Count of bi-modular hidden patterns under probabilistic dynamical systems

Manuel Lladser
University of Colorado - Boulder

In collaboration with L. Lhote

Frontier Probability Days 2014

Word Counting Problems

In what follows, $\mathcal{A}$ is a reference **alphabet**

Word Counting Problems

In what follows, $\mathcal{A}$ is a reference **alphabet**

Exact string matching : Given a text $\mathcal{T}$ and a pattern ω corresponding to a **single word** over the alphabet, find all the occurrences of ω in $\mathcal{T}$, where an occurrence is an exact matching with ω

ex : $\omega = aba$ occurs 3 times in the text $\mathcal{T} = \underline{aba}bcb\underline{ab\underline{a}ba}$

Multiple strings matching : Given a text $\mathcal{T}$ and a pattern $\omega = \{\omega_1, \omega_2, \ldots, \omega_\ell\}$ i.e. a **set of words** over the alphabet, find all the occurrences of words in $\mathcal{T}$

ex : $\omega = \{ab, ba\}$ occurs 8 times in the text $\mathcal{T} = \overline{\underline{ab}}\,\overline{\underline{ab}}cb\overline{\underline{ab}}\,\overline{\underline{ab}}\,\overline{\underline{ba}}$

Word Counting Problems

Hidden pattern matching : Given a text $\mathcal{T}$ and a **hidden pattern** $\omega = (\omega_1, \ldots, \omega_\ell)$ i.e. a **sequence of words** over the alphabet, find the occurrences of ω in $\mathcal{T}$, where an occurrence is a decomposition of the text of the form

$$\mathcal{T} = u_0 \cdot \omega_1 \cdot u_1 \cdot \omega_2 \cdot u_2 \cdots u_{\ell-1} \cdot \omega_\ell \cdot u_\ell, \text{ with } u_i \in \mathcal{A}^\star$$

Call $\omega_1, \ldots, \omega_\ell$ the **modules** of the pattern

ex : $\omega = (ab, ba)$ occurs 3 times in the text $\mathcal{T} = \underline{ab}c\underline{\overline{ab}}ac\overline{ba}$

Word Counting Problems

Hidden pattern matching : Given a text $\mathcal{T}$ and a **hidden pattern** $\omega = (\omega_1, \ldots, \omega_\ell)$ i.e. a **sequence of words** over the alphabet, find the occurrences of ω in $\mathcal{T}$, where an occurrence is a decomposition of the text of the form

$$\mathcal{T} = u_0 \cdot \omega_1 \cdot u_1 \cdot \omega_2 \cdot u_2 \cdots u_{\ell-1} \cdot \omega_\ell \cdot u_\ell, \text{ with } u_i \in \mathcal{A}^\star$$

Call $\omega_1, \ldots, \omega_\ell$ the **modules** of the pattern

$$\text{ex} : \omega = (ab, ba) \text{ occurs 3 times in the text } \mathcal{T} = \underline{ab}c\underline{\overline{ba}}c\overline{ba}$$

Generalized hidden pattern matching : Given a text $\mathcal{T}$ and a **generalized hidden pattern** $\omega = (\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_\ell)$, i.e. a **sequence of languages** over the alphabet, find the occurrences of ω in $\mathcal{T}$, where an occurrence is a match with any hidden pattern in $\mathcal{L}_1 \times \ldots \times \mathcal{L}_\ell$

$$\text{ex} : \omega = (\{ab, ba\}, \{cba\}) \text{ occurs 2 times in the text } \mathcal{T} = \underline{ab}\underline{ac}\underline{cba}a$$

Word Count Statistics

Fix a pattern ω. For each text $\mathcal{T}$ define :

$$|\mathcal{T}|_\omega \stackrel{def}{=} \#\text{ of occurrences of } \omega \text{ in } \mathcal{T}$$

	Memoryless source	Markov chain	Dynamical source
Exact Strings	**Régnier & Szpankowski (1997)**		**Bourdon & Vallée (2002)** **Bourdon & Vallée (2006)**
Multiple strings	**Régnier & Szpankowski (1997)**	**Bourdon & Vallée (2002)** **Bourdon & Vallée (2006)**	
Hidden pattern	**Flajolet, Szpankowski, Vallée (2002)**	**Bourdon & Vallée (2002)**	
Generalized hidden pattern	**Bourdon & Vallée (2002)**		

green=mean and variance red=limit law

Results for the number of occurrence positions e.g. by Nicodème, Salvy & Flajolet (2002)

Word Count Statistics

Fix a pattern ω. For each text $\mathcal{T}$ define :

$$|\mathcal{T}|_\omega \stackrel{def}{=} \#\text{ of occurrences of } \omega \text{ in } \mathcal{T}$$

	Memoryless source	Markov chain	Dynamical source
Exact Strings	**Régnier & Szpankowski (1997)**		**Bourdon & Vallée (2002)** **Bourdon & Vallée (2006)**
Multiple strings	**Régnier & Szpankowski (1997)**	**Bourdon & Vallée (2002)** **Bourdon & Vallée (2006)**	
Hidden pattern	**Flajolet, Szpankowski, Vallée (2002)**	**Bourdon & Vallée (2002)**	
Generalized hidden pattern	**Bourdon & Vallée (2002)**		

green=mean and variance red=limit law

Results for the number of occurrence positions e.g. by Nicodème, Salvy & Flajolet (2002)

Problem Statement

Chief goal : Prove the Gaussian limit law for the number of occurrences of generalized hidden patterns under probabilistic dynamical sources

Problem Statement

Chief goal : Prove the Gaussian limit law for the number of occurrences of generalized hidden patterns under probabilistic dynamical sources

Case study - simplifications :

- $\omega = (a, b)$, with a and b are different characters in $\mathcal{A}$; in particular, there cannot be overlap between the modules of the pattern

Problem Statement

Chief goal : Prove the Gaussian limit law for the number of occurrences of generalized hidden patterns under probabilistic dynamical sources

Case study - simplifications :

- $\omega = (a, b)$, with a and b are different characters in $\mathcal{A}$; in particular, there cannot be overlap between the modules of the pattern
- Normalized count : for technical reasons, we study the normalized count

$$C(\mathcal{T}) = \frac{|\mathcal{T}|_{(a,b)}}{|\mathcal{T}|_a}$$

Problem Statement

Chief goal : Prove the Gaussian limit law for the number of occurrences of generalized hidden patterns under probabilistic dynamical sources

Case study - simplifications :

- $\omega = (a, b)$, with a and b are different characters in $\mathcal{A}$; in particular, there cannot be overlap between the modules of the pattern
- Normalized count : for technical reasons, we study the normalized count

$$C(\mathcal{T}) = \frac{|\mathcal{T}|_{(a,b)}}{|\mathcal{T}|_a}$$

Theorem (Lhote-LI)

If C_n denotes the normalized count in a random text of length n produced by a holomorphic dynamical source then $\big(C_n - \mathbb{E}(C_n)\big)/\sqrt{\mathbb{V}(C_n)}$ is asymptotically Normal, where

$$\begin{aligned} \mathbb{E}(C_n) &\sim -\sigma'(1) \cdot n \\ \mathbb{V}(C_n) &\sim (\sigma'(1) - \sigma(1) - \sigma''(1)) \cdot n \end{aligned}$$

Problem Statement

Chief goal : Prove the Gaussian limit law for the number of occurrences of generalized hidden patterns under probabilistic dynamical sources

Case study - simplifications :

