Cost functionals for large random trees

Marion Sciauveau
Joint work with J-F. Delmas and J-S. Dhersin

CERMICS (ENPC) and LAGA (Paris 13)

Les probabilités de demain - IHÉS - 11 mai 2017

Introduction

- **Trees** have lot of applications in various fields such as computer science for data structure or in biology for genealogical or phylogenetic trees of extant species.
  Here we will consider the class of **binary trees** (under the Catalan model).

- **Cost functionals** are functions defined on the set of trees and described by a recurrence relation. They allow to represent the cost of many divide-and-conquer algorithms and to study the balance of trees.

Some notations for binary trees

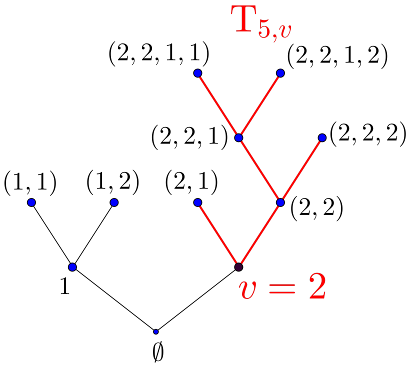


Figure: A binary tree with 5 internal nodes

- T_n rooted full binary ordered tree with n internal nodes
- $\boxed{|\mathrm{T}_n| = 2n+1}$: the cardinal of T_n
- $L(\mathrm{T}_n)$: the left-sub-tree of T_n
- $R(\mathrm{T}_n)$: the right-sub-tree of T_n
- the sub-tree $\mathrm{T}_{n,v}$ of T_n with root v

Random binary trees

A random binary tree is a binary tree selected at random from some probability distribution on binary trees. We often consider two models: **Catalan model** and **Random permutation model**. In what follows, we will only take interest in the Catalan model:

Catalan model: random tree uniformly distributed among the full binary ordered trees with given number of internals nodes. In other words, the probability that a particular tree occurs is $\frac{1}{C_n}$ where C_n is the n^{th} Catalan number: $C_n = \frac{1}{n+1}\binom{2n}{n}$.

1 Introduction

2 Binary trees and Brownian excursion

3 Results for binary trees

4 Conclusion

Definition of cost functionals

Additive functional

A functional F on binary trees is called an additive functional if it satisfies the following recurrence relation:

$$F(\mathrm{T}) = F(L(\mathrm{T})) + F(R(\mathrm{T})) + b_{|\mathrm{T}|}$$

for all trees T such that $|\mathrm{T}| \geq 1$ and with $F(\emptyset) = 0$.
$(b_k, k \geq 1)$ is called the toll function.

Remark:

$$F(\mathrm{T}_n) = \sum_{v \in \mathrm{T}_n} b_{|\mathrm{T}_{n,v}|}$$

Motivation (1)

Goal: study the asymptotics of cost functionals with toll function of type $b_k = k^\beta$ for $\beta > 0$.

Answer: For $\beta > 0$,

$$Z_\beta^{(n)} = \underbrace{|\mathrm{T}_n|^{-(\beta+\frac{1}{2})}}_{\text{scaling factor}} \underbrace{\sum_{v \in \mathrm{T}_n} |\mathrm{T}_{n,v}|^\beta}_{\text{additive functional}} \underset{n\to\infty}{\longrightarrow} 2\, Z_\beta$$

- For $\beta > 0$, **Fill and Kapur** (2003) showed that $Z_\beta^{(n)}$ converges in distribution to Z_β. But Z_β was only characterized by its moments.
- **Fill and Janson** (2007) announced that for $\beta > \frac{1}{2}$, Z_β can be represented as a functional of the normalized Brownian excursion e.

