The growing tree distribution over Boolean functions.

ALÉA'11
CIRM, Marseille

Brigitte Chauvin, Danièle Gardy and Cécile Mailler

Laboratoire de Mathématiques de Versailles
Université de Versailles-Saint-Quentin

March 10, 2011

Our interest

Our aim: defining a random distribution over the set $\mathcal{F}_k$ of Boolean functions of k variables.

What's a Boolean function ?

$$\begin{aligned} f: \quad \{0,1\}^k &\longrightarrow \{0,1\} \\ (x_1,\ldots,x_k) &\longrightarrow f(x_1,\ldots,x_k) \end{aligned}$$

$1 = $ *True*
$0 = $ *False*

Our interest

Our aim: defining a random distribution over the set $\mathcal{F}_k$ of Boolean functions of k variables.

What's a Boolean function ?

$$f: \begin{array}{ccc} \{0,1\}^k & \longrightarrow & \{0,1\} \\ (x_1,\dots,x_k) & \longrightarrow & f(x_1,\dots,x_k) \end{array}$$

1	=	*True*
0	=	*False*

- The uniform law over the set of Boolean functions of k variables.
  [Shannon. *The synthesis of two-terminal switching circuits*]

Complexity

The complexity of f, denoted by $L(f)$, is the minimal number of connectives needed to represent f by a Boolean expression. E.g. $xXORy = (x \wedge \overline{y}) \vee (\overline{x} \wedge y)$ has complexity 3.

Shannon effect: *most* Boolean functions have exponential complexity in k - asymptotically when $k \longrightarrow \infty$.

How can we define other laws over $\mathcal{F}_k$?

Idea: consider the tree representation of a Boolean expression.

[Lefmann, Savický. *Some typical properties of large And/Or Boolean formulas*]

[Chauvin, Flajolet, Gardy, Gittenberger. *And/or trees revisited*]

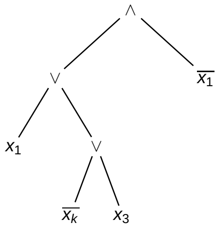
How can we define other laws over $\mathcal{F}_k$?

Idea: consider the tree representation of a Boolean expression.

[Lefmann, Savický. *Some typical properties of large And/Or Boolean formulas*]

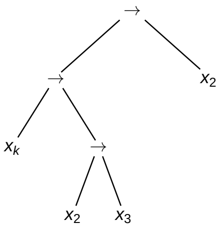
[Chauvin, Flajolet, Gardy, Gittenberger. *And/or trees revisited*]

The $\wedge/\vee$ model.



connectives: $\wedge$ or $\vee$
positive and negative literals

The implication model.



unique connective: $\to$
only positive literals

Some definitions

Size of a binary tree:

The size of a binary tree is the number of its internal nodes.

Complexity of a function f:

The complexity of f is the size of the minimal trees computing f.

Some definitions

Size of a binary tree:

The size of a binary tree is the number of its internal nodes.

Complexity of a function f:

The complexity of f is the size of the minimal trees computing f.

a Boolean expression = a binary planar labelled tree	$\mathcal{E}_k$
$\downarrow$ "represents"	$\downarrow \Phi$
a Boolean function	$\mathcal{F}_k$

Φ is not injective.

The Catalan trees law

the set of all trees of size n labelled with k variables	uniform law $\mathbb{U}_{n,k}$
$\downarrow \Phi$	$\downarrow \Phi$
$\mathcal{F}_k$	$\mu_{n,k}$

The Catalan trees law

the set of all trees of size n labelled with k variables $\downarrow \Phi$ $\mathcal{F}_k$

uniform law $\mathbb{U}_{n,k}$ $\downarrow \Phi$ $\mu_{n,k}$

Theorem:

The sequence $(\mu_{n,k})_n$ converges to a probability law μ_k when $n \longrightarrow +\infty$.

[Lefmann, Savický] [Chauvin, Flajolet, Gardy, Gittenberger.]

Theorem:

We have $\forall f \in \mathcal{F}_k$: $\mu_k(f) = \Theta\left(\frac{1}{k^{L(f)+2}}\right)$.

