Partial match queries: a limit process

Nicolas Broutin Ralph Neininger Henning Sulzbach

Background/Introduction

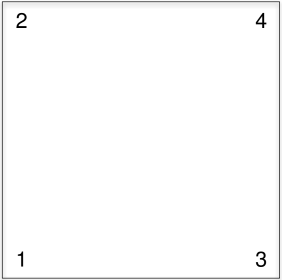
Data structures/Algorithms

- Analysis of costs/running times in natural conditions
- expected cost
- performance guarantee provided by concentration

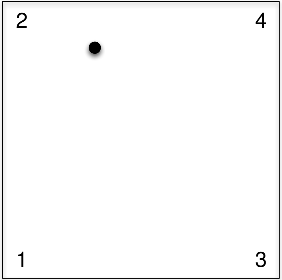
Methodology

- complex “objects” that decompose recursively (tree like, or related)
- general approach for convergence using contractions

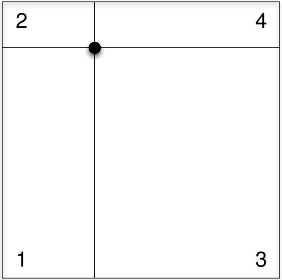
Searching geometric data and quadtrees

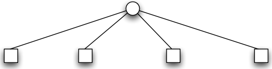


Searching geometric data and quadtrees

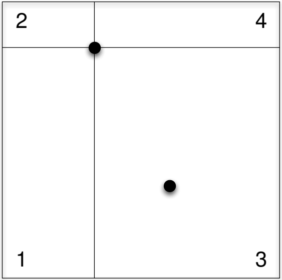


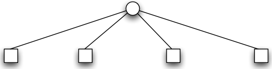
Searching geometric data and quadtrees



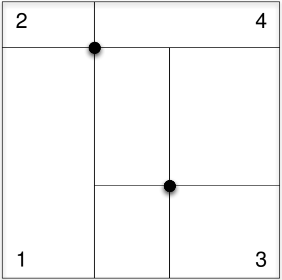


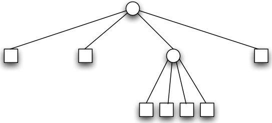
Searching geometric data and quadtrees



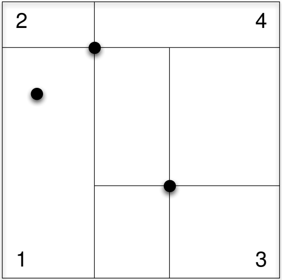


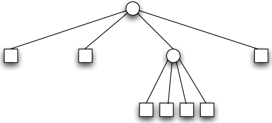
Searching geometric data and quadtrees



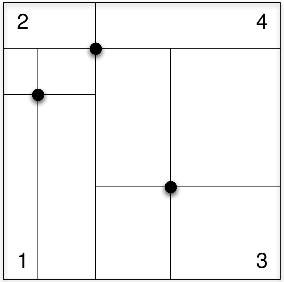


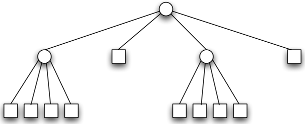
Searching geometric data and quadtrees