- $\omega = (a, b)$, with a and b are different characters in $\mathcal{A}$; in particular, there cannot be overlap between the modules of the pattern
- Normalized count : for technical reasons, we study the normalized count

$$C(\mathcal{T}) = \frac{|\mathcal{T}|_{(a,b)}}{|\mathcal{T}|_a}$$

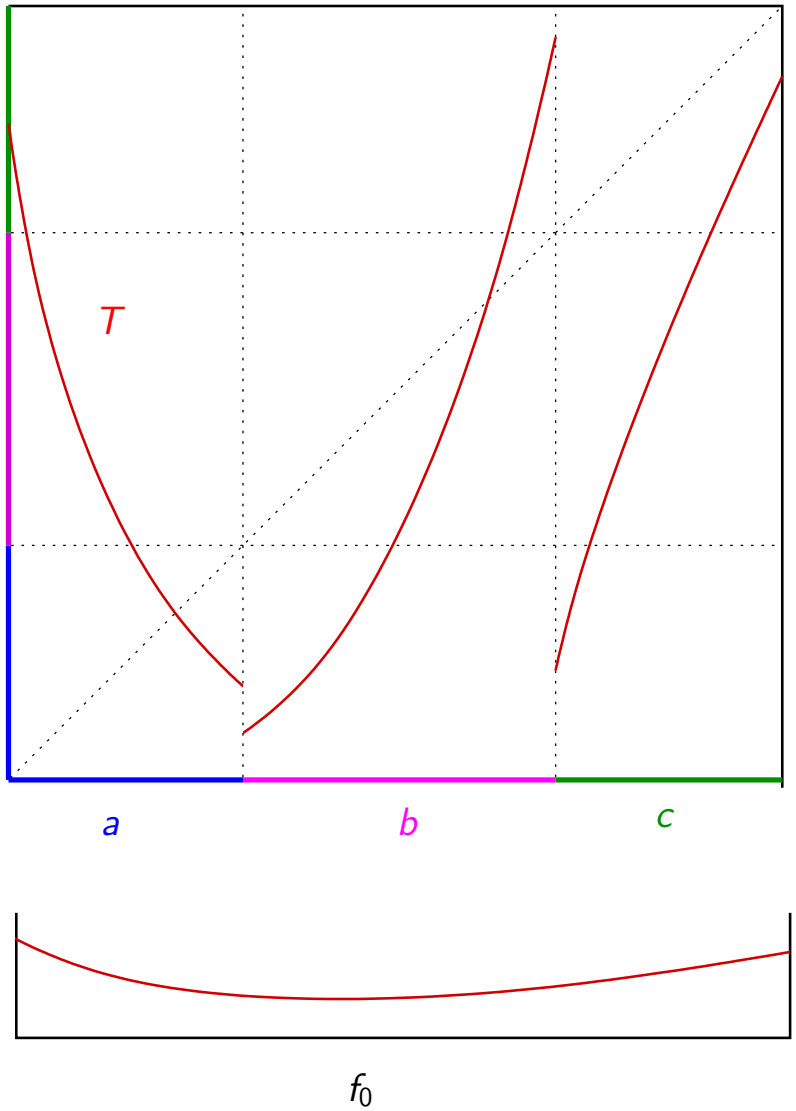
Theorem (Lhote-LI)

If C_n denotes the normalized count in a random text of length n produced by a holomorphic dynamical source then $\big(C_n - \mathbb{E}(C_n)\big)/\sqrt{\mathbb{V}(C_n)}$ is asymptotically Normal, where

$$\begin{aligned} \mathbb{E}(C_n) &\sim -\sigma'(1) \cdot n \\ \mathbb{V}(C_n) &\sim (\sigma'(1) - \sigma(1) - \sigma''(1)) \cdot n \end{aligned}$$

The interest in this result is only in the method of proof !

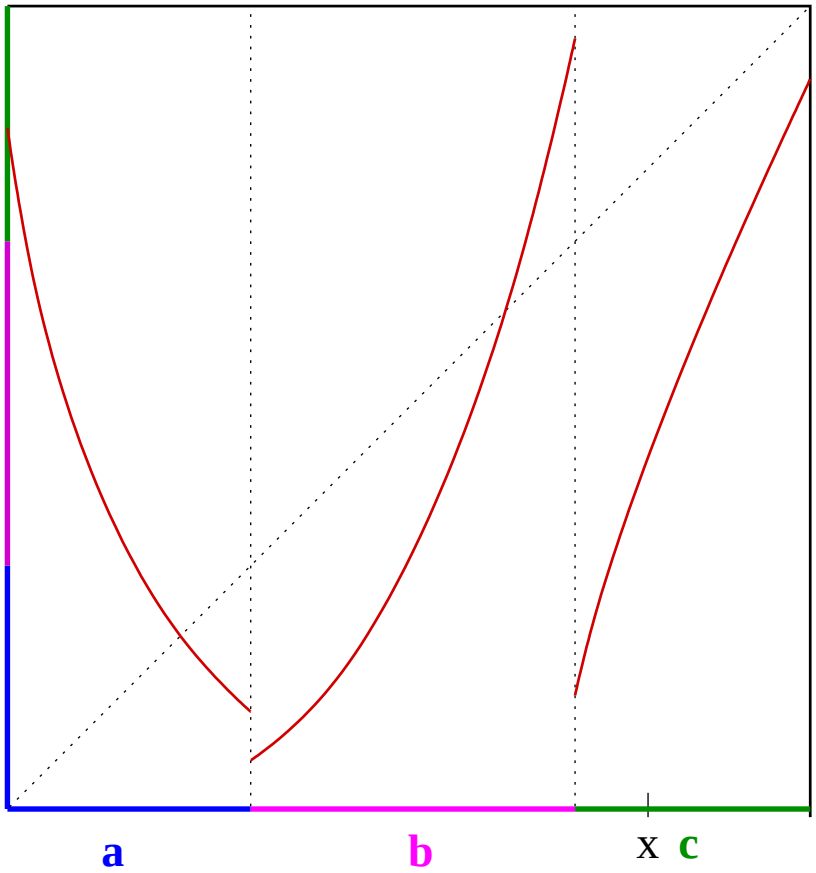
Definition of Dynamical Source (DS)



- an alphabet $\mathcal{A}$ e.g. $\{a, b, c\}$
- the open interval $I = (0, 1)$
- a topological partition $(I_\omega)_{\omega \in \mathcal{A}}$ of I
- a piecewise monotone and twice continuously differentiable function $T : I \to I$
- an initial probability density function f_0 over I

