Department of Statistics and Operations Research,
University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark, www.stat.ku.dk

Niels Richard Hansen
September 17, 2002

Stochastic Grammars and Maximum Likelihood Inference

Abstract

Stochastic derivations of strings written in a finite alphabet are considered. For the class of models called stochastic context-free grammars we discuss whether such derivations terminates with probability one. It is finally shown that for stochastic context-free grammars maximum-likelihood estimation with either full or partial observations will automatically yield terminating derivations.

1 Introduction

In computer science and the theory of languages there has been a lot of interest in studying formal production rules capable of deriving or producing a specific language of strings. These formal rules are called the *grammar* of the language.

The formal languages are in a sense not very flexible – either a string is in a language or not – and some attempts have been made to assign probabilities to the production rules, so that different strings can have different probabilities of occurring. These stochastic models are called *stochastic grammars*. A more recent application is the area of biological sequence analysis like the description of RNA-molecules (Eddy & Durbin, 1994), (Yasubumi et. al., 1994), (Durbin et. al., 1998).

For ordinary grammars we usually consider the set of whatever we can derive in a finite number of steps using the grammar. For stochastic grammars, however, we need to consider the derivations that occur with probability one, and we could face the problem that with positive probability some derivations continue forever never yielding a final product. To have a meaningful model the derivations should terminate with probability one. This problem was mentioned by Sankoff (1971). He remarks that a related branching process should be subcritical, but this seems to have been overlooked in the more recent literature.

The main purpose of this paper is to reconsider the problem of termination of stochastic derivations from a stochastic grammar and especially relations to statistical inference. For the class of grammars considered in this paper called *stochastic context-free grammars* a rather complete treatment is given. After a construction of stochastic context-free grammars in section 2, we show how to characterize the

terminating stochastic context-free grammars. Then we show in section 3 that the maximum likelihood estimator under a complete parameterization of a finite set of production rules always gives a terminating stochastic context-free grammar. In section 4 we give a simple example of a stochastic grammar, and the reader is invited to jump to this example after reading section 2. Finally in section 5 we discuss some results related to the length of the produced strings.

2 Basic Definitions and Results

2.1 Alphabets, Strings and Languages

For any set W (finite or infinite) we will use W^* to denote the set of finite sequences or strings from W. The set W should be thought of as an alphabet and the set of strings W^* are then the possible words we can write in this alphabet. A subset of W^* is called a language.

We want to construct a language in a systematic way using formal production rules – a so called grammar. In this paper we want to assign probabilities to the words in the language by using a stochastic grammar, where the production rules are chosen randomly according to some probability distribution.

Usually the elements in W^* will be called strings and the convention that a substring of a string is contiguous will be used. At some point the set $W^{**} := (W^*)^*$ of finite sequences of strings, strings of strings, strings from the alphabet W^* or whatever one prefers will be needed. In this paper, elements in W^{**} will be called finite sequences of strings to distinguish these from the more basic strings being finite sequences from the alphabet W.

Whenever necessary, the letters in a string $\alpha \in W^*$ will be labelled using $\mathbb{N}$, so that $\alpha = \alpha_1 \dots \alpha_m$ where $\alpha_i \in W$ for $i = 1, \dots, m$. The length of string α is denoted by $|\alpha|$ and equals the number of letters in α. In general, the Greek letters α, β, γ will be used to denote strings or single letters from the alphabet W. Furthermore, elements in W^{**} are written as $(\alpha_1, \dots, \alpha_k)$ with $\alpha_i \in W^*$ for $i = 1, \dots, k$.

2.2 Grammars

Consider a finite set E with n elements and when necessary label the elements in E by $\{x_1, \dots, x_n\}$. The elements in E will be called *non-terminals*. We will also consider another finite set A where the elements in A will be called terminals. We will use $E \cup A$ to denote the disjoint union of E and A. Elements in E will be denoted x, y, z and elements in A will be denoted a, b, c.

One should think about A as the alphabet of interest and E as some additional letters, which can be used as *placeholders* in a formal derivation of strings in A^*. Thus, we are really interested in subsets of A^* – languages written in the alphabet

A – but for the construction of such languages we need the additional non-terminals in E.

In the rest of the paper $\mathcal{S} = (E \cup A)^*$ will denote the strings written in the alphabet $E \cup A$.