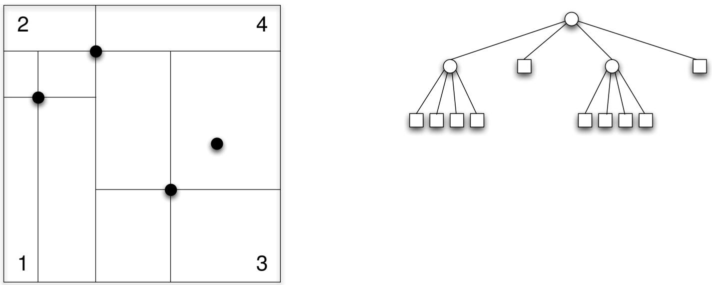


Searching geometric data and quadtrees

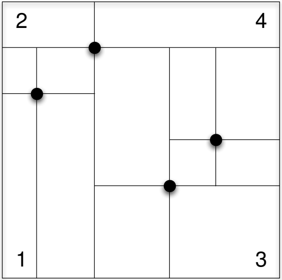


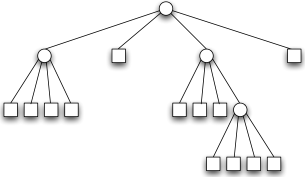


Searching geometric data and quadtrees

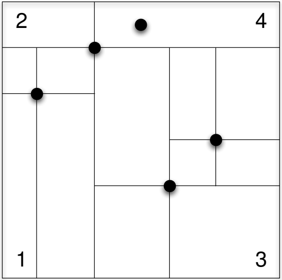


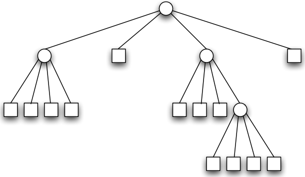
Searching geometric data and quadtrees



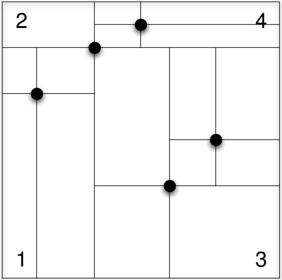


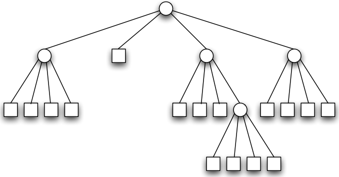
Searching geometric data and quadtrees



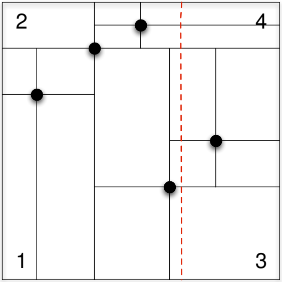


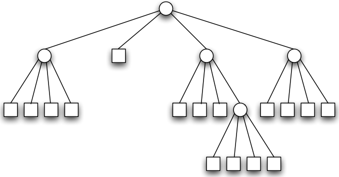
Searching geometric data and quadtrees



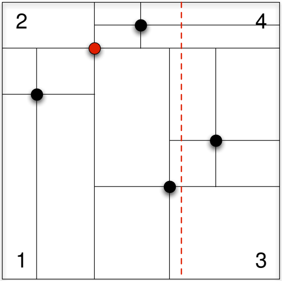


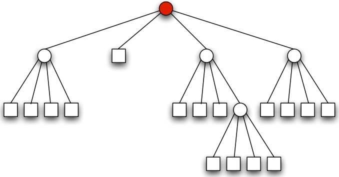
Searching geometric data and quadtrees



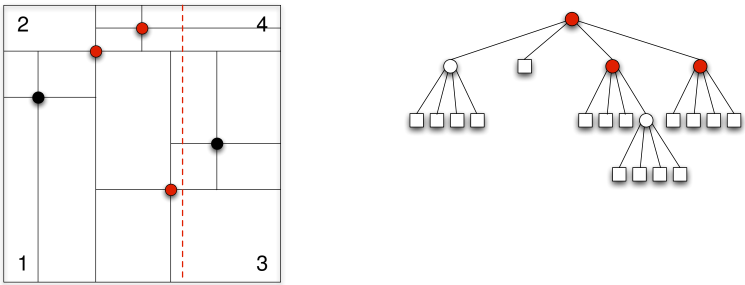


Searching geometric data and quadtrees

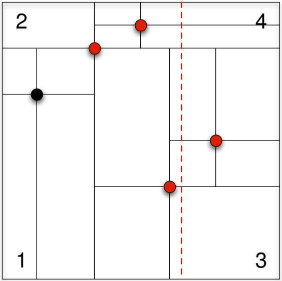


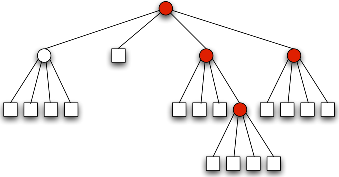


Searching geometric data and quadtrees



Searching geometric data and quadtrees





Model and Previous results

Point set = $\{(U_i, V_i), i \geq 1\}$ iid uniform in $[0,1]^2$

$C_n(s)$ the number of lines intersecting $\{x = s\}$ in a quadtree of size n

Theorem (Flajolet, Gonnet, Puech and Robson (1993))