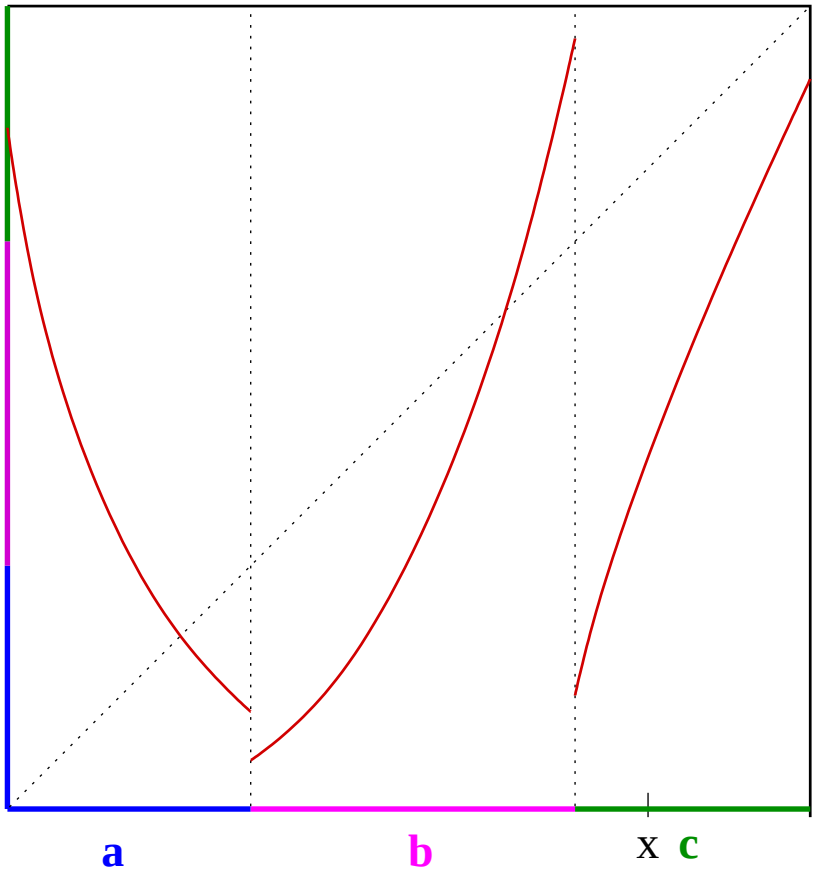
See Vallée (2001) for details

DS Character Emissions



Initialization : select X at random using probability densitiy function f_0

DS Character Emissions

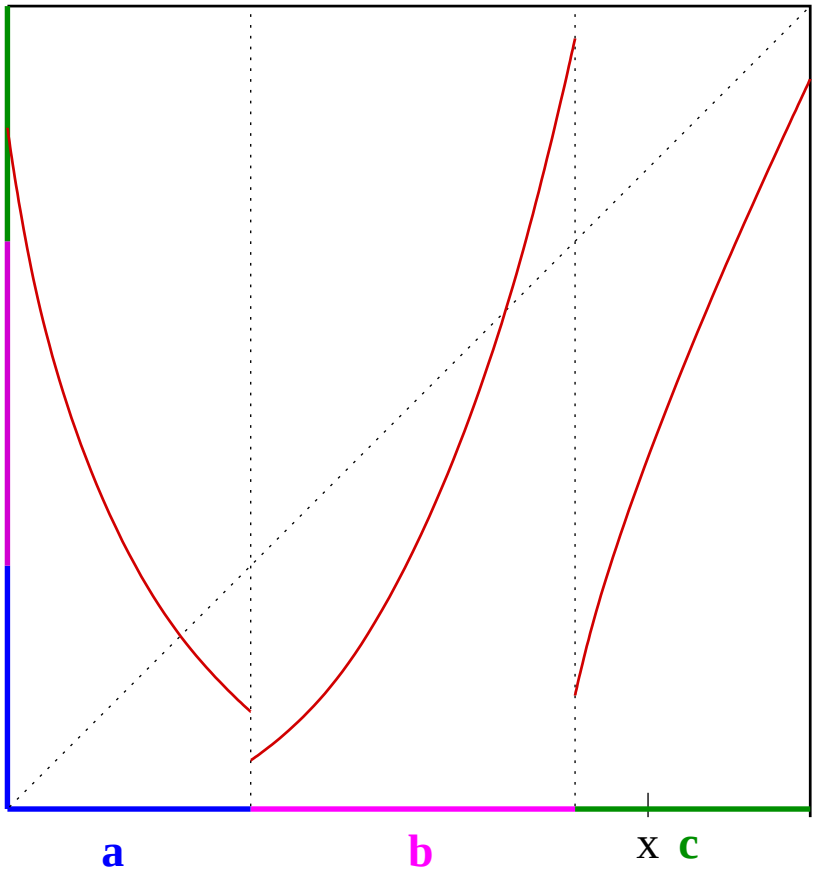


Initialization : select X at random using probability densitiy function f_0

Text emitted :

$$\mathcal{T} = \mathcal{T}(X) =$$

DS Character Emissions

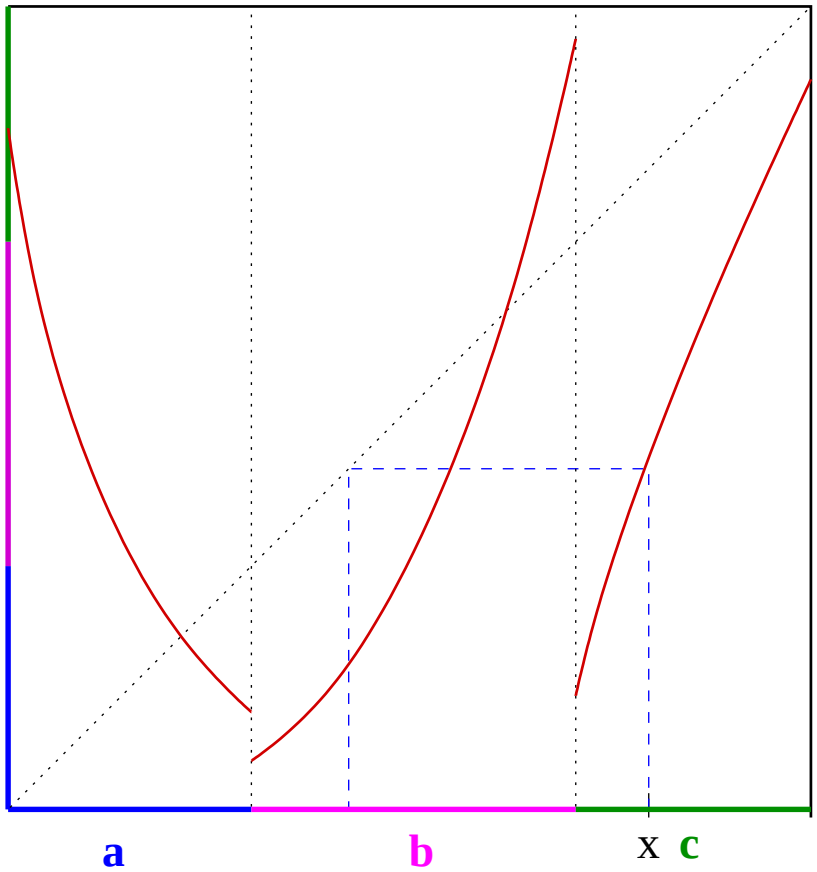


Initialization : select X at random using probability densitiy function f_0

Text emitted :