Definition 2.1 *A subset $R \subseteq E \times \mathcal{S}$ fulfilling that for all $x \in E$ there exits a $\gamma \in \mathcal{S}$ such that $(x, \gamma) \in R$ is called a set of context-free production rules.*

Definition 2.2 *A context-free grammar is a triple $\mathcal{G} = (E, A, R)$ where E is a set of non-terminals, A is a set of terminals and R is a set of production rules.*

Given a grammar $\mathcal{G}$ we can recursively derive a string from A^* as follows: We start with one non-terminal $x_0 \in E$ interpreted as a string in $\mathcal{S}$ containing one letter[1]. Then for a given string $\alpha \in \mathcal{S}$, if $\alpha \in A^*$ stop the derivation, otherwise substitute each occurrence of a non-terminal x in α with a string γ for which $(x, \gamma) \in R$. Such recursive derivations might in principle continue forever, but we are only interested in those that stop eventually, i.e. those for which we end up with a string in A^*. A derivation that stops eventually is also said to terminate. Denote by $\mathcal{L} = \mathcal{L}(\mathcal{G})$ the language derived using $\mathcal{G}$, i.e. the strings produced using $\mathcal{G}$ by a derivation that terminates.

The grammar is called context-free because a string γ, which can be produced from a given non-terminal x do not depend on the context in which x is found, only on the fact that (x, γ) should belong to R.

2.3 Stochastic Grammars

To define a stochastic grammar we need to assign probabilities to the production rules. Thus let P be a Markov kernel from E to $\mathcal{S}$ such that

$$\sum_{\gamma:(x,\gamma)\in R} P(x, \gamma) = 1 \tag{1}$$

for all $x \in E$.

Definition 2.3 *A stochastic context-free grammar is a quadruple (E, A, R, P) where E is a set of non-terminals, A is a set of terminals, R is a set of production rules and P is a Markov kernel from E to $\mathcal{S}$ satisfying* (1).

We then want to interpret the stochastic derivations as drawing an initial non-terminal using some distribution ν on E and then given a string $\alpha \in \mathcal{S}$, independently for each non-terminal x in α (if any) substitute x by a string drawn from

[1] sometimes the initial non-terminal x_0 is assumed to be fixed and given together with the grammar (E, A, R)

$P(x, \cdot)$. We also want to define a suitable framework for making statistical inference about unknown parameters given some observed or partially observed stochastic derivations. We will understand stochastic derivations as a Markov chain, not on $\mathcal{S}$, but on $\mathcal{S}^*$ in fact. The reason for this will be discussed below.

Extend P to be defined on $E \cup A$ by letting $P(a, a) = 1$ for all $a \in A$. Define the *concatenation* map $c : \mathcal{S}^* \to \mathcal{S}$ by

$$c(\alpha_1, \ldots, \alpha_m) = \alpha_1 \ldots \alpha_m$$

and define a Markov kernel Q_0 from $\mathcal{S}$ to $\mathcal{S}^*$ by

$$Q_0(\alpha, (\gamma_1, \ldots, \gamma_m)) = \prod_{i=1}^{m} P(\alpha_i, \gamma_i)$$

if $|\alpha| = m$. Finally define a Markov kernel Q from $\mathcal{S}^*$ to $\mathcal{S}^*$ by

$$\begin{aligned} Q((\alpha_1, \ldots, \alpha_m), (\gamma_1, \ldots, \gamma_k)) &= Q_0(c(\alpha_1, \ldots, \alpha_m), (\gamma_1, \ldots, \gamma_k)) \\ &= Q_0(\alpha_1 \ldots \alpha_m, (\gamma_1, \ldots, \gamma_k)) \end{aligned}$$

if $|\alpha_1| + \cdots + |\alpha_m| = k$.

Definition 2.4 *Let (E, A, R, P) be a given stochastic context-free grammar, then a stochastic derivation with initial distribution ν on E is a Markov chain on $\mathcal{S}^*$ with transition probabilities given by the Markov kernel Q defined above and initial distribution given by ν – interpreted as a distribution on $\mathcal{S}^*$.*

Each step in the Markov chain can be seen as composed of a stochastic production of strings and a deterministic concatenation of these strings, thus the chain moves as the following diagram indicates:

$$\begin{array}{ccccccccc} & & \mathcal{S}^* & & \mathcal{S}^* & & \mathcal{S}^* & & \\ & \nearrow & \downarrow & \nearrow & \downarrow & \nearrow & \downarrow & \nearrow & \\ \mathcal{S} & & \mathcal{S} & & \mathcal{S} & & \mathcal{S} & & \end{array}$$

If $(X_n)_{n \in \mathbb{N}_0}$ denotes the Markov chain, we see by the definition of Q that the conditional distribution of X_{n+1} given X_n only depends upon $c(X_n)$, hence the process $(c(X_n))_{n \in \mathbb{N}}$ is also a Markov chain on $\mathcal{S}$. One might think that $(c(X_n))_{n \in \mathbb{N}}$ is a more natural candidate for a stochastic derivation from the stochastic context-free grammar (E, A, R, P) than $(X_n)_{n \in \mathbb{N}}$ is. However, $(c(X_n))_{n \in \mathbb{N}}$ is not in general sufficient for parameter estimation as $(X_n)_{n \in \mathbb{N}}$ is, since we might not be able to deduce the production rules used from $(c(X_n))_{n \in \mathbb{N}}$ alone.