For ξ uniform independent of $\{(U_i, V_i), i \geq 1\}$

$$\mathbf{E}\left[C_n(\xi)\right] \sim \kappa n^{\beta} \qquad \text{where} \quad \kappa = \frac{\Gamma(2\beta+2)}{2\Gamma(\beta+1)^2}, \quad \beta = \frac{\sqrt{17}-3}{2}$$

Theorem (Chern and Hwang (2003))

Let $\phi(z) = (z+1)(z+2) - 4$ and $\beta > \beta'$ the roots of ϕ. For ξ uniform independent of $\{(U_i, V_i), i \geq 1\}$, one has the exact expression

$$\mathbf{E}\left[C_n(\xi)\right] = \sum_{1 \leq k \leq n} \binom{n}{k} (-1)^{k+1} \frac{2(1-\beta)_{k-1}(1-\beta')_{k-1}}{k!(k+1)!}$$

Corollary (Chern and Hwang (2003))

For ξ uniform independent of $\{(U_i, V_i), i \geq 1\}$

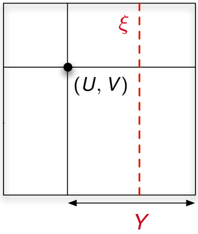
$$\mathbf{E}\left[C_n(\xi)\right] = \kappa n^{\beta} - 1 + O(n^{\beta-1})$$

Idea of the method / heuristic for the constants

Recursive decomposition

We have $Y \overset{d}{=} \max\{U_1, U_2\}$ and
$(I, J) = \text{Mult}(\text{Bin}(n-1, Y); V, (1-V))$ then

$$C_n(\xi) \overset{d}{=} 1 + C_I(\xi') + C_J(\xi')$$



$$\Rightarrow \quad \mathbf{E}[C_n(\xi)] \approx 2\mathbf{E}[C_{nYV}(\xi')]$$

Plugging $\mathbf{E}[C_n(\xi)] = \kappa n^\beta$ yields

$$1 = 2\mathbf{E}[Y^\beta V^\beta] = 2\mathbf{E}[Y^\beta] \cdot \mathbf{E}[V^\beta] = \frac{4}{(\beta+2)(\beta+1)} \quad \Rightarrow \quad \beta = \frac{\sqrt{17}-3}{2}$$

About the variance $\mathbf{Var}\,(C_n(\xi))$

Even when conditioning on the first point, the two terms are still dependent on the query line

The cost at a fixed query line

Idea:

- if the query line is fixed at $s \in (0,1)$, then we do have independence
- however, its relative position changes in the subproblems
- $\Rightarrow$ consider the entire process $(C_n(s), s \in (0,1))$

Theorem (Flajolet, Labelle, Laforest and Salvy 1995)

$$\mathbf{E}\left[C_n(0)\right] = \Theta(n^{\sqrt{2}-1}) = o(n^\beta)$$

Note: in particular, $\mathbf{E}[C_n(U_1)] = o(n^\beta)$, and $C_n(s)$ is not concentrated.

Theorem (Curien and Joseph (2011))

For every fixed $s \in (0,1)$, one has

$$\mathbf{E}\left[C_n(s)\right] \sim K_1 (s(1-s))^{\beta/2} n^\beta, \qquad K_1 = \frac{\Gamma(2\beta+2)\Gamma(\beta+2)}{2\Gamma(\beta+1)^3\Gamma(\beta/2+1)^2}.$$

Main result

Theorem

There exists a random continuous function Z such that, as $n \to \infty$,

$$\left(\frac{C_n(s)}{K_1 n^{\beta}}, s \in [0,1] \right) \xrightarrow{d} (Z(s), s \in [0,1]). \qquad (1)$$

This convergence in distribution holds in the Banach space $(\mathcal{D}[0,1], \|\cdot\|)$ of right-continuous functions with left limits (càdlàg) equipped with the supremum norm.