Some examples of additive functionals

1. **Total path length** [Aldous (1991) and Takàcs (1994)]

$$P(\mathrm{T}_n) = \sum_{v \in \mathrm{T}_n} d(\emptyset, v) = \underbrace{\sum_{u \in \mathrm{T}_n} |\mathrm{T}_{n,u}|}_{b_k = k} - |\mathrm{T}_n| \sim |\mathrm{T}_n|^{\frac{3}{2}} Z_1^{(n)}$$

$$\underbrace{|\mathrm{T}_n|^{-\frac{3}{2}}}_{\text{scaling factor}} \quad P(\mathrm{T}_n) \xrightarrow[n\to\infty]{a.s.} 2\, Z_1 = 2 \int_0^1 e(s) ds$$

2. **Wiener index** [Janson (2003) and Chassaing (2004)]

$$W(\mathrm{T}_n) = \sum_{u,v \in \mathrm{T}_n} d(u,v) = 2|\mathrm{T}_n| \underbrace{\sum_{w \in \mathrm{T}_n} |\mathrm{T}_{n,w}|}_{b_k = k} - 2 \underbrace{\sum_{w \in \mathrm{T}_n} |\mathrm{T}_{n,w}|^2}_{b_k = k^2}$$

$$\sim 2|\mathrm{T}_n|^{\frac{5}{2}} \left(Z_1^{(n)} - Z_2^{(n)} \right)$$

$$\underbrace{|\mathrm{T}_n|^{-\frac{5}{2}}}_{\text{scaling factor}} \quad W(\mathrm{T}_n) \xrightarrow[n\to\infty]{a.s.} 4\,(Z_1 - Z_2)$$

Motivation (2)

- We study

$$\underbrace{|\mathrm{T}_n|^{-\frac{3}{2}}}_{\text{scaling factor}} \quad \underbrace{\sum_{v \in \mathrm{T}_n} |\mathrm{T}_{n,v}| f\left(\frac{|\mathrm{T}_{n,v}|}{|\mathrm{T}_n|}\right)}_{\text{unnormalized additive functional}}$$

  for f satisfying smooth conditions.

- **Aim**: derive an invariance principle for such tree functionals.

- **Model**: Catalan model

1 Introduction

2 Binary trees and Brownian excursion

3 Results for binary trees

4 Conclusion

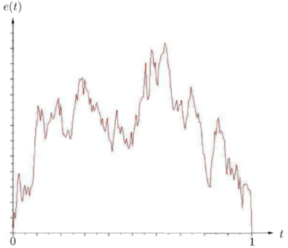
Brownian tree associated to the normalized Brownian excursion

Let e be a normalized Brownian excursion on $[0,1]$ i.e. a standard Brownian motion on $[0,1]$ conditioned on being nonnegative on $[0,1]$ and on taking the value 0 at 1.

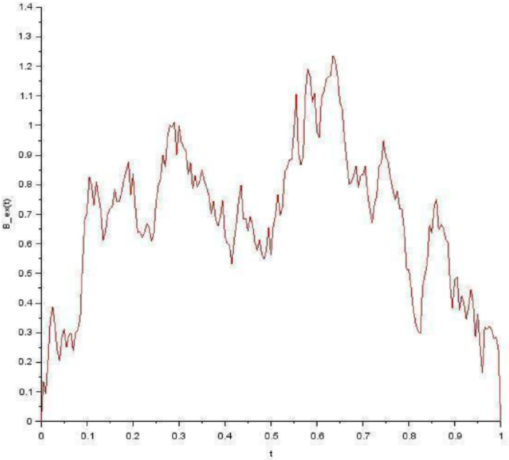
For $s,t \in [0,1]$, $s<t$, we define

$$d_e(s,t) = e(s) + e(t) - 2 \inf_{s<u<t} e(u)\cdot$$

The Browian tree is defined as $\mathcal{T}_e = [0,1]/\sim_e$ where $s \sim_e t \Leftrightarrow d_e(s,t) = 0$ and we still denote by d_e, the induced distance on the quotient. We denote by $\mathbf{p}$ the canonical projection from $[0,1]$ to $\mathcal{T}_e$.