Note that all elements $a \in A$ occurring in a string $\alpha \in \mathcal{S}$ just reproduce themselves at each step in the Markov chain and that each sequence in A^{**} is therefore absorbing. Thus absorbtion in a sequence from A^{**} corresponds to termination of the stochastic derivation.

2.4 Termination of Stochastic Derivations

Definition 2.5 *Put* $\tau = \inf\{n \mid X_n \in A^{**}\}$. *A stochastic context-free grammar is said to be terminating with probability one or just terminating if*

$$\mathbb{P}_\nu(\tau < \infty) = 1$$

for all initial distributions ν.

It turns out that the theory of multitype branching processes can be used to answer whether a stochastic context-free grammar is terminating, since the non-terminals occurring in the Markov chain give rise to a multitype branching process.

With $E = \{x_1, \ldots, x_n\}$ the set of non-terminals let Λ_{ij} denote the expected number[2] of non-terminals x_j produced from a non-terminal x_i for $i, j = 1, \ldots, n$ and let $\Lambda = (\Lambda_{ij})_{i,j}$ be the matrix of expectations. The spectral radius of Λ (which we will denote by $\rho(\Lambda)$) then determines whether or not the stochastic grammar is terminating – at least to a very large extend. The precise statement follows below, but first we will exclude a special class of grammars, which will not be terminating.

Definition 2.6 *A stochastic context-free grammar is said to be singular if there exists a set* $E_0 \subseteq E$ *such that for all* $x \in E_0$ *and all* $\gamma \in \mathcal{S}$ *with* $P(x, \gamma) > 0$ *the string* γ *contains exactly one state from* E_0. *The grammar is called nonsingular, if it is not singular.*

Theorem 2.7 *A nonsingular stochastic context-free grammar is terminating if and only if* $\rho(\Lambda) \leq 1$.

Proof: Define the map $\kappa : \mathcal{S} \to \mathbb{N}_0^n$ by $\kappa(\alpha) = (\kappa_1, \ldots, \kappa_n)$ if α contains κ_i non-terminals x_i for $i = 1, \ldots, n$. We lift κ to be defined on $\mathcal{S}^*$ too by composition with c and denote the lifted map by κ also. Then if $(X_n)_{n\in\mathbb{N}_0}$ is a stochastic derivation from a stochastic context-free grammar, $(\kappa(X_n)_n)_{n\in\mathbb{N}_0}$ is a multitype branching process with n types. With $H_{ij}, i, j = 1, \ldots, n$ denoting the offspring distributions for this branching process, i.e. H_{ij} is the distribution of the number of non-terminals x_i produced from a non-terminal x_j, we see that Λ is the matrix of expectations given by the distributions H_{ij}. Also if the stochastic context-free grammar is non-singular no subset of E will reproduce exactly one non-terminal from that subset with probability one, and from theorem 10.1, chapter II in (Harris, 1963) we find that this branching process will become extinct with probability one if and only if $\rho(\Lambda) \leq 1$ – see also chapter 2 in (Mode, 1971) and chapter 4 in (Jagers, 1975). Since termination of the stochastic context-free grammar is the same as extinction of the branching process $(\kappa(X_n))_{n\in\mathbb{N}}$ the result follows. □

[2] this could in principle be ∞ but later on the set G will be finite, in which case the expectations will be finite

Clearly, a singular context-free grammar will not be terminating, so we will not be interested in such grammars. The theorem above then tells us exactly which grammars are terminating.

2.5 Statistical Implications

If $\Theta \ni \theta \mapsto P_\theta$ denotes a parameterization of the Markov kernel P from E to $\mathcal{S}$ in the stochastic context-free grammar, then Θ might in general contain parameters, which don't make the stochastic context-free grammar terminating, so we should restrict our parameter space to those that make it terminating. Considering only those parameters θ with $\rho(\theta) \leq 1$ – with $\rho(\theta)$ the spectral radius of the corresponding mean value matrix $\Lambda(\theta)$ – we get a restriction of the parameter space, which besides singularity problems gives terminating stochastic grammers.

One could argue that this restriction problem is artificial, since we could just condition on termination of the derivations, i.e. condition on absorption in A^{**}. Then all parameters gives rise to a model, which produces finite strings. There are, however, two reasons why this is not tractable. First of all a purely practical reason is that to condition on termination we should be able to compute the parameter dependent probability of termination, which seems impossible in general. Secondly, even if we were able to compute this probability, we would then get a parameter estimate that might be difficult to interpret, since in the conditional model $P_\theta(x, \cdot)$ is no longer the distribution of products from the non-terminal x.