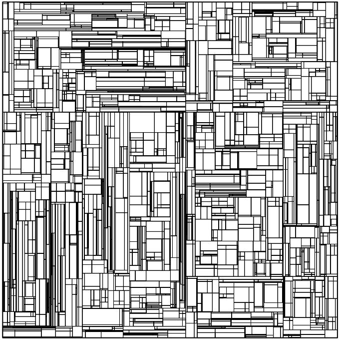
Proposition

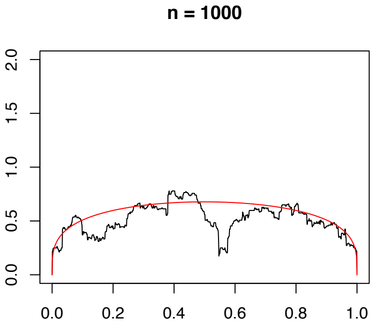
The distribution of the random function Z in (1) is a fixed point of the following equation

$$Z(s) \stackrel{d}{=} \mathbf{1}_{\{s<U\}} \left[(UV)^{\beta} Z^{(1)} \left(\frac{s}{U} \right) + (U(1-V))^{\beta} Z^{(2)} \left(\frac{s}{U} \right) \right] + \mathbf{1}_{\{s \geq U\}} \left[((1-U)V)^{\beta} Z^{(3)} \left(\frac{s-U}{1-U} \right) + ((1-U)(1-V))^{\beta} Z^{(4)} \left(\frac{s-U}{1-U} \right) \right],$$

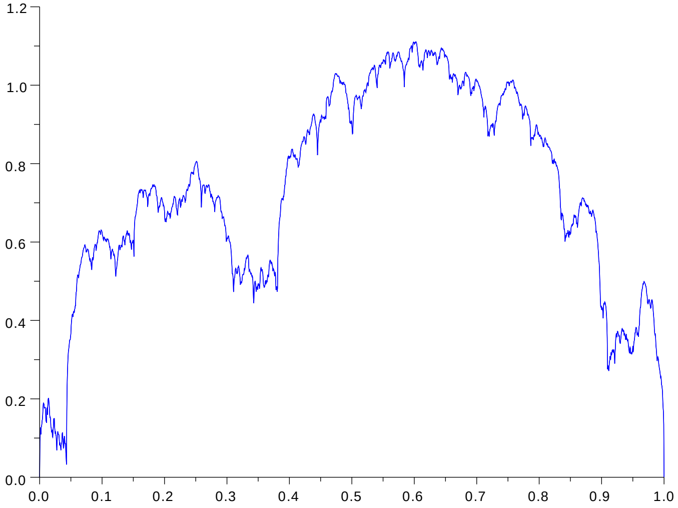
where U and V are independent $[0,1]$-uniform random variables and $Z^{(i)}$, $i = 1, \dots, 4$ are independent copies of the process Z, which are also independent of U and V. Furthermore, Z in (1) is the only solution such that $\mathbf{E}[Z(s)] = (s(1-s))^{\beta/2}$ for all $s \in [0,1]$ and $\mathbf{E}[\|Z\|^2] < \infty$.

What does it look like I





What does it look like II



Moments and supremum

Theorem

We have for all $s \in (0,1)$*, as* $n \to \infty$*,*

$$\mathbf{Var}\,(C_n(s)) \sim \left(2\mathrm{B}(\beta+1,\beta+1)\frac{2\beta+1}{3(1-\beta)} - 1\right)(s(1-s))^{\beta} n^{2\beta}.$$

Here, $B(a,b) := \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$ *denotes the Eulerian beta integral* ($a, b > 0$)*.*

Theorem

Let $S_n = \sup_{s\in[0,1]} C_n(s)$*. Then, as* $n \to \infty$*,*

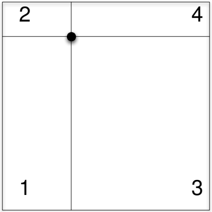
$$n^{-\beta} S_n \xrightarrow{d} S = \sup_{s\in[0,1]} Z(s) \qquad \textit{and} \qquad \mathbf{E}[S_n] \sim n^{\beta}\mathbf{E}[S], \qquad \mathbf{Var}(S_n) \sim n^{2\beta}\mathbf{Var}(S).$$

Convergence in distribution by contraction I.