$$\mathcal{T} = \mathcal{T}(X) = c$$

DS Character Emissions

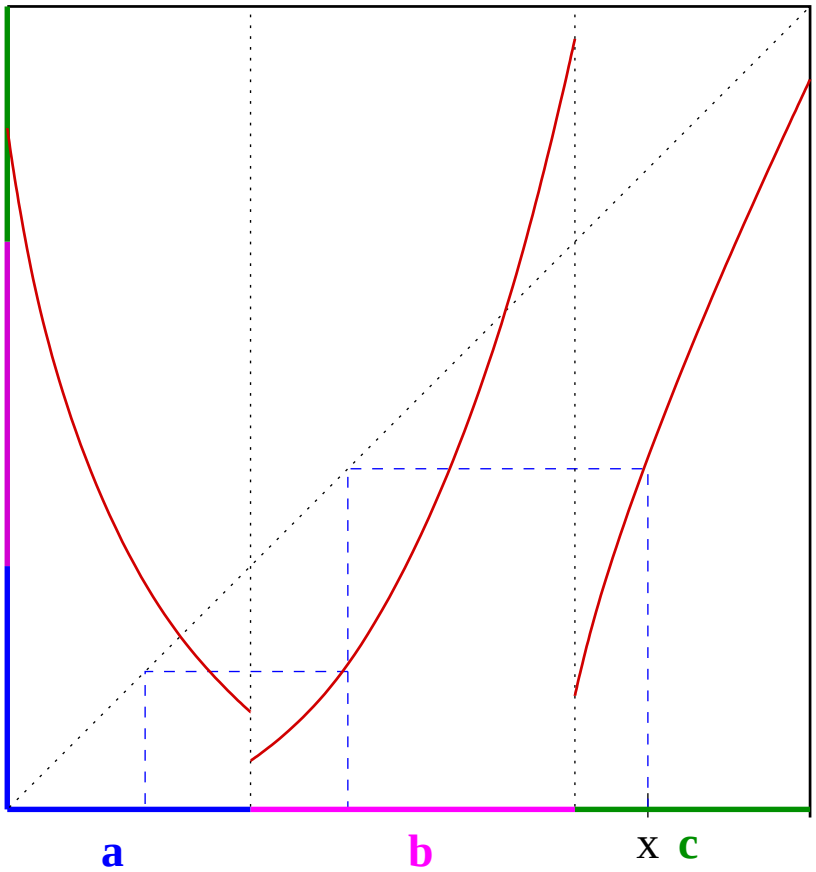


Initialization : select X at random using probability densitiy function f_0

Text emitted :

$$\mathcal{T} = \mathcal{T}(X) = cb$$

DS Character Emissions

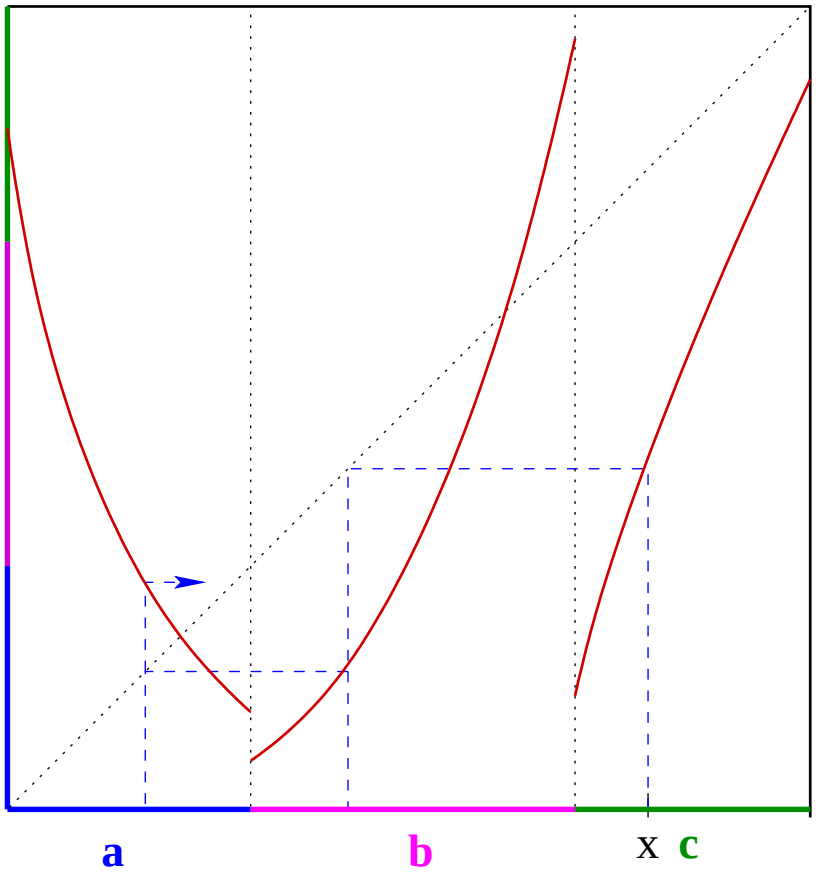


Initialization : select X at random using probability densitiy function f_0

Text emitted :

$$\mathcal{T} = \mathcal{T}(X) = cba$$

DS Character Emissions

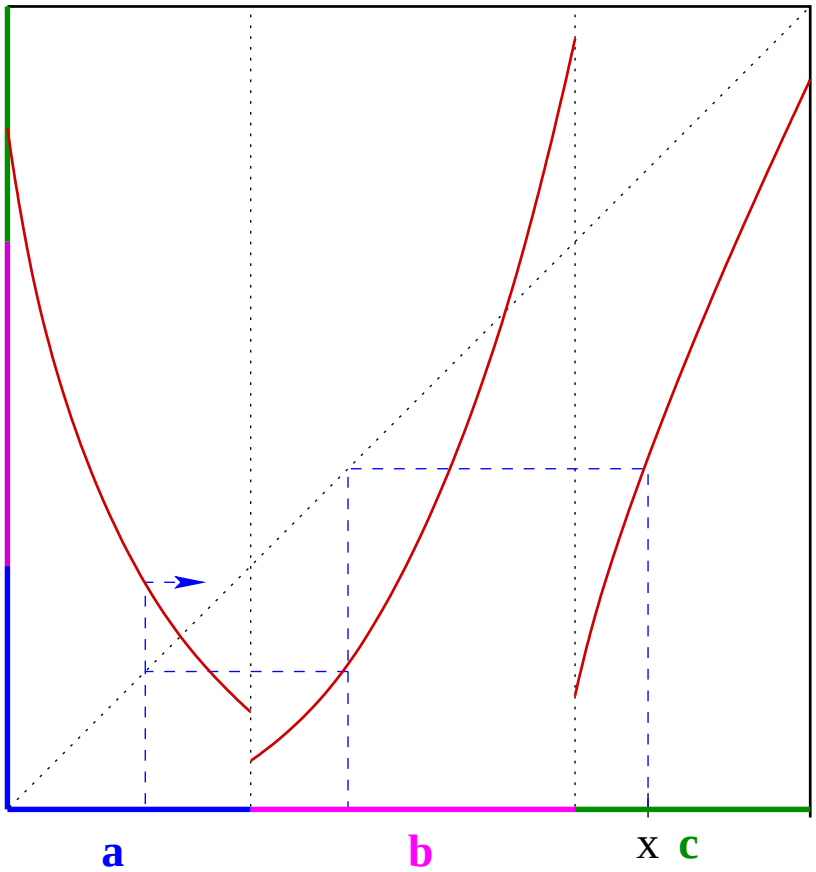


Initialization : select X at random using probability densitiy function f_0

Text emitted :