3 Maximum Likelihood Inference

In this section we show that if we consider a finite set G of context-free production rules and the full parameterization Θ, i.e. all Markov kernels P giving a stochastic context-free grammar, then if we observe a Markov chain absorbed in A^{**}, the maximum likelihood estimator $\hat{\theta}$ satisfies $\rho(\hat{\theta}) \leq 1$, i.e. the corresponding stochastic context-free grammar is terminating – provided nonsingularity. We will in this section assume that the initial distribution ν is fixed and given – only the production probabilities will be estimated.

First we consider the case with complete observations and then the case where we observe the Markov chain partially.

3.1 With Complete Observations

A *finite* set G of production rules is given, for which we label the rules $g_{ij} = (x_i, \gamma_{ij})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, k_i$. Thus, for all $x_i \in E$ we have k_i different allowed production rules. Denote by n_{ijl} the number of non-terminals x_l produced from

non-terminal x_i using rule g_{ij}, and for $i = 1, \ldots, n$ let

$$N_i = \begin{pmatrix} n_{i11} & \ldots & n_{i1n} \\ \vdots & & \vdots \\ n_{ik_i1} & \ldots & n_{ik_in} \end{pmatrix} \tag{2}$$

be the $k_i \times n$ matrix of the number of produced non-terminals from non-terminal x_i. Furthermore, let

$$N = \begin{pmatrix} N_1 \\ \vdots \\ N_n \end{pmatrix} \tag{3}$$

be the matrix of all the matrices N_i.

Probabilities are assigned to each of the production rules. Let p_{ij} be the probability of choosing the production rule g_{ij} given non-terminal x_i. Then we have that $\sum_j p_{ij} = 1$ for all $i = 1, \ldots, n$ and $\Theta = \{p = (p_{ij}) | \sum_j p_{ij} = 1;\ 0 \leq p_{ij} \leq 1\}$. Let for $i = 1, \ldots, n$

$$p_i = (p_{i1}, \ldots, p_{ik_i}) \tag{4}$$

and let

$$P = \operatorname{diag}(p_1, \ldots, p_n) \tag{5}$$

be the block-diagonal matrix, with the blocks being the vectors p_i. Then $\Lambda_{il} = \sum_j p_{ij} n_{ijl}$ and hence

$$\Lambda = PN = \begin{pmatrix} p_1 N_1 \\ \vdots \\ p_n N_n \end{pmatrix} \tag{6}$$

where Λ is the expectation matrix.

Assume that we have observed $\mathbb{X} = (X_n)_{n \in \mathbb{N}_0}$ being a Markov chain on $\mathcal{S}^*$ absorbed in A^{**} from a certain point. This observation is transformed into a sufficient statistic of counts. Thus, let $c_{ij} = c_{ij}(\mathbb{X})$ be the number of times production rule g_{ij} has been used in $\mathbb{X}$. Then the log-likelihood is

$$l(p) = \sum_{i=1}^{n} \sum_{j=1}^{k_i} c_{ij} \log(p_{ij}), \tag{7}$$

which of course gives the usual maximum likelihood estimators;

$$\hat{p}_{ij} \quad = \quad {s_i}^{-1} c_{ij} \tag{8}$$

where $s_i = \sum_j c_{ij}$ for $i = 1, \ldots, n$.

Putting

$$c_i = (c_{i1}, \ldots, c_{ik_i}) \tag{9}$$

together with the block diagonal matrix

$$C = \text{diag}(c_1, \ldots, c_n) \tag{10}$$

and the diagonal matrix

$$S = \text{diag}(s_1, \ldots, s_n) \tag{11}$$

we get that

$$\widehat{P} = S^{-1}C, \tag{12}$$

and therefore we have the estimated expectation matrix

$$\widehat{\Lambda} = S^{-1}CN. \tag{13}$$

Let

$$c = (c_1, \ldots, c_n) \quad \text{and} \quad s = (s_1, \ldots, s_n) \tag{14}$$

be the collection of the $c_1, \ldots, c_n$ vectors and the $s_1, \ldots, s_n$ numbers in two vectors (as opposed to the matrices C and S). Because the Markov chain $\mathbb{X}$ is absorbed in A^{**} each state x_i appearing somewhere in $\mathbb{X}$ has to have been used for some production, so the number of times x_i has been produced (for $x_i = X_0$ this number includes the initial production of x_i) equals the number of times x_i has been used for a production. Thus for $l = 1, \ldots, n$

$$\delta_l + \sum_{i=1}^{n} \sum_{j=1}^{k_i} c_{ij} n_{ijl} = \sum_{j=1}^{k_l} c_{lj} = s_l, \tag{15}$$

where $\delta_l = \delta_l(\mathbb{X})$ equals 1 if $X_0 = x_l$ and 0 otherwise, i.e. δ_l counts which non-terminal initiates the derivation. In matrix notation this amounts to the equation

$$\delta + cN = s. \tag{16}$$

with δ the vector of δ_l's.