Cost of the construction of the quadtree / path length

$P_n = \sum_{i=1}^{n} D_i$ with D_i the depth of the i-th inserted point

- I_n^r the number of points inside the r-th child cell
- Q^r the volume or the r-th child cell



We have

$$P_n \stackrel{d}{=} \sum_{r=1}^{4} P_{I_n^r} + n - 1 \quad \text{and write} \quad X_n = \frac{P_n - \alpha n \log n}{n}$$

$$(I_n^1, \ldots, I_n^4) \stackrel{d}{=} \mathrm{Mult}(n-1; UV, U(1-V), (1-U)(1-V), (1-U)V).$$

Shifting and rescaling we obtain:

$$\underbrace{\frac{P_n - \alpha n \log n}{n}}_{X_n} = \sum_{r=1}^{4} \underbrace{\left(\frac{I_n^r}{n}\right)}_{A_n^r} \frac{P_{I_n^r} - \alpha I_n^r \log I_n^r}{I_n^r} + \underbrace{\frac{n-1}{n} - \frac{\alpha \log n}{n} + \alpha \sum_{r=1}^{4} \left(\frac{I_n^r}{n}\right) \log\left(\frac{I_n^r}{n}\right)}_{b_n}$$

Convergence in distribution by contraction II.

General problem:

A recursive family of equations $X_n \overset{d}{=} \sum_{r=1}^{4} A_n^r \cdot X_{I_n^r}^r + b_n$ with

- $(A_n^1, \ldots, A_n^4, I_n^1, \ldots, I_n^4, b_n)$ independent of $((X^1), \ldots, (X^4))$
- $(X_n^r, n \geq 1)$ iid copies of (X)

The equation "converges" to a limit equation:

$$A_n^r = \frac{I_n^r}{n} \to \mathrm{Leb}(Q_r)$$

$$b_n = \frac{n-1}{n} - \frac{\alpha \log n}{n} + \alpha \sum_{r=1}^{4} \left(\frac{I_n^r}{n}\right) \log\left(\frac{I_n^r}{n}\right) \to 1 + \alpha \sum_{r=1}^{4} \mathrm{Leb}(Q_r) \log \mathrm{Leb}(Q_r)$$

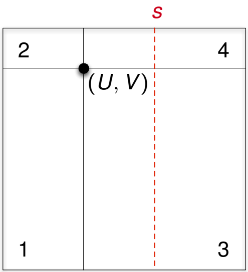
$$X \overset{d}{=} \sum_{r=1}^{4} \mathrm{Leb}(Q_r) \cdot X^r + 1 + \alpha \sum_{r=1}^{4} \mathrm{Leb}(Q_r) \log \mathrm{Leb}(Q_r) \qquad (2)$$

Formalization: (2) a transfer map on a **space of probability measures on** $\mathbb{R}$.

$d_2(\phi, \varphi) = \inf\{\|X - Y\|_2 : \mathscr{L}(X) = \phi, \mathscr{L}(Y) = \varphi\}$

- on $\mathscr{M}_2 = \{\text{probability measures } \mu : \int x^2 d\mu < \infty\}$ no contraction (can shift!)
- on $\mathscr{M}_2^0 = \{\mu \in \mathscr{M}_2 : \int x d\mu = 0\}$ contraction