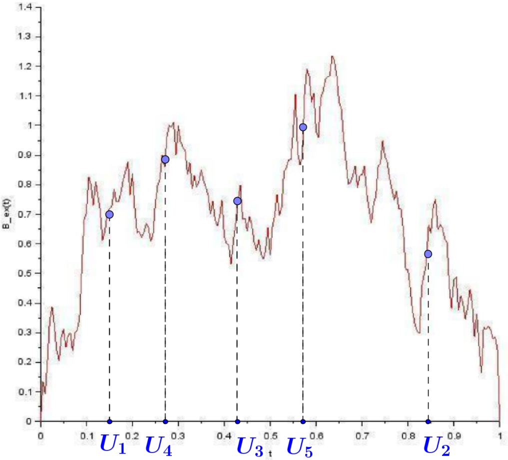


Natural embedding of binary trees into the Brownian excursion e (1)



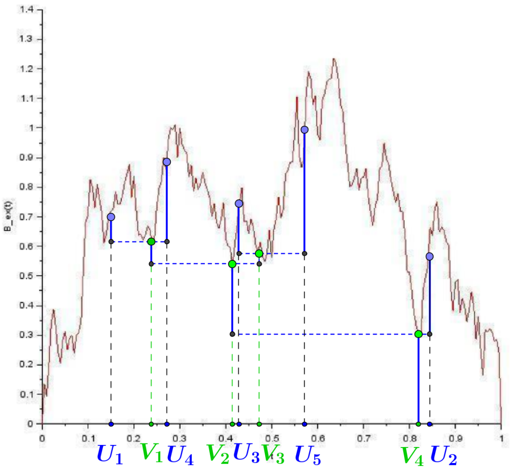
- Normalized Brownian excursion: e

Natural embedding of binary trees into the Brownian excursion e (1)



- Normalized Brownian excursion: e
- $(U_i)_{1 \leq i \leq 5}$ i.i.d. uniform on $[0, 1]$ and indep. of e

Natural embedding of binary trees into the Brownian excursion e (1)



- Normalized Brownian excursion: e
- $(U_i)_{1\leq i\leq 5}$ i.i.d. uniform on $[0,1]$ and indep. of e
- $(V_i)_{1\leq i\leq 4}$ such that

$$e(V_i) = \min_{u\in[U_{(i)},U_{(i+1)}]} e(u)$$

Natural embedding of binary trees into the Brownian excursion e (2)

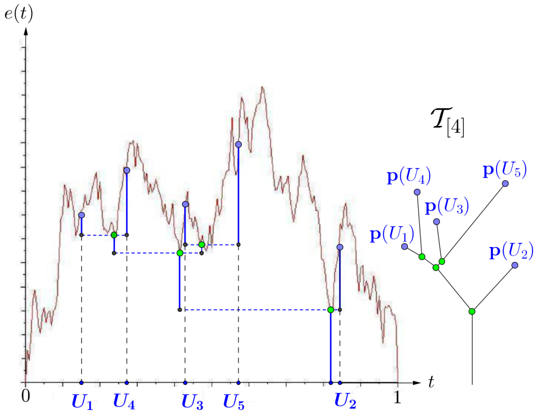


Figure: The Brownian excursion and $\mathcal{T}_{[n]}$ for $n = 4$

Natural embedding of binary trees into the Brownian excursion e (2)