The goal is to show that $\widehat{\Lambda}$ has spectral radius less than or equal to 1. To do so the Perron-Frobenius theory is applied. Matrices are assumed to be square in the following whenever necessary. A matrix is called nonnegative if all entries are nonnegative, and the spectral radius of A will be denoted by $\rho(A)$ (which is the largest modulus of the eigenvalues of A). If A is nonnegative it follows from the Perron-Frobenius theory that $\rho(A)$ is it-self an eigenvalue with (at least one) nonnegative eigenvector for both A and A^T – see theorem 1.1 in chapter 2 of (Berman & Plemmons, 1979). If A is irreducible and aperiodic $\rho(A)$ is furthermore the unique eigenvalue of this modulus with both geometric and algebraic multiplicity one. For a nonnegative and irreducible matrix A the numerically largest eigenvalue $\rho(A)$ is often called the Perron-Frobenius eigenvalue. We will need results for nonnegative matrices, which are not necessarily irreducible.

Lemma 3.1 *If A is a nonnegative matrix then*

$$\rho(A) \leq \max_j \sum_i A_{ij}. \tag{17}$$

If A is irreducible the inequality in (17) is strict unless all the column sums are equal.

Proof: This is an inequality from Perron-Frobenius theory. For irreducible matrices – see for instance corollary 1 and theorem 1.5 in chapter 1 in (Seneta, 1981). It holds, however, without the irreducibility assumption. To see this we copy a small part of the proof for the Perron-Frobenius theorem as given in (Seneta, 1981).

For any nonnegative nonzero (column)vector $x = (x_1, \ldots, x_n)^T$ (where $x_i \geq 0$ and $x \neq 0$) we define

$$r(x) = \min_i \frac{\sum_j A_{ij} x_j}{x_i},$$

which gives that

$$r(x) \sum_i x_i \leq \sum_j x_j \sum_i A_{ij}.$$

And therefore

$$r(x) \leq \frac{\sum_j x_j \sum_i A_{ij}}{\sum_i x_i} \leq \max_j \sum_i A_{ij}$$

where the last inequality is strict unless the column sums are equal for those i with $x_i > 0$. As mentioned above, from chapter 2 theorem 1.1 in (Berman & Plemmons, 1979), the matrix A has a nonnegative *right* eigenvector – call this vector x_0 – with eigenvalue $\rho(A)$, and it follows that

$$\rho(A) = r(x_0) \leq \max_j \sum_i A_{ij}.$$

If A is irreducible all entries in x_0 are strictly positive (Berman & Plemmons, 1979, chapter 2 theorem 1.3) and the inequality is strict unless all column sums are equal. □

The following lemma will also be used. The proof is elementary – see for instance exercise I.3.7 in (Bhatia, 1997).

Lemma 3.2 *For matrices A and B it holds that $\rho(AB) = \rho(BA)$.*

Theorem 3.3 *For any vector of counts fulfilling equation (16) and with $\widehat{\Lambda}$ given by (13) it holds that $\rho(\widehat{\Lambda}) \leq 1$.*

Proof: Using lemma 3.2 we find that

$$\rho(\widehat{\Lambda}) = \rho(S^{-1}CN) = \rho(CNS^{-1}).$$

Denote by $\sigma : M(n_1, n_2) \to \mathbb{R}^{n_2}$ the mapping which takes an $n_1 \times n_2$ matrix into the (row)vector of column sums, then it is easy to see that $\sigma(AB) = \sigma(A)B$ for all matrices $A \in M(n_1, n_2)$ and $B \in M(n_2, n_3)$. Furthermore, by the definition of C and c it follows that $\sigma(C) = c$. Therefore

$$\begin{aligned} \sigma(CNS^{-1}) &= \sigma(C)NS^{-1} \\ &= cNS^{-1} \\ &= (s - \delta)S^{-1}, \end{aligned}$$

where the last equality follows from (16). If we denote by $\mathbf{1}$ the vector of all ones, it follows that

$$\sigma(CNS^{-1}) = \mathbf{1} - \delta S^{-1}$$

and $\rho(\widehat{\Lambda}) = \rho(S^{-1}NC) \leq 1$ from lemma 3.1. □

We can also observe, that if $\widehat{\Lambda}$ is irreducible, then $\rho(\widehat{\Lambda}) < 1$ also from lemma 3.1, since all column sums are 1 except the column l with $\delta_l = 1$, which is $1 - s_l^{-1} < 1$.