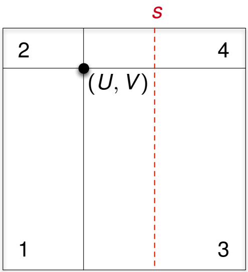
Convergence for partial match processes



$$(I_n^{(1)}, \dots, I_n^{(4)}) \stackrel{d}{=} \mathrm{Mult}(n-1; UV, U(1-V), (1-U)(1-V), (1-U)V)$$

$$C_n(s) \stackrel{d}{=} 1 + \mathbf{1}_{\{s<U\}} \left[C^{(1)}_{I_n^{(1)}}\left(\frac{s}{U}\right) + C^{(2)}_{I_n^{(2)}}\left(\frac{s}{U}\right) \right] + \mathbf{1}_{\{s \geq U\}} \left[C^{(3)}_{I_n^{(3)}}\left(\frac{1-s}{1-U}\right) + C^{(4)}_{I_n^{(4)}}\left(\frac{1-s}{1-U}\right) \right]$$

Convergence for partial match processes



$$(I_n^{(1)}, \dots, I_n^{(4)}) \stackrel{d}{=} \mathrm{Mult}(n-1; UV, U(1-V), (1-U)(1-V), (1-U)V)$$

$$C_n(s) \stackrel{d}{=} 1 + \mathbf{1}_{\{s<U\}} \left[C^{(1)}_{I_n^{(1)}}\left(\frac{s}{U}\right) + C^{(2)}_{I_n^{(2)}}\left(\frac{s}{U}\right) \right] + \mathbf{1}_{\{s\ge U\}} \left[C^{(3)}_{I_n^{(3)}}\left(\frac{1-s}{1-U}\right) + C^{(4)}_{I_n^{(4)}}\left(\frac{1-s}{1-U}\right) \right]$$

Heuristic: If $n^{-\beta} C_n(\cdot)$ converges, we should have $n^{-\beta} C_n(\cdot) \to Z(\cdot)$ satisfying

$$Z(s) \stackrel{d}{=} \mathbf{1}_{\{s<U\}} \left[(UV)^\beta Z^{(1)}\left(\frac{s}{U}\right) + (U(1-V))^\beta Z^{(2)}\left(\frac{s}{U}\right) \right] + \mathbf{1}_{\{s\ge U\}} \left[((1-U)V)^\beta Z^{(3)}\left(\frac{s-U}{1-U}\right) + ((1-U)(1-V))^\beta Z^{(4)}\left(\frac{s-U}{1-U}\right) \right]$$

Convergence in $\mathcal{D}[0,1]$ by contraction arguments I.

Neininger and Sulzbach (2011+)

Let (X_n) be $\mathcal{D}[0,1]$-valued random variables with

$$X_n \stackrel{d}{=} \sum_{r=1}^{K} A_n^{(r)} \circ X^{(r)}_{I_n^{(r)}} + b_n, \quad n \geq 1,$$

where

- $(A_n^{(1)}, \ldots, A_n^{(K)})$ are random linear and continuous operators on $\mathcal{D}[0,1]$
- b_n is a $\mathcal{D}[0,1]$-valued random variable
- $I_n^{(1)}, \ldots, I_n^{(K)}$ are random integers between 0 and $n-1$
- $(X_n^{(1)}), \ldots, (X_n^{(K)})$ are distributed like (X_n)
- $(A_n^{(1)}, \ldots, A_n^{(K)}, b_n, I_n^{(1)}, \ldots, I_n^{(K)}), (X_n^{(1)}), \ldots, (X_n^{(K)})$ are independent

Convergence in $\mathcal{D}[0,1]$ by contraction arguments I.

Neininger and Sulzbach (2011+)

Let (X_n) be $\mathcal{D}[0,1]$-valued random variables with

$$X_n \stackrel{d}{=} \sum_{r=1}^{K} A_n^{(r)} \circ X^{(r)}_{I_n^{(r)}} + b_n, \quad n \geq 1,$$

where

- $(A_n^{(1)}, \dots, A_n^{(K)})$ are random linear and continuous operators on $\mathcal{D}[0,1]$
- b_n is a $\mathcal{D}[0,1]$-valued random variable
- $I_n^{(1)}, \dots, I_n^{(K)}$ are random integers between 0 and $n-1$
- $(X_n^{(1)}), \dots, (X_n^{(K)})$ are distributed like (X_n)
- $(A_n^{(1)}, \dots, A_n^{(K)}, b_n, I_n^{(1)}, \dots, I_n^{(K)}), (X_n^{(1)}), \dots, (X_n^{(K)})$ are independent

Example: here, because of the rescaling, we have

$$A_n^{(1)} : f \mapsto \mathbf{1}_{\{\,\cdot\, \leq U\}} \left(\frac{I_n^{(1)}}{n}\right)^{\beta} f\left(\frac{\cdot}{U}\right)$$

Convergence in $\mathcal{D}[0,1]$ by contraction arguments I.