[Kozik. *Subcritical pattern languages for and/or trees*]

Moreover, there is no Shannon effect. [Genitrini, Gittenberger. *No Shannon effect on probability distributions on Boolean functions induced by random expressions.*]

The Galton-Watson model

- Let us consider a critical Galton-Watson process.
- Let us label it according to one of the two labelling models, choosing for each node **uniformly** at random among the possible labels.

$\rightarrow$ a random tree labelled with k variables, random variable in $\mathcal{E}_k$ which law is denoted by Π_k.

$\rightarrow$ we denote by π_k the law induced by Π_k over $\mathcal{F}_k$ through Φ.
[Chauvin, Flajolet, Gardy, Gittenberger. *And/or trees revisited*]

Theorem [Chauvin et al.][Kozik][Genitrini, Gittenberger]:

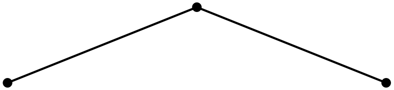
The distribution π_k gives more weight to low complexity functions, though it weights all functions that can be represented in the chosen logical system.
There is no Shannon effect.

The growing tree

This model is inspired of the random BST growing process.
A growing tree grows choosing uniformly the leaf which grows:

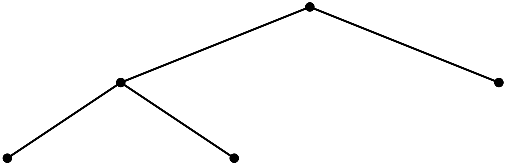
The growing tree

This model is inspired of the random BST growing process.
A growing tree grows choosing uniformly the leaf which grows:



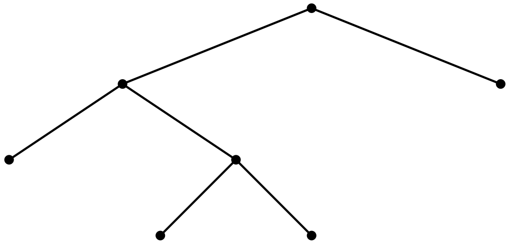
The growing tree

This model is inspired of the random BST growing process.
A growing tree grows choosing uniformly the leaf which grows:



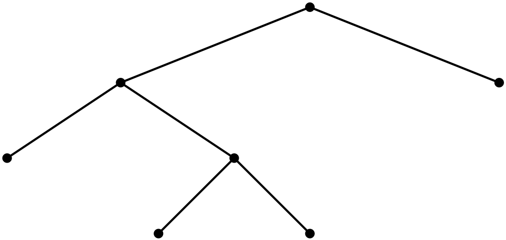
The growing tree

This model is inspired of the random BST growing process.
A growing tree grows choosing uniformly the leaf which grows:



The growing tree

This model is inspired of the random BST growing process.
A growing tree grows choosing uniformly the leaf which grows:



We stop after the n^{th} step.

$\longrightarrow$ non labelled random tree.
$\longrightarrow$ we label it "uniformly" at random: labelled random tree with law $\mathbb{P}_{n,k}$.
$\longrightarrow$ $p_{n,k}$ is the law induced by $\mathbb{P}_{n,k}$ over $\mathcal{F}_k$ through Φ.

$\longrightarrow$ non labelled random tree.
$\longrightarrow$ we label it "uniformly" at random: labelled random tree with law $\mathbb{P}_{n,k}$.
$\longrightarrow$ $p_{n,k}$ is the law induced by $\mathbb{P}_{n,k}$ over $\mathcal{F}_k$ through Φ.

Questions

- Does $p_{n,k}$ converge when n tends to $+\infty$?
- If it does, what is the asymptotic law?
- And what is the speed of convergence?

Methods

- Analytic combinatorics.
- Probabilistic method via Yule trees.