$$\mathcal{T} = \mathcal{T}(X) = cbaa\ldots$$

DS Character Emissions



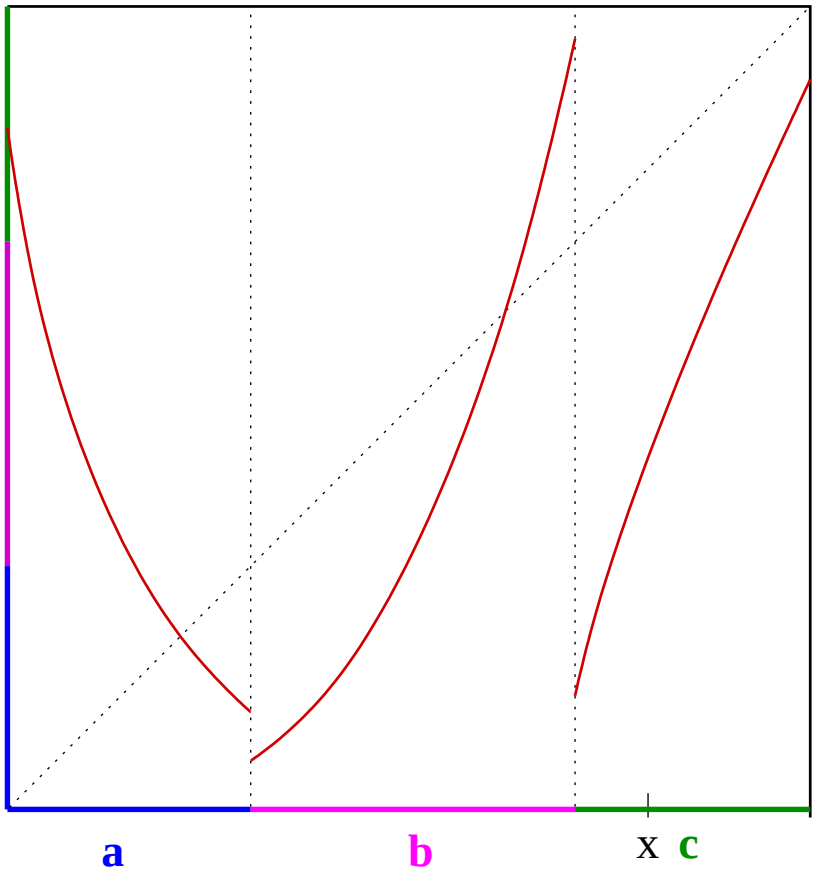
Initialization : select X at random using probability densitiy function f_0

Text emitted :

$$\mathcal{T} = \mathcal{T}(X) = cbaa\ldots$$

What's the chance that $\mathcal{T} = cbaa$...?

DS Character Emissions



Initialization : select X at random using probability densitiy function f_0
Text emitted :

$$\mathcal{T} = \mathcal{T}(X) = cbaa\ldots$$

What's the chance that $\mathcal{T} = cbaa$...?

ANS : For each character ω, need the **inverse functions** of the branches of T, denoted h_a, h_b and h_c respectively

DS Character Emissions

In what follows, $\mathcal{A}^*$ is the set of all words of finite length over the alphabet $\mathcal{A}$

Fundamental Identities (Bourdon-Vallée 2002)

For each $\omega \in \mathcal{A}^*$:

$$p_\omega \stackrel{def}{=} \mathbb{P}(\mathcal{T} = \omega\ldots) = \int_0^1 \mathbf{G}_{[\omega]}[f_0](t)dt$$

where

- for all $\alpha \in \mathcal{A}$: $(\mathbf{G}_{[\alpha]}f)(s) \stackrel{def}{=} (f \circ h_\alpha)(s) \cdot |h'_\alpha(s)|$
- for all $\omega_1, \omega_2 \in \mathcal{A}^\star$: $\mathbf{G}_{[\omega_1\omega_2]} = \mathbf{G}_{[\omega_2]} \circ \mathbf{G}_{[\omega_1]}$

DS Character Emissions

In what follows, $\mathcal{A}^*$ is the set of all words of finite length over the alphabet $\mathcal{A}$

Fundamental Identities (Bourdon-Vallée 2002)

For each $\omega \in \mathcal{A}^*$:

$$p_\omega \stackrel{def}{=} \mathbb{P}(\mathcal{T} = \omega\ldots) = \int_0^1 \mathbf{G}_{[\omega]}[f_0](t)dt$$

where

- for all $\alpha \in \mathcal{A}$: $(\mathbf{G}_{[\alpha]}f)(s) \stackrel{def}{=} (f \circ h_\alpha)(s) \cdot |h'_\alpha(s)|$
- for all $\omega_1, \omega_2 \in \mathcal{A}^\star$: $\mathbf{G}_{[\omega_1\omega_2]} = \mathbf{G}_{[\omega_2]} \circ \mathbf{G}_{[\omega_1]}$ **(note the reverse order !)**

DS Character Emissions

In what follows, $\mathcal{A}^*$ is the set of all words of finite length over the alphabet $\mathcal{A}$

Fundamental Identities (Bourdon-Vallée 2002)

For each $\omega \in \mathcal{A}^*$:

$$p_\omega \stackrel{def}{=} \mathbb{P}(\mathcal{T} = \omega \ldots) = \int_0^1 \mathbf{G}_{[\omega]}[f_0](t)dt$$

where

- for all $\alpha \in \mathcal{A}$: $(\mathbf{G}_{[\alpha]}f)(s) \stackrel{def}{=} (f \circ h_\alpha)(s) \cdot |h'_\alpha(s)|$
- for all $\omega_1, \omega_2 \in \mathcal{A}^\star$: $\mathbf{G}_{[\omega_1\omega_2]} = \mathbf{G}_{[\omega_2]} \circ \mathbf{G}_{[\omega_1]}$ **(note the reverse order !)**

Remark 1 : the above identity generalizes the equality $p_{\omega_1\omega_2} = p_{\omega_1} \cdot p_{\omega_2}$ for memoryless sources

DS Character Emissions

In what follows, $\mathcal{A}^*$ is the set of all words of finite length over the alphabet $\mathcal{A}$

Fundamental Identities (Bourdon-Vallée 2002)

For each $\omega \in \mathcal{A}^*$:

$$p_\omega \stackrel{def}{=} \mathbb{P}(\mathcal{T} = \omega \ldots) = \int_0^1 \mathbf{G}_{[\omega]}[f_0](t)dt$$

where

- for all $\alpha \in \mathcal{A}$: $(\mathbf{G}_{[\alpha]}f)(s) \stackrel{def}{=} (f \circ h_\alpha)(s) \cdot |h'_\alpha(s)|$
- for all $\omega_1, \omega_2 \in \mathcal{A}^\star$: $\mathbf{G}_{[\omega_1\omega_2]} = \mathbf{G}_{[\omega_2]} \circ \mathbf{G}_{[\omega_1]}$ **(note the reverse order !)**

Remark 1 : the above identity generalizes the equality $p_{\omega_1\omega_2} = p_{\omega_1} \cdot p_{\omega_2}$ for memoryless sources

Remark 2 : for sources with memory the operators $\mathbf{G}_{[\omega_2]}$ and $\mathbf{G}_{[\omega_1]}$ do not typically commute

Generating Functions Associated with Languages

Notation :

- $|\omega|$ = **size** of the word ω
- $c : \mathcal{A}^\star \to \mathbb{R}^+$ a **cost function** over $\mathcal{A}^\star$ e.g.

$$|\omega|_\ell := \text{ number of occurrences of the letter } \ell \text{ in } \omega$$

Generating Functions Associated with Languages

Notation :

- $|\omega|$ = **size** of the word ω
- $c : \mathcal{A}^\star \to \mathbb{R}^+$ a **cost function** over $\mathcal{A}^\star$ e.g.

$$|\omega|_\ell := \text{ number of occurrences of the letter } \ell \text{ in } \omega$$

Generating Operators (Bourdon-Vallée 2002)

The generating function relative to a language $\mathcal{L}$ and cost c is

$$L(z,u) \stackrel{def}{=} \sum_{\omega\in\mathcal{L}} z^{|\omega|} u^{c(\omega)} p_\omega = \int_0^1 \mathbf{L}(z,u)[f_0](t)dt,$$

where $\mathbf{L}(z,u)$ is the generating operator

$$\mathbf{L}(z,u) \stackrel{def}{=} \sum_{\omega\in\mathcal{L}} z^{|\omega|} u^{c(\omega)} \mathbf{G}_{[\omega]}$$