Neininger and Sulzbach (2011+)

Let (X_n) be $\mathcal{D}[0,1]$-valued random variables with

$$X_n \stackrel{d}{=} \sum_{r=1}^{K} A_n^{(r)} \circ X^{(r)}_{I_n^{(r)}} + b_n, \quad n \geq 1,$$

where

- $(A_n^{(1)}, \dots, A_n^{(K)})$ are random linear and continuous operators on $\mathcal{D}[0,1]$
- b_n is a $\mathcal{D}[0,1]$-valued random variable
- $I_n^{(1)}, \dots, I_n^{(K)}$ are random integers between 0 and $n-1$
- $(X_n^{(1)}), \dots, (X_n^{(K)})$ are distributed like (X_n)
- $(A_n^{(1)}, \dots, A_n^{(K)}, b_n, I_n^{(1)}, \dots, I_n^{(K)}), (X_n^{(1)}), \dots, (X_n^{(K)})$ are independent

Example: here, because of the rescaling, we have

$$A_n^{(1)} : f \mapsto \mathbf{1}_{\{\,\cdot\, \leq U\}} \left(\frac{I_n^{(1)}}{n}\right)^{\beta} f\left(\frac{\cdot}{U}\right)$$

Keep in mind

We want contraction in a **space of probability measures on $\mathcal{D}[0,1]$**.

Convergence in $\mathcal{D}[0,1]$ by contraction arguments II.

Neininger and Sulzbach (2011+)

For a random linear operator A write

$$\|A\|_2 := \mathbf{E}[\|A\|_{\mathrm{op}}^2]^{1/2} \quad \text{with} \quad \|A\|_{\mathrm{op}} := \sup_{\|x\|=1} \|A(x)\|$$

(A1) CONVERGENCE AND CONTRACTION (SIMPLIFIED).

- we have $\|A_n^{(r)}\|_2, \|b_n\|_2 < \infty$ for all $r = 1, \ldots, K$ and $n \geq 0$
- there exist random operators $A^{(1)}, \ldots, A^{(K)}$ on $\mathcal{D}[0,1]$ and a $\mathcal{D}[0,1]$-valued random variable b such that

$$\|b_n - b\|_2 + \sum_{r=1}^{K} \left(\|A_n^{(r)} - A^{(r)}\|_2 \right) \leq R(n) \qquad R(n) \to 0$$

- for all $\ell \in \mathbb{N}$,

$$L^* = \limsup_{n\to\infty} \mathbf{E}\left[\sum_{r=1}^{K} \|A_r^{(n)}\|_{\mathrm{op}}^2 \right] < 1.$$

(A2) EXISTENCE AND EQUALITY OF MOMENTS. $\mathbf{E}[\|X_n\|^2] < \infty$ for all n and $\mathbf{E}[X_{n_1}(t)] = \mathbf{E}[X_{n_2}(t)]$ for all $n_1, n_2 \in \mathbb{N}_0, t \in [0,1]$.

Convergence in $\mathcal{D}[0,1]$ by contraction arguments III.

Neininger and Sulzbach (2011+)

(A3) EXISTENCE OF A CONTINUOUS SOLUTION. There exists a solution X of the fixed-point equation

$$X \stackrel{d}{=} \sum_{r=1}^{K} A_r \circ X^{(r)} + b$$

with continuous paths, $\mathbf{E}[\|X\|^2] < \infty$ and $\mathbf{E}[X(t)] = \mathbf{E}[X_1(t)]$ for all $t \in [0,1]$.

