about } \gamma = 0.4. \text{ If we let } k = 1, \text{ then this maximum is about } 0.11. \text{ So the maximum value for } \eta_0^k \text{ is about } 0.11^k.

[Note the change in the graph at } \gamma = 2/3; \text{ this is where } \alpha_1 \left(= \min[\frac{\gamma}{2}, 1 - \gamma]\right) \text{ changes from } \frac{\gamma}{2} \text{ to } 1 - \gamma ]\]
4.2. PART II. To put the estimate for $E_{S_L}$ in PART I into perspective we develop a different – artificial – method for “learning” (or, more accurately, ‘finding’) the target grammar. We suppose that there is a computer having all possible grammars in memory and it just tries them out, one after another, as sentences come in. So, for the first sentence encountered it tries to parse it with grammar 1 (the grammars are lined up in some random order). If it succeeds, it uses grammar 1 for the next sentence. If it fails, grammar 1 is discarded permanently and it tries grammar 2 for this sentence and those that follow, until it fails – whence it is discarded and we move on to grammar 3. The computer proceeds through the list of grammars in this fashion until it reaches a grammar that keeps succeeding and it just stays on that grammar ‘forever’, just as the child, at some point, stays with some grammar ‘forever’ – which, for the child, is what we call ‘learning the target grammar’. For the computer (or, rather, its algorithm), this ‘staying’ on the grammar is what we call ‘finding the target grammar’. The algorithm, written in pseudo-code, is given below, along with a derivation for the following estimate:

$$E[\# \text{ of steps until find target grammar}] \leq \frac{2^k-1}{\lambda_0}. \quad (11)$$

Algorithm `FindGrammar( )`

```plaintext
{  
g = 1; // g ∈ \{1, 2, \ldots, 2^k\}, i.e., an index into the list of 2^k grammars  
for n = 1 \ldots \infty {  
    if grammar[g] fails to parse sentence[n]  
    g = g + 1;  
}
}
```
Notice that this algorithm, if successful, does not halt. This is not usually a desired quality for an algorithm, but fine for a language learner who wants to keep using the language he’s learned.

The derivation of the estimate above for the computer method is as follows.

Assume we have a list of grammars \( \{G_1, G_2, \ldots, G_{2^k}\} \), randomly ordered.

Define the random variables:

\( H = \) the position of the target grammar in the list \( \{G_1, G_2, \ldots, G_{2^k}\} \)

\( Y_i = \) # of steps until grammar \( G_i \) fails. Recall that \( P(G_i \text{ fails}) = \lambda_i \), except for the (unknown) target grammar.

\( S_h = Y_1 + Y_2 + \cdots + Y_{h-1} \), when \( H = h \)

So, \( S_h \) is the number of steps until the target grammar is found, given that \( H = h \). And then \( S_H \) is just the number of steps until the target grammar is found. Thus, the estimate being verified can be written as

\[
ES_H \leq \frac{2^{k-1}}{\lambda_0}.
\]

We have \( EY_i = 1/\lambda_i \), since \( Y_i \) is a geometric random variable.

And

\[
ES_H = \sum_{h=1}^{2^k} P(H = h) \ E \left[ S_H \mid H = h \right]
\]
\[ BOUNDING \ THE \ EXPECTED \ TIME \ UNTIL \ CONVERGENCE \]
\[ \approx \frac{2^{k-1}}{\lambda_0} \]

And we have the estimate given in (11).

Sequential search is the optimal method for an unordered list, so our computer method gives us the optimal expectation for finding or learning the target grammar, which can be thought of as just a search through a list.
5. Submartingale Convergence Applies Only for $k \leq 2$

Yang’s algorithm is certainly intended to be more biologically realistic than the computer ‘program’ just presented. The change in mental state is predicted to be gradual rather than abrupt. We might also hope to show that the child makes progress as time goes on, perhaps in the sense that the expected probabilities for the correct settings of the parameters increase with exposure to more sentences.

For the case of one parameter, we’d like to show that $EX_n$ increases with $n$. In fact we can show that $E \left[ X_{n+1} \mid X_n \right] \geq X_n$, so that $X_n$ is a submartingale. This implies that $X_n$ must converge and it isn’t hard, then, to show (what we already know by Theorem 1) that the limit is 1.

In the case $k = 2$, we can show that $S_n = X_{n,1} + X_{n,2}$ is a submartingale, invoke the (sub)Martingale Convergence Theorem to show its convergence, and, again, use that to prove convergence to 1 of the $X_{n,i}$. For $k \geq 3$, $S_n$ is not a submartingale and so another approach, that taken in our proof, was used to show convergence of the $X_{n,i}$.