Symbolic Specifications

The identity $\mathbf{G}_{[\omega_1\omega_2]} = \mathbf{G}_{[\omega_2]} \circ \mathbf{G}_{[\omega_1]}$, for all $\omega_1, \omega_2 \in \mathcal{A}^\star$, allows the direct computation of the generating operators of certain new languages from the generating operators of old ones

New Language	**Memoryless source**	**Dynamical Source**
$\mathcal{L}_1 \cup \mathcal{L}_2$ (disjoint union)	$L_1(z,u) + L_2(z,u)$	$\mathbf{L}_1(z,u) + \mathbf{L}_2(z,u)$
$\mathcal{L}_1 \cdot \mathcal{L}_2$ (concatenation)	$L_1(z,u) \cdot L_2(z,u)$	$\mathbf{L}_2(z,u) \circ \mathbf{L}_1(z,u)$ (not commutative)
$\mathcal{L}^\star = \cup_{k\geq 0}\mathcal{L}^k$ (star operation)	$\dfrac{1}{1-L(z,u)}$	$(\mathbf{I} - \mathbf{L}(z,u))^{-1}$ (quasi-inverse)

See Bourdon-Vallée (2002, 2006) for details

Back to the "toy" problem

- (a, b) is the bi-modular pattern with $a \neq b$

Back to the "toy" problem

- (a, b) is the bi-modular pattern with $a \neq b$
- $\mathcal{B} = \mathcal{A} \backslash \{a\}$
- $\mathcal{C} = \mathcal{A} \backslash \{a, b\}$

Back to the "toy" problem

- (a, b) is the bi-modular pattern with $a \neq b$
- $\mathcal{B} = \mathcal{A} \backslash \{a\}$
- $\mathcal{C} = \mathcal{A} \backslash \{a, b\}$
- $\mathcal{K} = a \cdot \mathcal{B}^{\star}$

Back to the "toy" problem

- (a, b) is the bi-modular pattern with $a \neq b$
- $\mathcal{B} = \mathcal{A} \backslash \{a\}$
- $\mathcal{C} = \mathcal{A} \backslash \{a, b\}$
- $\mathcal{K} = a \cdot \mathcal{B}^\star$
- Generating operator of $\mathcal{K}$ where z marks "length" and u the cost $|\cdot|_b$:

$$\mathbf{K}(z, u) = \sum_{\omega \in \mathcal{K}} z^{|\omega|} u^{|\omega|_b} \mathbf{G}_{[\omega]}$$

Back to the "toy" problem

- (a, b) is the bi-modular pattern with $a \neq b$
- $\mathcal{B} = \mathcal{A} \backslash \{a\}$
- $\mathcal{C} = \mathcal{A} \backslash \{a, b\}$
- $\mathcal{K} = a \cdot \mathcal{B}^\star$
- Generating operator of $\mathcal{K}$ where z marks "length" and u the cost $|\cdot|_b$:

$$\mathbf{K}(z, u) = \sum_{\omega \in \mathcal{K}} z^{|\omega|} u^{|\omega|_b} \mathbf{G}_{[\omega]}$$

Proposition (Lhote-LI)

$$\mathbf{K}(z, u) = \Big(\mathbf{I} - z(\mathbf{C} + u\mathbf{G}_{[b]})\Big)^{-1} \circ z\mathbf{G}_{[a]}$$

Back to the "toy" problem

- the language of words with exactly m occurrences of the character a is

$$\mathcal{L}_m = \mathcal{B}^\star \mathcal{K}^m = \mathcal{B}^\star \cdot a\mathcal{B}^\star \cdot a\mathcal{B}^\star \cdots a\mathcal{B}^\star$$

Back to the "toy" problem

- the language of words with exactly m occurrences of the character a is

$$\mathcal{L}_m = \mathcal{B}^\star \mathcal{K}^m = \mathcal{B}^\star \cdot a\mathcal{B}^\star \cdot a\mathcal{B}^\star \cdots a\mathcal{B}^\star$$

- each $\omega \in \mathcal{L}_m$ may be uniquely written in the form

$$\omega = \omega_0 \cdot a\omega_1 \cdot a\omega_2 \cdots a\omega_m,$$

with $\omega_i \in \mathcal{B}^\star$, for all i. Thus, for $\mathcal{T} \in \mathcal{L}_m$:

$$\begin{aligned} C(\mathcal{T}) &\stackrel{def}{=} \frac{|\mathcal{T}|_{(a,b)}}{|\mathcal{T}|_a} \\ &= \frac{1}{m}|\omega_1|_b + \frac{2}{m}|\omega_2|_b + \cdots + \frac{m}{m}|\omega_m|_b \end{aligned}$$

Back to the "toy" problem

- the language of words with exactly m occurrences of the character a is

$$\mathcal{L}_m = \mathcal{B}^\star \mathcal{K}^m = \mathcal{B}^\star \cdot a\mathcal{B}^\star \cdot a\mathcal{B}^\star \cdots a\mathcal{B}^\star$$

- each $\omega \in \mathcal{L}_m$ may be uniquely written in the form

$$\omega = \omega_0 \cdot a\omega_1 \cdot a\omega_2 \cdots a\omega_m,$$

with $\omega_i \in \mathcal{B}^\star$, for all i. Thus, for $\mathcal{T} \in \mathcal{L}_m$:

$$\begin{aligned} C(\mathcal{T}) &\stackrel{def}{=} \frac{|\mathcal{T}|_{(a,b)}}{|\mathcal{T}|_a} \\ &= \frac{1}{m}|\omega_1|_b + \frac{2}{m}|\omega_2|_b + \cdots + \frac{m}{m}|\omega_m|_b \end{aligned}$$

Proposition (Lhote-LI)

Recall that $\mathcal{K} = a\mathcal{B}^*$. For $m \geq 1$, the generating operator relative to $\mathcal{L}_m$ and C is

$$\mathbf{L}_m(z,u) = \mathbf{K}(z, u^{\frac{m}{m}}) \circ \cdots \circ \mathbf{K}(z, u^{\frac{2}{m}}) \circ \mathbf{K}(z, u^{\frac{1}{m}}) \circ (\mathbf{I} - z\mathbf{B})^{-1}$$

Back to the "toy" problem

Recall that $\mathcal{L}_m = \mathcal{B}^* \mathcal{K}^m$, where $\mathcal{K} = a\mathcal{B}^*$

Why do we care about the language $\mathcal{L}_m$ and its associated operator $\mathbf{L}_m(z, u)$?

Back to the "toy" problem

Recall that $\mathcal{L}_m = \mathcal{B}^* \mathcal{K}^m$, where $\mathcal{K} = a\mathcal{B}^*$

Why do we care about the language $\mathcal{L}_m$ and its associated operator $\mathbf{L}_m(z,u)$?

ANS : To understand the asymptotic behaviour for u real near 1 of

$$\begin{aligned}
\mathbb{E}(u^{C_n}) &= [z^n] \sum_{\omega \in \mathcal{A}^*} p_\omega z^{|\omega|} u^{C(\omega)} \\
&= [z^n] \sum_{m=0}^{\infty} L_m(z,u) \\
&= \sum_{m=0}^{\infty} \int_0^1 [z^n] \mathbf{L}_m(z,u)[f](t)\, dt
\end{aligned}$$

If only the source was memoryless ...