Main results

Theorem:

In the $\wedge/\vee$ labelling model,
$p_{n,k} \longrightarrow p_k = \frac{1}{2}\delta_{True} + \frac{1}{2}\delta_{False}$ when n tends to $+\infty$, and

$$\|p_{n,k} - p_k\|_\infty = O\left(\frac{1}{\ln n}\right).$$

Theorem:

In the implication model,
$p_{n,k} \longrightarrow p_k = \delta_{True}$ when n tends to $+\infty$, and

$$\|p_{n,k} - p_k\|_\infty = O\left(\frac{1}{\ln n}\right).$$

Main results

Theorem:

In the $\wedge/\vee$ labelling model,
$p_{n,k} \longrightarrow p_k = \frac{1}{2}\delta_{True} + \frac{1}{2}\delta_{False}$ when n tends to $+\infty$, and

$$\|p_{n,k} - p_k\|_\infty = O\left(\frac{1}{\ln n}\right).$$

Theorem:

In the implication model,
$p_{n,k} \longrightarrow p_k = \delta_{True}$ when n tends to $+\infty$, and

$$\|p_{n,k} - p_k\|_\infty = O\left(\frac{1}{\ln n}\right).$$

There is no Shannon effect.
The model is very simple.

Study of the growing tree law. The $\wedge/\vee$ labelling.

Analytic combinatorics approach.

Study of generating functions

$$\Phi_f(z) = \sum_{n\geq 0} p_{n,k}(f) z^n$$

Let $n \in \mathbb{N}^*$ and let $f \in \mathcal{F}_k$. We have:

$$p_{n+1,k}(f) = \frac{1}{2} \sum_{g\wedge h=f} \sum_{i=0}^{n} \frac{1}{n+1} p_{i,k}(g) p_{n-i,k}(h) + \frac{1}{2} \sum_{g\vee h=f} \sum_{i=0}^{n} \frac{1}{n+1} p_{i,k}(g) p_{n-i,k}(h).$$

Study of generating functions

$$\Phi_f(z) = \sum_{n\geq 0} p_{n,k}(f) z^n$$

Let $n \in \mathbb{N}^*$ and let $f \in \mathcal{F}_k$. We have:

$$p_{n+1,k}(f) = \frac{1}{2} \sum_{g\wedge h=f} \sum_{i=0}^{n} \frac{1}{n+1} p_{i,k}(g) p_{n-i,k}(h) + \frac{1}{2} \sum_{g\vee h=f} \sum_{i=0}^{n} \frac{1}{n+1} p_{i,k}(g) p_{n-i,k}(h).$$

Summing over n and then deriving, we obtain a **differential** system:

$$\boxed{2\phi_f'(z) = \sum_{g\wedge h=f} \phi_g(z)\phi_h(z) + \sum_{g\vee h=f} \phi_g(z)\phi_h(z)}$$

Study of generating functions

$$\Phi_f(z) = \sum_{n\geq 0} p_{n,k}(f) z^n$$

Let $n \in \mathbb{N}^*$ and let $f \in \mathcal{F}_k$. We have:

$$p_{n+1,k}(f) = \frac{1}{2} \sum_{g\wedge h=f} \sum_{i=0}^{n} \frac{1}{n+1} p_{i,k}(g) p_{n-i,k}(h) + \frac{1}{2} \sum_{g\vee h=f} \sum_{i=0}^{n} \frac{1}{n+1} p_{i,k}(g) p_{n-i,k}(h).$$

Summing over n and then deriving, we obtain a **differential** system:

$$\boxed{2\phi_f'(z) = \sum_{g\wedge h=f} \phi_g(z)\phi_h(z) + \sum_{g\vee h=f} \phi_g(z)\phi_h(z)}$$

Simplifying the system considering equivalency classes of functions that share the same generating functions.
For example: $f \sim \bar{f}$.

Study of the differential system

$$\begin{cases} \phi_T' &= P_T(\phi_T, \phi_1, \dots, \phi_q) \\ \phi_1' &= P_1(\phi_T, \phi_1, \dots, \phi_q) \\ &\vdots \\ \phi_q' &= P_q(\phi_T, \phi_1, \dots, \phi_q) \end{cases}$$

where $P_T, P_1, \dots, P_q$ are homogeneous quadratic polynomials.