3.2 With Partial Observations

Suppose that we do not observe the Markov chain $\mathbb{X}$ completely but instead some transformation $T(\mathbb{X})$. Then if the sufficient counts $c_{ij}(\mathbb{X})$ are functions of $T(\mathbb{X})$ – in which case T is a sufficient transformation – we can of course still estimate parameters using maximum-likelihood as above, but if $T(\mathbb{X})$ is a truly non-sufficient partial observation of the Markov chain we need for instance the EM-algorithm to find the maximum-likelihood estimate. We will in the following assume that the transformation T takes values in some discrete space F.

Definition 3.4 *A value $t \in F$ is called absorbing if the set $T^{-1}(t)$ consists entirely of sequences from $\mathcal{S}^*$ absorbed in A^{**} from a certain point. A value $t \in F$ is called possible if $T^{-1}(t)$ is non-empty.*

Remark 3.5 We will be mostly interested in transformations defined a priori only for those Markov chains that get absorbed in A^{**}, but if we join an element Δ to F and let $T(\mathbb{X}) = \Delta$ if $\mathbb{X}$ is not absorbed in A^{**}, we get a transformation T defined on the whole set $(\mathcal{S}^*)^{\mathbb{N}_0}$ with values in $F \cup \{\Delta\}$. In this case all elements in F are absorbing and the element Δ is not.

Example 3.6 The single most important transformation is

$$T(\mathbb{X}) = c(X_\tau)$$

where $\tau = \inf\{n \mid X_n \in A^{**}\}$ as previously defined is the point of termination and c is the concatenation map. The transformation T takes its values in A^* and is only defined whenever $\tau < \infty$. The transformation is enlarged as described above to take some value Δ if $\tau = \infty$, and we observe that all strings in A^* are absorbing. However, some strings might not be possible.

Let

$$c_{ij}(p,t) := \mathbb{E}_p(c_{ij}(\mathbb{X})|t)$$
$$\delta_l(p,t) := \mathbb{E}_p(\delta_l(\mathbb{X})|t)$$

be the conditional expectation of the counts c_{ij} and δ_l given the partial observation $t \in F$ under the measure given by the parameter p. Then the EM-algorithm updates the parameters as follows; given p_n

1. Compute $c_{ij}(p_n, t)$
2. Maximize the log-likelihood

$$l(p) = \sum_{i=1}^{n} \sum_{j=1}^{k_i} c_{ij}(p_n, t) \log(p_{ij}),$$

   i.e. put

$$p_{ij,n+1} \quad = \quad s_i(p_n,t)^{-1} c_{ij}(p_n,t)$$

   where $s_i(p_n,t) = \sum_j c_{ij}(p_n,t)$ for $i = 1, \ldots, n$.

It is well known that the sequence $(p_n)_{n\in\mathbb{N}}$ of parameters gives an increasing likelihood (Dempster et. al., 1977), (Lari & Young, 1990), and we see from above that the maximization part of the EM-algorithm can be done explicitly in each iteration of the algorithm. The only problem is the computation of $c_{ij}(p_n, t)$ in each iteration. Depending on the transformation T and the stochastic context-free grammar, a number of different algorithms exists. For details the reader is referred to the literature (Durbin et. al., 1998), (Lari & Young, 1990), (Jelinek, 1985) and (Baker, 1979). It should, however, be mentioned that these algorithms have a computationally complexity cubic in the length of the observed string, so that each iteration of the EM-algorithm will become quite computationally demanding.

What is of interest to us is that even if we observe only $t = T(\mathbb{X})$ instead of $\mathbb{X}$ and if t is absorbing, then using the EM-algorithm to estimate the parameter p we will get a stochastic context-free grammar, which is terminating.

Theorem 3.7 *Suppose that $t \in F$ is possible and absorbing. With*

$$\Theta = \{p = (p_{ij}) \mid \sum_j p_{ij} = 1;\ 0 \leq p_{ij} \leq 1\}$$

the parameter space for the productions, then for any starting value $p_0 \in \Theta$ the EM-algorithm will give a sequence $(p_n)_{n\in\mathbb{N}}$ satisfying $\rho(p_n) \leq 1$ for $n \geq 1$ and if

$$p_n \to p$$

for $n \to \infty$ then $\rho(p) \leq 1$.