We sketch a demonstration that, for $k = 2$, $S_n$ is a submartingale. First, simplify the notation:

Replace $X_{n,1}, X_{n,2}$ with $X_n, Y_n$ and let $p, q$ be any values they could assume (recall that these values are themselves used as probabilities).

The 3 fail-to-parse probabilities, $\lambda_2, \lambda_3, \lambda_4,$

and $\gamma$, the adjustment-after-(failure-or-success) factor, stay the same.

Now, recall the submartingale criteria: $E|S_n| < \infty$ and

$$E \left[ S_{n+1} \mid X_n, Y_n \right] \geq S_n \text{ a.s., } n = 1, 2, \ldots$$
5. Submartingale convergence applies only for \( k \leq 2 \)

The first criterion is clearly met since \(|S_n| \leq 2, \forall n\). For the second criterion, we do a direct calculation and find that

\[
E \left[ S_{n+1} \middle| X_n = p, Y_n = q \right] = (p+q) [1 - 2\lambda_4(1 - \gamma)] + pq [2\lambda_4(1 - \gamma)] + 2\lambda_4(1 - \gamma)
\]

Checking whether

\[
(p + q) [1 - 2\lambda_4(1 - \gamma)] + pq [2\lambda_4(1 - \gamma)] + 2\lambda_4(1 - \gamma) \geq p + q,
\]

yields

\[
2\lambda_4(1 - \gamma) pq - 2\lambda_4(1 - \gamma)[p + q] + 2\lambda_4(1 - \gamma) \geq 0,
\]

giving

\[
pq - (p + q) + 1 \geq 0
\]

and

\[(p - 1)(q - 1) \geq 0,
\]

which is clearly true. So, \( S_n \) is a submartingale.

We demonstrate now that, for \( k = 3 \), \( S_n \) is not a submartingale, no matter what values are assigned to the \( \lambda_i, i = 2, \ldots, 2^3 \). (A similar demonstration could be made for \( k > 3 \).)

Again, replace \( X_{n,1}, X_{n,2}, X_{n,3} \) with \( X_n, Y_n, Z_n \) and let \( p, q, r \) be their values, while \( \lambda_2, \ldots, \lambda_8 \) and \( \gamma \) stay the same. Let \( S_n = X_n + Y_n + Z_n \).

The submartingale criteria are, again,

\[
E|S_n| < \infty \quad \text{and} \quad E \left[ S_{n+1} \middle| X_n, Y_n, Z_n \right] \geq S_n \quad \text{a.s.,} \quad n = 1, 2, \ldots
\]

A long direct calculation results in

\[
E \left[ S_{n+1} \middle| X_n, Y_n, Z_n \right] = S_n + (1 - \gamma) C_n
\]

where

\[
C_n = 3\lambda_8(1 - p)(1 - q)(1 - r)
\]
5. Submartingale convergence applies only for $k \leq 2$

$$+ \lambda_4 p(1-q)(1-r) + \lambda_6 q(1-p)(1-r) + \lambda_7 r(1-p)(1-q)$$

$$- \lambda_2 pq(1-r) - \lambda_3 pr(1-q) - \lambda_5 qr(1-p)$$

The question then is, given certain values of $\lambda_2, \ldots, \lambda_8$, is $C_n \geq 0$? Our answer is no: it doesn’t matter which values are chosen (or determined) for the $\lambda_2, \ldots, \lambda_8$, $C_n$ cannot be assured to be $\geq 0$. Note that in proving our assertion, we can let $p, q,$ and $r$ range over all their possible values.

Let $p = q = r$,

which reduces $C_n$ to:

$$C_n = (1-p)^3 \left[ 3\lambda_8 \right] + (1-p)^2 \left[ \lambda_4 + \lambda_6 + \lambda_7 \right] - p^2 (1-p) \left[ \lambda_2 + \lambda_3 + \lambda_5 \right]$$

If $C_n \geq 0$, then we have

$$(1-p)^3 \left[ 3\lambda_8 \right] + (1-p)^2 \left[ \lambda_4 + \lambda_6 + \lambda_7 \right] \geq p^2 (1-p) \left[ \lambda_2 + \lambda_3 + \lambda_5 \right]$$

or

$$\left( \frac{1-p}{p} \right)^2 \left[ 3\lambda_8 \right] + \left( \frac{1-p}{p} \right) \left[ \lambda_4 + \lambda_6 + \lambda_7 \right] \geq \lambda_2 + \lambda_3 + \lambda_5$$

Since we can make $p$ arbitrarily close to 1, and we have $\lambda_2 + \lambda_3 + \lambda_5 > 0$, the last inequality is impossible to maintain. Thus, $S_n$ is not a submartingale.

Of course this doesn’t mean that the probabilities of correct settings don’t stochastically increase with $n$, in some sense. A more exact understanding of how monotone (and quick) the convergence is, is a topic for further research.
Bibliography