Study of the differential system

$$\begin{cases} \phi_T' &= P_T(\phi_T, \phi_1, \dots, \phi_q) \\ \phi_1' &= P_1(\phi_T, \phi_1, \dots, \phi_q) \\ &\vdots \\ \phi_q' &= P_q(\phi_T, \phi_1, \dots, \phi_q) \end{cases}$$

where $P_T, P_1, \dots, P_q$ are homogeneous quadratic polynomials.
Remarking that the monomial Φ_T^2 only appears in the first equation, standard arguments of analysis lead to:

$$\phi_T(u) \overset{u \longrightarrow u_0}{\sim} \frac{c}{(u_0 - u)},$$

but this is only true on the real line !

Study of the differential system

$$\begin{cases} \phi'_T &= P_T(\phi_T, \phi_1, \dots, \phi_q) \\ \phi'_1 &= P_1(\phi_T, \phi_1, \dots, \phi_q) \\ &\vdots \\ \phi'_q &= P_q(\phi_T, \phi_1, \dots, \phi_q) \end{cases}$$

where $P_T, P_1, \dots, P_q$ are homogeneous quadratic polynomials.
Remarking that the monomial Φ_T^2 only appears in the first equation, standard arguments of analysis lead to:

$$\phi_T(u) \overset{u \longrightarrow u_0}{\sim} \frac{c}{(u_0 - u)},$$

but this is only true on the real line !

Let us apply a tauberian theorem...

Concluding the proof

$$\begin{cases} \sum_{i=1}^{n} p_{i,k}(True)u_0^i = \sum_{i=1}^{n} p_{i,k}(False)u_0^i \sim cn \\ \sum_{i=1}^{n} p_{i,k}(f)u_0^i = o(n) \text{ for all } f \notin \{True, False\}. \end{cases}$$

We easily show that $c = \frac{1}{2}$ and $u_0 = 1$.
Looking back at the induction formula over the coefficients $p_{n,k}(True)$, we get that:

$$p_{n+1,k}(True) \geq \frac{1}{n+1} \sum_{i=1}^{n} p_{n,i}(False) \xrightarrow{n\to\infty} \frac{1}{2}.$$

Concluding the proof

$$\begin{cases} \sum_{i=1}^{n} p_{i,k}(True)u_0^i = \sum_{i=1}^{n} p_{i,k}(False)u_0^i \sim cn \\ \sum_{i=1}^{n} p_{i,k}(f)u_0^i = o(n) \text{ for all } f \notin \{True, False\}. \end{cases}$$

We easily show that $c = \frac{1}{2}$ and $u_0 = 1$.
Looking back at the induction formula over the coefficients $p_{n,k}(True)$, we get that:

$$p_{n+1,k}(True) \geq \frac{1}{n+1} \sum_{i=1}^{n} p_{n,i}(False) \xrightarrow{n\to\infty} \frac{1}{2}.$$

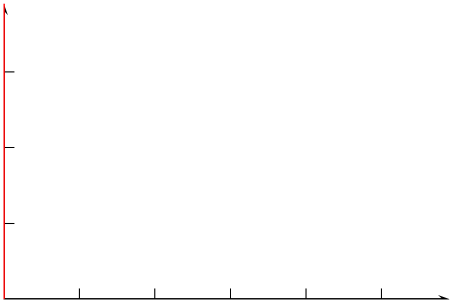
Theorem:

$p_{n,k} \longrightarrow \frac{1}{2}\delta_{True} + \frac{1}{2}\delta_{False}$ when $n \longrightarrow +\infty$.

Study of the growing tree law. The $\wedge/\vee$ labelling.

Probabilistic approach via Yule trees.

Yule trees



Labelled Yule tree

A Yule tree is a continuous time process $(\mathcal{Y}_t)_{t\geq 0}$ of binary trees.

Labelled Yule tree

A Yule tree is a continuous time process $(\mathcal{Y}_t)_{t\geq 0}$ of binary trees.

We label the process:

- as soon as a leaf appears, it chooses a label uniformly at random among literals.
- as soon as a leaf becomes an internal node, it chooses a label uniformly at random among $\wedge$ and $\vee$.

Labelled Yule tree

A Yule tree is a continuous time process $(\mathcal{Y}_t)_{t\geq 0}$ of binary trees.

We label the process:

- as soon as a leaf appears, it chooses a label uniformly at random among literals.
- as soon as a leaf becomes an internal node, it chooses a label uniformly at random among $\wedge$ and $\vee$.