Proof: Since t is absorbing, we have that the equation (16)

$$\delta(\mathbb{X}) + c(\mathbb{X})N = s(\mathbb{X})$$

holds for all $\mathbb{X}$ with $T(\mathbb{X}) = t$. Since t is possible, such $\mathbb{X}$'s exist too. The linearity of conditional expectations then gives that for all p

$$\delta(p,t) + c(p,t)N = s(p,t) \tag{18}$$

where $\delta(p,t)$, $c(p,t)$ and $s(p,t)$ are the collections of $\delta_l(p,t)$, $c_{ij}(p,t)$ and $s_i(p,t)$ into vectors exactly as in the previous section. Letting $\Lambda(p)$ denote the expectation matrix for the parameter p and $\rho(p)$ the corresponding spectral radius, it follows from theorem 3.3 that

$$\rho(p_n) \leq 1$$

for all $n \geq 1$ since (18) holds. Continuity of Λ as a function of p as well as continuity of the spectral radius map shows that for an eventual limit p also $\rho(p) \leq 1$ holds. □

Remark 3.8 This gives that the maximum likelihood estimator essentially always gives a terminating stochastic context-free grammar, and from a practical point of view – using the EM-algorithm for numerical optimization – any estimate we produce, also in a finite number of iterations, will give a terminating stochastic context-free grammar. This would not necessarily be true, if we used another numerical maximization algorithm.

4 Example

Example 4.1 We will consider a stochastic context-free grammar with two states and two symbols. Thus let $E = \{x_1, x_2\}$ and $A = \{a, b\}$ and let the non-trivial production rules be

$$\begin{aligned} g_{11} &= (x_1, x_1x_2) \\ g_{12} &= (x_1, b) \\ g_{21} &= (x_2, x_1) \\ g_{22} &= (x_2, a) \end{aligned}$$

and let the parameterization of the production probabilities be given as

$$\begin{aligned} P(x_1, x_1x_2) &= 1 - P(x_1, b) = p_1 \\ P(x_2, x_1) &= 1 - P(x_2, a) = p_2 \end{aligned}$$

with $0 < p_1, p_2 < 1$. Let the initial distribution be fixed $\nu = \delta_{x_1}$.

Consider the transformation T as in example 3.6 given by

$$T(\mathbb{X}) = c(X_\tau) = \underbrace{b \ldots b}_{n_1}\underbrace{a \ldots a}_{m_1} \ldots \underbrace{b \ldots b}_{n_k}\underbrace{a \ldots a}_{m_k}$$

where $n_1 > 0$. Observe that any sequence starting with b can occur from these production rules. Observe also that the counts c_{ij} will always satisfy the following two equations

$$\begin{aligned} c_{22} &= c_{11} - c_{21} \\ c_{12} &= 1 + c_{21} \end{aligned}$$

and that

$$\begin{aligned} n &:= \sum_i^k n_i = c_{12} \\ m &:= \sum_i^k m_i = c_{22}. \end{aligned}$$

Hence $c_{21} = n - 1$ and $c_{11} = m + n - 1$. In this example we see that T is in fact sufficient and that the maximum-likelihood estimators for p_1 and p_2 are given as

$$\begin{aligned} \hat{p}_1 &= \frac{m + n - 1}{m + 2n - 1} \\ \hat{p}_2 &= \frac{n - 1}{m + n - 1}. \end{aligned}$$

The expectation matrix is found to be

$$\Lambda = \begin{pmatrix} p_1 & p_1 \\ p_2 & 0 \end{pmatrix}$$

for which we find that $\rho(\Lambda) \leq 1$ if and only if

$$p_2 \leq \frac{1}{p_1} - 1.$$

Finally observe that

$$\frac{1}{\hat{p}_1} - 1 = \frac{n}{m + n - 1} > \frac{n - 1}{m + n - 1} = \hat{p}_2$$

so that the maximum likelihood estimator gives, as proved in section 3, a terminating stochastic grammar.

5 The Length of the Produced Sequences

This section is mostly of a speculative nature!

It could be of some interest to study the distribution of the length of the sequences produced by a terminating stochastic context-free grammar. This problem is closely related to the study of the total number of terminals/non-terminals of each type produced by the branching process. One must especially expect a fundamental difference between the distributions corresponding to a *critical* ($\rho = 1$) and a *sub-critical* ($\rho < 1$) branching process.

Letting H_i denote the offspring distribution of the branching process from non-terminal x_i, i.e.

$$H_i(m_1, \ldots, m_n) = \sum_{j:n_{ij1}=m_1,\ldots,n_{ijn}=m_n} p_{ij},$$

and letting h_i denote the corresponding generating function, i.e.

$$h_i(z_1, \ldots, z_n) = \sum_{m_1,\ldots,m_n} z_1^{m_1} \cdots z_n^{m_n} H_i(m_1, \ldots, m_n)$$

one can easily prove that the distributions R_i of the total number of non-terminals produced given that we start with one non-terminal x_i have generating functions r_i fulfilling the equations

$$r_i(z) = z_i h_i(r_1(z), \ldots, r_n(z))$$

with $z = (z_1, \ldots, z_n)$ and $i = 1, \ldots, n$.