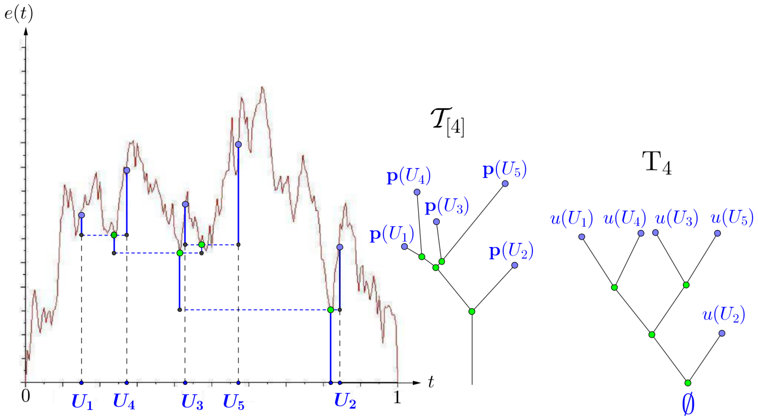


Figure: The Brownian excursion, $\mathcal{T}_{[n]}$ (for $n = 4$) and T_n

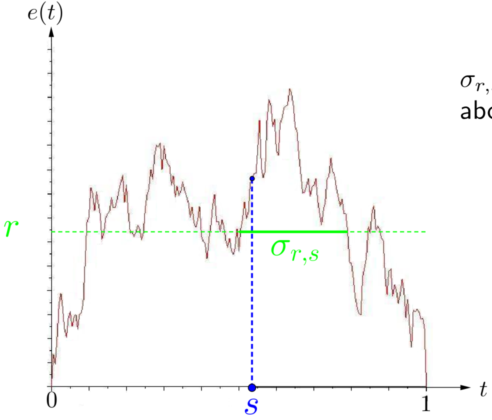
1 Introduction

2 Binary trees and Brownian excursion

3 Results for binary trees

4 Conclusion

Length of a subexcursion



$\sigma_{r,s}$ = length of the excursion of e above level r straddling s

$$\sigma_{r,s} = \int_0^1 dt\, \mathbf{1}_{\{\min_e(s,t) \geq r\}}$$

Invariance principle

Let

$$A_n(f) = |\mathrm{T}_n|^{-\frac{3}{2}} \sum_{v \in \mathrm{T}_n} |\mathrm{T}_{n,v}| f\left(\frac{|\mathrm{T}_{n,v}|}{|\mathrm{T}_n|}\right)$$

and

$$\Phi_e(f) = \int_0^1 ds \int_0^{e_s} dr \; f(\sigma_{r,s})$$

Theorem

A.s., $\forall f \in \mathcal{C}((0,1])$ *s.t.* $\lim_{x \downarrow 0^+} x^a f(x) = 0$ *for some* $0 \le a < \frac{1}{2}$*, we have:*

$$\lim_{n \to +\infty} A_n(f) = 2\,\Phi_e(f)$$

Application for $f(x) = x^{\beta-1}$

For$\beta > 0$ and $n \in \mathbb{N}^*$, we set:

$$Z_\beta = \int_0^1 ds \int_0^{e_s} dr\, \sigma_{r,s}^{\beta-1} \quad \text{and} \quad Z_\beta^{(n)} = |\mathrm{T}_n|^{-(\beta+\frac{1}{2})} \sum_{v\in\mathrm{T}_n} |\mathrm{T}_{n,v}|^\beta$$

Theorem

We have a.s., $\forall \beta > 0$,

$$\lim_{n\to+\infty} Z_\beta^{(n)} = 2\, Z_\beta$$

Application for $f(x) = x^{\beta-1}$

For$\beta > 0$ and $n \in \mathbb{N}^*$, we set:

$$Z_\beta = \int_0^1 ds \int_0^{e_s} dr\, \sigma_{r,s}^{\beta-1} \quad \text{and} \quad Z_\beta^{(n)} = |\mathrm{T}_n|^{-(\beta+\frac{1}{2})} \sum_{v \in \mathrm{T}_n} |\mathrm{T}_{n,v}|^\beta$$

Theorem

We have a.s., $\forall \beta > 0$,

$$\lim_{n \to +\infty} Z_\beta^{(n)} = 2\, Z_\beta$$

Lemma

If $\beta > \frac{1}{2}$,

$$\text{a.s. } Z_\beta < +\infty \text{ and } \mathbb{E}[Z_\beta] < +\infty$$

Otherwise,

$$\text{a.s. } Z_\beta = +\infty$$

Fluctuations of the invariance principle

Theorem

Let $f \in \mathcal{C}([0,1])$ *be locally Lipschitz continuous on* $(0,1]$ *with* $\|x^a f^{'}\|_{essup} < +\infty$ *for some* $a \in (0,1)$*. We have*

$$\Big(\underbrace{|T_n|^{1/4}}_{speed\ of\ CV} (A_n - 2\Phi_e)(f),\, A_n \Big) \xrightarrow[n\to\infty]{\mathcal{L}} \left(\sqrt{2}\sqrt{\Phi_e(xf^2)}\, G, 2\Phi_e \right),$$

where $G \sim \mathcal{N}(0,1)$ *and is independent of the excursion* e*.*

Results and ongoing work

- **Results**:
  - invariance principle for more general additive functionals and for two classes of trees: the binary trees under the Catalan model and some simply generated trees.
  - recover some classical results on additive functional (e.g. total size, total path length ...)
  - fluctuations coming from the approximation of the branch lengths by their mean in the binary case
- **Ongoing work**: study asymetric cost functionals depending on the cardinal of the left and right sub-tree of each nodes.

Thank you for your attention !