$\rightarrow$ We thus have $\mathcal{E}_t$, a continuous time process of labelled binary trees. For all $t \geq 0$, this process define a law over $\mathcal{E}_k$ which image by Φ over $\mathcal{F}_k$ by P_t.

$\rightarrow$ $P_t = p_{n(t),k}$ where $n(t)$ is the size of $\mathcal{E}_t$.

In a Yule tree, the right and left subtrees at each node are **independent**.

In a Yule tree, the right and left subtrees at each node are **independent**.

Strategy of proof

- We have to show that only constant functions are weighted by P_t.
- Let us consider $P_t^{10} := P_t(f(a) = 1 \text{ and } f(b) = 0)$ for fixed a and $b \in \{0,1\}^k$ two assignments of the k variables.
- Let us show that P_t^{10} tends to zero when $t \longrightarrow +\infty$, independently from the choice of a and b.

Let $P_t^{\alpha\beta} = P_t(f(a) = \alpha \text{ and } f(b) = \beta)$ with fixed a and $b \in \{0,1\}^k$.

Let $P_t^{\alpha\beta} = P_t(f(a) = \alpha \text{ and } f(b) = \beta)$ with fixed a and $b \in \{0,1\}^k$.
Conditionning by the time when the root grows, we get:

$$P_t^{10} = \sum_{i=1}^{k} \frac{e^{-t}}{2k} \left(\mathbb{1}_{\{a_i=1 \text{ and } b_i=0\}} + \mathbb{1}_{\{a_i=0 \text{ and } b_i=1\}}\right)$$

$$+\frac{1}{2}\int_0^t \Big(P_{t-s}^{11}P_{t-s}^{10} + P_{t-s}^{10}(P_{t-s}^{11}+P_{t-s}^{01}) + P_{t-s}^{01}(P_{t-s}^{00}+P_{t-s}^{10}) + P_{t-s}^{00}P_{t-s}^{10}\Big)e^{-s}ds$$

Let $P_t^{\alpha\beta} = P_t(f(a) = \alpha \text{ and } f(b) = \beta)$ with fixed a and $b \in \{0,1\}^k$.
Conditionning by the time when the root grows, we get:

$$P_t^{10} = \sum_{i=1}^{k} \frac{e^{-t}}{2k} \left(\mathbb{1}_{\{a_i=1 \text{ and } b_i=0\}} + \mathbb{1}_{\{a_i=0 \text{ and } b_i=1\}} \right)$$

$$+\tfrac{1}{2}\int_0^t \left(P_{t-s}^{11}P_{t-s}^{10} + P_{t-s}^{10}(P_{t-s}^{11}+P_{t-s}^{01}) + P_{t-s}^{01}(P_{t-s}^{00}+P_{t-s}^{10}) + P_{t-s}^{00}P_{t-s}^{10}\right)e^{-s}ds$$

We can simplify this expression and deduce from it that, if $\pi(t) = \mathbb{P}_t^{10}$:

$$\pi' + \pi^2 = 0 \rightarrow \pi(t) = \frac{1}{t+cst}$$

Let $P_t^{\alpha\beta} = P_t(f(a) = \alpha \text{ and } f(b) = \beta)$ with fixed a and $b \in \{0,1\}^k$.
Conditionning by the time when the root grows, we get:

$$P_t^{10} = \sum_{i=1}^{k} \frac{e^{-t}}{2k} \left(\mathbb{1}_{\{a_i=1 \text{ and } b_i=0\}} + \mathbb{1}_{\{a_i=0 \text{ and } b_i=1\}}\right)$$

$$+\frac{1}{2}\int_0^t \left(P_{t-s}^{11}P_{t-s}^{10} + P_{t-s}^{10}(P_{t-s}^{11}+P_{t-s}^{01}) + P_{t-s}^{01}(P_{t-s}^{00}+P_{t-s}^{10}) + P_{t-s}^{00}P_{t-s}^{10}\right)e^{-s}ds$$

We can simplify this expression and deduce from it that, if $\pi(t) = \mathbb{P}_t^{10}$:

$$\pi' + \pi^2 = 0 \rightarrow \pi(t) = \frac{1}{t+cst}$$

Remark: If $n(t)$ is the number of leaves of $\mathcal{E}_t$, then $n(t) \sim e^t$ a.s., thus: "$t \sim \ln n(t)$".