(Otter, 1949) used this equation for $n = 1$ to show that for $m \to \infty$

$$R_1(m) = c\alpha^{-m} m^{-\frac{3}{2}} + O(\alpha^{-m} m^{-\frac{5}{2}}) \quad m \equiv 1 (\text{mod } q) \tag{19}$$

for some constants c, $\alpha \geq 1$ and q the period of H_1. One should especially know that the critical case corresponds to $\alpha = 1$, so the tails of R_1 are very slowly decaying in the critical case and exponentially decaying in the sub-critical case. For multi-type branching processes there doesn't seem to exist a result like that. However, in the paper (Good, 1958) it is proved that $R_i(m_1, \ldots, m_n)$ is the coefficient of $z_1^{m_1} \cdots z_i^{m_i - 1} \cdots z_n^{m_n}$ in

$$r_1^{m_1} \cdots r_n^{m_n} |I - \{\frac{z_i}{r_i} \partial_j r_i\}_{i,j}| \tag{20}$$

with $|\cdot|$ denoting determinant.

Returning to the example of the preceding section, the generating functions become

$$\begin{aligned} r_1(z_1, z_2) &= q_1 + p_1 z_1 z_2 \\ r_2(z_1, z_2) &= q_2 + p_2 z_1 \end{aligned}$$

with $q_i = 1 - p_i$. Using (20), this gives

$$R_1(m_1, m_2) = \binom{m_1}{m_2} \binom{m_2}{2m_2 - m_1 + 1} \left(1 - \frac{m_1 - 1}{m_1}\right) p_1^{m_2} p_2^{m_1 - m_2 - 1} q_1^{m_1 - m_2} q_2^{2m_2 - m_1 + 1}$$

Figure 1: The log-marginals plotted against $\log(m)$. In both cases we see a tail decay thats asymptotically linear on this log-log-plot, thus the decay is like a power function. The straight line has slope $-3/2$

for $1+m_2 \leq m_1 \leq 2m_2+1$. It is not easy to compute the marginals from this expression analytically. But if we consider an example taking $p_1 = 3/4$ and $p_2 = 1/3$ (which is a critical example), we can then check numerically if the same kind of tail behavior is present as in the one-type case. Thus on figure 1 we see plot $\log(\sum_{m_2} R_1(m_1, m_2))$ against $\log(m_1)$ and $\log(\sum_{m_1} R_1(m_1, m_2))$ against $\log(m_2)$ compared to a straight line with slope $-3/2$. On both pictures we see that asymptotically, the decay of the marginals of R_1 is like $n^{-3/2}$, which confirms that in the multi-type setting we should have the same tail behavior as for the one-type case.

References

Baker, James K. (1979). *Trainable grammars for speech recognition.* Speech Communication Papers for the 97th Meeting of the Acoustical Society of America (eds; Klatt, D.H.. and Wolf, J.J.). 547-550.

Baldi, Pierre and Brunak, Søren. (1998). *Bioinformatics. The machine learning approach.* MIT Press.

Berman, Abraham and Plemmons, Robert J. (1979). *Nonnegative Matrices in the Mathematical Sciences.* Academic Press.

Bhatia, Rajendra. (1997). *Matrix Analysis.* Springer Verlag.

Dempster, A. P., Laird, N.M., Rubin, D.B. (1977). *Maximum Likelihood from Incomplete Data via the EM algorithm.* J.Roy.Statist.Soc. vol. 39, 1-38.

Durbin, R., Eddy, S., Krogh, A. and Mitchinson, G. (1998). *Biological Sequence Analysis. Probabilistic models of proteins and nucleic acids.* Cambridge University Press.

Eddy, Sean, R. Durbin, Richard (1994). *RNA sequence analysis using covariance models.* Nucleic Acids Research. Vol. 22. No. 11. 2079-2088.

Theodore E. Harris (1963). *The Theory of Branching Processes.* Springer-Verlag.

Jagers, Peter (1975). *Branching Processes with Biological Applications.* John Wiley and Sons.

Jelinke, Frederick. (1985). *Markov source modeling of text generation.* Impact of Processing Techniques on Communication. 569-598.

Lari, K. and Young, S. J. (1990). *The estimation of stochastic context-free grammars using the Inside-Outside algorithm.* Computer Speech and Language. Vol. 4, 35-56.

Mode, Charles J. (1971). *Multitype Branching Processes.* Elsevier.

Sankoff, David. (1971). *Branching processes with terminal types: Application to context-free grammars.* J. Appl. Prop. 8, 233-240.

Seneta, E. (1981). *Non-negative Matrices and Markov Chains. Second edition.* Springer Verlag.

Yasubumi, Sakakibara et.al. (1994) *Stochastic context-free grammars for tRNA modeling.* Nucleic Acids Research. Vol. 22. No. 23. 5112-5120.

Otter, Richard (1949). *The Multiplicative Process.* Ann.Maht.Statist. 20, 206-224.

Good, I. J. (1958). *Generalizations to several variables of Lagranges expansion, with applications to stochastic processes.*??.