Generating operators :

$$\mathbf{L}_m(z,u) = \mathbf{K}(z,u^{\frac{m}{m}}) \circ \dots \mathbf{K}(z,u^{\frac{2}{m}}) \circ \mathbf{K}(z,u^{\frac{1}{m}}) \circ (\mathbf{I} - z\mathbf{B})^{-1}$$

Generating functions of memoryless sources :

$$\begin{aligned} L_m(z,u) &= \frac{1}{1-zB} \prod_{k=1}^{m} K(z,u^{\frac{k}{m}}) \\ &= \frac{1}{1-zB} \exp\left[\sum_{k=1}^{m} \log K(z,u^{\frac{k}{m}})\right] \\ &= \frac{1}{1-zB} \exp\left[m \int_0^1 \log K(z,u^t)dt + (\text{explicit error terms})\right] \end{aligned}$$

If only the source was memoryless ...

Generating operators :

$$\mathbf{L}_m(z,u) = \mathbf{K}(z,u^{\frac{m}{m}}) \circ \dots \mathbf{K}(z,u^{\frac{2}{m}}) \circ \mathbf{K}(z,u^{\frac{1}{m}}) \circ (\mathbf{I} - z\mathbf{B})^{-1}$$

Generating functions of memoryless sources :

$$\begin{aligned} L_m(z,u) &= \frac{1}{1-zB} \prod_{k=1}^{m} K(z,u^{\frac{k}{m}}) \\ &= \frac{1}{1-zB} \exp\left[\sum_{k=1}^{m} \log K(z,u^{\frac{k}{m}})\right] \\ &= \frac{1}{1-zB} \exp\left[m \int_0^1 \log K(z,u^t)dt + (\text{explicit error terms})\right] \end{aligned}$$

Technical Bottleneck

In the context of dynamical sources the exp-log transformation of a composition of non-commuting operators is not defined !

Asymptotic Normality : proof sketch

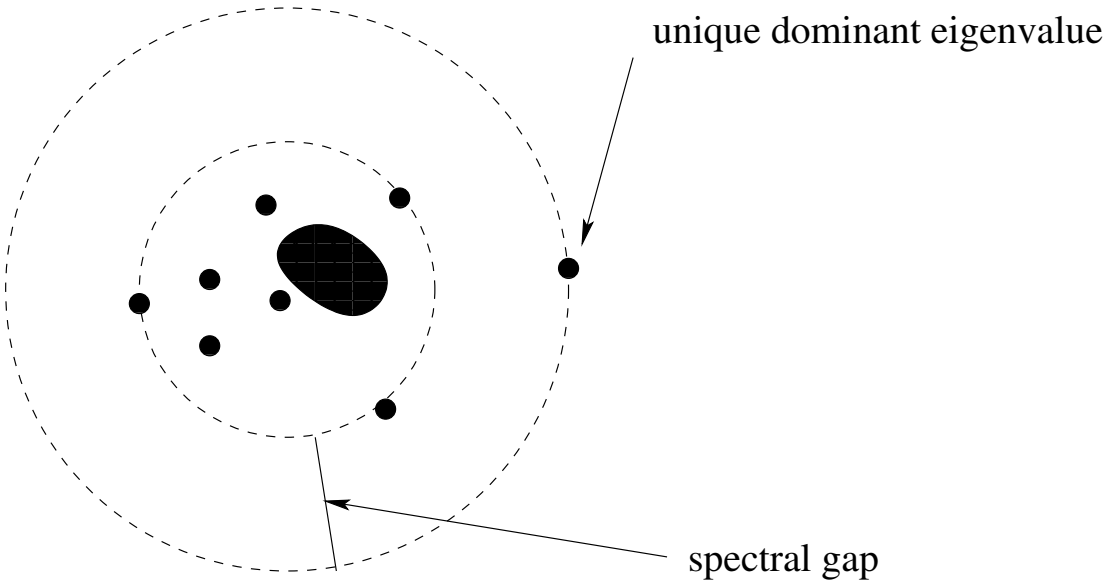
Recall that

$$\mathbf{K}(z,u) = \Big(\mathbf{I} - z(\mathbf{C} + u\mathbf{G}_{[b]})\Big)^{-1} \circ z\mathbf{G}_{[a]}$$

is the operator associated with $a\mathcal{B}^*$, where z marks "length" and u the character b

Proposition (Lhote-LI)

The operator $\mathbf{K}(z,u)$ admits a unique simple dominant eigenvalue $\lambda(z,u)$ isolated from the remainder of the spectrum by a spectral gap



Asymptotic Normality : proof sketch

Hence the spectral decomposition :

$$\mathbf{K}(z,u) = \lambda(z,u)\mathbf{P}(z,u) + \mathbf{R}(z,u),$$

with $\mathbf{P}(z,u) \circ \mathbf{P}(z,u) = \mathbf{P}(z,u)$ and $\mathbf{P}(z,u) \circ \mathbf{R}(z,u) = \mathbf{R}(z,u) \circ \mathbf{P}(z,u) = 0$

Asymptotic Normality : proof sketch

Hence the spectral decomposition :

$$\mathbf{K}(z,u) = \lambda(z,u)\mathbf{P}(z,u) + \mathbf{R}(z,u),$$

with $\mathbf{P}(z,u) \circ \mathbf{P}(z,u) = \mathbf{P}(z,u)$ and $\mathbf{P}(z,u) \circ \mathbf{R}(z,u) = \mathbf{R}(z,u) \circ \mathbf{P}(z,u) = 0$

Main Result : Asymptotics for a product of non-commutative operators

$$\begin{aligned}\prod_{j=m}^{1} \mathbf{K}(z,u^{\frac{j}{m}}) &= \prod_{j=1}^{m} \lambda(z,u^{\frac{j}{m}}) \times \prod_{j=m}^{1} \mathbf{P}(z,u^{\frac{j}{m}}) \times \left(1 + O\left(\frac{1}{m}\right)\right) \\ &= \prod_{j=1}^{m} \lambda(z,u^{\frac{j}{m}}) \times \mathbf{d}(z,u) \times \left(1 + O\left(\frac{\log u}{m}\right)\right)\end{aligned}$$

where the operator $\mathbf{d}(z,u)$ does not depend on m

See Lhote-Lladser (2012) for details

Asymptotic Normality : proof sketch

Using the Euler Mac-Laurin formula, we can now prove :

Corollary (Lhote-LI)

The generating function the language $\mathcal{L}_m = \mathcal{B}^*\mathcal{K}^m$, with $\mathcal{K} = a\mathcal{B}^*$, satisfies