Let $P_t^{\alpha\beta} = P_t(f(a) = \alpha \text{ and } f(b) = \beta)$ with fixed a and $b \in \{0,1\}^k$.
Conditionning by the time when the root grows, we get:

$$P_t^{10} = \sum_{i=1}^{k} \frac{e^{-t}}{2k} \left(\mathbb{1}_{\{a_i=1 \text{ and } b_i=0\}} + \mathbb{1}_{\{a_i=0 \text{ and } b_i=1\}} \right)$$
$$+ \tfrac{1}{2} \int_0^t \left(P_{t-s}^{11} P_{t-s}^{10} + P_{t-s}^{10} (P_{t-s}^{11} + P_{t-s}^{01}) + P_{t-s}^{01} (P_{t-s}^{00} + P_{t-s}^{10}) + P_{t-s}^{00} P_{t-s}^{10} \right) e^{-s} ds$$

We can simplify this expression and deduce from it that, if $\pi(t) = \mathbb{P}_t^{10}$:

$$\pi' + \pi^2 = 0 \rightarrow \pi(t) = \tfrac{1}{t+cst}$$

Remark: If $n(t)$ is the number of leaves of $\mathcal{E}_t$, then $n(t) \sim e^t$ a.s., thus: "$t \sim \ln n(t)$".

Theorem:

$p_{n,k} \longrightarrow p_k = \tfrac{1}{2}\delta_{True} + \tfrac{1}{2}\delta_{False}$ when n tends to $+\infty$ and

$$\|p_{n,k} - p_k\|_\infty = O\left(\frac{1}{\ln n}\right)$$

- Biaising the law over literals: $\mathbb{P}(x_i) = \mathbb{P}(\overline{x_i}) = \nu(x_i) > 0$.
  No changes.

- Biaising the law over literals: $\mathbb{P}(x_i) = \mathbb{P}(\overline{x_i}) = \nu(x_i) > 0$.
  No changes.
- Biasing the law over the connectives: $\mathbb{P}(\wedge) = q$ and $\mathbb{P}(\vee) = 1 - q$.

Theorem:

- If $q > \frac{1}{2}$, then $p_{n,k} \longrightarrow \delta_{False}$.
- If $q < \frac{1}{2}$, then $p_{n,k} \longrightarrow \delta_{True}$.
- If $q = \frac{1}{2}$, then $p_{n,k} \longrightarrow \frac{1}{2}(\delta_{True} + \delta_{False})$.

And the speed of convergence is of order $O\left(\frac{1}{n^{|2q-1|}}\right)$ if $q \neq \frac{1}{2}$ and of order $O\left(\frac{1}{\ln n}\right)$ otherwise.

- Biaising the law over literals: $\mathbb{P}(x_i) = \mathbb{P}(\overline{x_i}) = \nu(x_i) > 0$.
  No changes.
- Biasing the law over the connectives: $\mathbb{P}(\wedge) = q$ and $\mathbb{P}(\vee) = 1 - q$.
- Biasing the law over the connective, with only positive litterals.

Theorem:

- If $q > \frac{1}{2}$, then $p_{n,k} \longrightarrow \delta_{x_1 \wedge \ldots \wedge x_k}$.
- If $q < \frac{1}{2}$, then $p_{n,k} \longrightarrow \delta_{x_1 \vee \ldots \vee x_k}$.
- If $q = \frac{1}{2}$, then $p_{n,k} \longrightarrow p_k$ where p_k only weights 2^{2^k} threshold functions.

The growing tree law has the same behaviour as the law defined from balanced trees!

[Fournier, Gardy, Genitrini. *Balanced And/Or trees and linear threshold functions*]

Conclusion

- State a *meta-theorem* which would explain the link between the saturation level of a random tree and the behaviour of the law induced over $\mathcal{F}_k$.

Conclusion

- State a *meta-theorem* which would explain the link between the saturation level of a random tree and the behaviour of the law induced over $\mathcal{F}_k$.

Thanks for your attention!