$$L_m(z,u) = A(z,u)^m d(z,u)(1 + O(m^{-1})),$$

where

$$A(z,u) \stackrel{def}{=} \exp\left[\int_0^1 \log\lambda(z,u^t)dt\right],$$

and $d(z,u)$ is an analytic function that does not depend on m

Asymptotic Normality : proof sketch

1. We have $L_m(z,u) = A(z,u)^m d(z,u) + O\left(\frac{A(z,u)^m}{m}\right)$
2. It follows that

$$L(z,u) = \frac{d(z,u)}{1-A(z,u)} + O\left(\log \frac{1}{1-|A(z,u)|}\right),$$

for (z,u) such that $|A(z,u)| < 1$

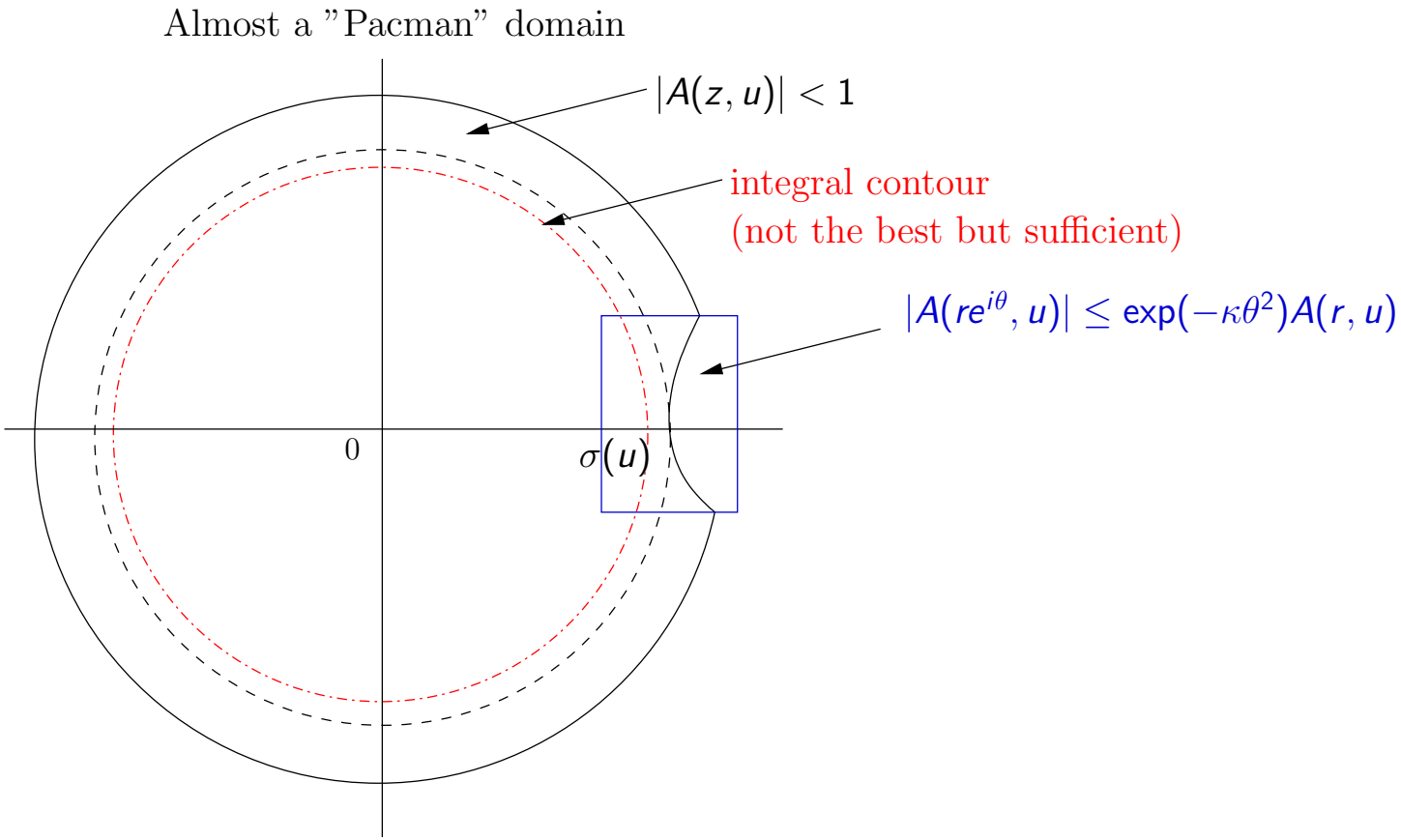
3. $$\left\{ \begin{array}{c} A(1,1) = 1 \\ A'_z(1,1) \cdot A'_u(1,1) \neq 0 \end{array} \right\} \Rightarrow \exists \sigma \ s.t. \ A(\sigma(u),u) = 1$$

4. $\sigma(u)$ is a simple pole of the "dominant part" of $L(z,u)$ so that

$$[z^n] \frac{d(z,u)}{1-A(z,u)} = \frac{d(\sigma(u),u)}{A'_z(\sigma(u),u)} \sigma(u)^{-n-1}(1+O(\rho^n))$$

with $\rho < 1$.

The O-part : large exponent m



Localization of Cauchy's formula over the red-contour gives "good" bounds for large powers of $A(z,u)$ when m is linear in n, for instance to extract the coefficient of z^n of $\frac{A(z,u)^m}{m}$

The O-part : small exponent m

Large deviations results on "number of occurrences of a" give :

Lemma (Lhote-LI)

There exist $\epsilon > 0$ and $\rho_\epsilon < 1$ such that

$$[z^n] \sum_{m \leq \epsilon n} L_m(z, u) = O(\rho_\epsilon^n \sigma(u)^{-n}).$$

Finally the Gaussian limit ...

Proposition (Lhote-LI)

There exist a real neighborhood $\mathcal{U}$ of $u = 1$ such that, for each $u \in \mathcal{U}$ and $\xi > 0$,

$$\mathrm{E}(u^{C_n}) = [z^n]L(z,u) = \alpha(u) \cdot \sigma(u)^{-n}(1 + O(n^{-\frac{1}{2}+\xi} \log n)),$$

where the constant in the O-term is uniform for all $u \in \mathcal{U}$, and

$$\alpha(u) = \frac{d(\sigma(u), u)}{\sigma(u)\, A'_z(\sigma(u), u)}.$$

Finally the Gaussian limit ...

Proposition (Lhote-LI)

There exist a real neighborhood $\mathcal{U}$ of $u = 1$ such that, for each $u \in \mathcal{U}$ and $\xi > 0$,

$$\mathrm{E}(u^{C_n}) = [z^n]L(z,u) = \alpha(u) \cdot \sigma(u)^{-n}(1 + O(n^{-\frac{1}{2}+\xi} \log n)),$$

where the constant in the O-term is uniform for all $u \in \mathcal{U}$, and

$$\alpha(u) = \frac{d(\sigma(u), u)}{\sigma(u)\, A'_z(\sigma(u), u)}.$$

Define $U(s) = -\log \sigma(e^s)$ and $V(s) = \log \alpha(e^s)$. If $C_n^\star = \frac{C_n - nU'(0)}{\sqrt{nU''(0)}}$ then

$$\mathrm{E}(e^{sC_n^\star}) = \frac{s^2}{2} + O\left(\frac{1}{\sqrt{n}} + \frac{\log n}{n^{\frac{1}{2}-\xi}}\right)$$

By the Lévy's Continuity Theorem, $C_n^\star$ is asymptotically Normal !

Finally the Gaussian limit ...

Proposition (Lhote-LI)

There exist a real neighborhood $\mathcal{U}$ of $u = 1$ such that, for each $u \in \mathcal{U}$ and $\xi > 0$,

$$\mathrm{E}(u^{C_n}) = [z^n]L(z,u) = \alpha(u) \cdot \sigma(u)^{-n}(1 + O(n^{-\frac{1}{2}+\xi} \log n)),$$

where the constant in the O-term is uniform for all $u \in \mathcal{U}$, and

$$\alpha(u) = \frac{d(\sigma(u), u)}{\sigma(u)\, A'_z(\sigma(u), u)}.$$

Define $U(s) = -\log \sigma(e^s)$ and $V(s) = \log \alpha(e^s)$. If $C_n^\star = \frac{C_n - nU'(0)}{\sqrt{nU''(0)}}$ then

$$\mathrm{E}(e^{sC_n^\star}) = \frac{s^2}{2} + O\left(\frac{1}{\sqrt{n}} + \frac{\log n}{n^{\frac{1}{2}-\xi}}\right)$$

By the Lévy's Continuity Theorem, $C_n^\star$ is asymptotically Normal !

... Thank you !