(A4) PERTURBATION CONDITION. $X_n = W_n + h_n$ where $\|h_n - h\| \to 0$ with $h \in \mathcal{D}[0,1]$ and random variables W_n in $\mathcal{D}[0,1]$ such that there exists a sequence (r_n) with, as $n \to \infty$,

$$\mathbf{P}\left(W_n \notin \mathcal{D}_{r_n}[0,1]\right) \to 0.$$

Here, $\mathcal{D}_{r_n}[0,1] \subset \mathcal{D}[0,1]$ denotes the set of functions on the unit interval, for which there is a decomposition of $[0,1]$ into intervals of length as least r_n on which they are constant.

(A5) RATE OF CONVERGENCE. $R(n) = o\left(\log^{-m}(1/r_n)\right)$.

Existence of a continuous solution

Define

- a complete tree $T = \bigcup_{n\geq 0}\{1,2,3,4\}^n$ with (U_u, V_u), $u \in \mathcal{T}$, iid uniform on $[0,1]$
- a starting function $h(s) = (s(1-s))^{\beta/2}$
- an iteration/mixing operator $G : [0,1]^2 \times \mathcal{C}[0,1]^4 \to \mathcal{C}[0,1]$

$$
\begin{aligned}
&G(x, y; f_1, f_2, f_3, f_4)(s) = \\
&\mathbf{1}_{\{s<x\}} \left[(xy)^\beta f_1\left(\frac{s}{x}\right) + (x(1-y))^\beta f_2\left(\frac{s}{x}\right)\right] \\
&+ \mathbf{1}_{\{s\geq x\}} \left[((1-x)y)^\beta f_3\left(\frac{s-x}{1-x}\right) + ((1-x)(1-y))^\beta f_4\left(\frac{s-x}{1-x}\right)\right]
\end{aligned}
$$

For every node $u \in \mathcal{T}$, let

$$
\begin{aligned}
Z_0^u &= h \\
Z_{n+1}^u &= G(U_u, V_u; Z_n^{u1}, Z_n^{u2}, Z_n^{u3}, Z_n^{u4})
\end{aligned}
$$

Lemma

$Z_n = Z_n^\varnothing$, $n \geq 0$, is a non-negative martingale

Existence of a continuous solution

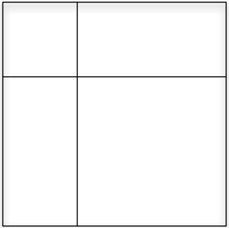
Define

- a complete tree $T = \bigcup_{n\geq 0}\{1,2,3,4\}^n$ with (U_u, V_u), $u \in \mathcal{T}$, iid uniform on $[0,1]$
- a starting function $h(s) = (s(1-s))^{\beta/2}$
- an iteration/mixing operator $G : [0,1]^2 \times \mathcal{C}[0,1]^4 \to \mathcal{C}[0,1]$

$$G(x,y;f_1,f_2,f_3,f_4)(s) =$$
$$\mathbf{1}_{\{s<x\}}\left[(xy)^\beta f_1\left(\frac{s}{x}\right) + (x(1-y))^\beta f_2\left(\frac{s}{x}\right)\right]$$
$$+\,\mathbf{1}_{\{s\geq x\}}\left[((1-x)y)^\beta f_3\left(\frac{s-x}{1-x}\right) + ((1-x)(1-y))^\beta f_4\left(\frac{s-x}{1-x}\right)\right]$$

For every node $u \in \mathcal{T}$, let

$$Z_0^u = h$$
$$Z_{n+1}^u = G(U_u, V_u; Z_n^{u1}, Z_n^{u2}, Z_n^{u3}, Z_n^{u4})$$



Lemma

$Z_n = Z_n^\varnothing$, $n \geq 0$, is a non-negative martingale

Existence of a continuous solution

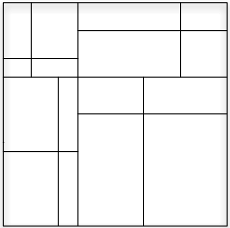
Define

- a complete tree $T = \bigcup_{n\geq 0}\{1,2,3,4\}^n$ with (U_u, V_u), $u \in \mathcal{T}$, iid uniform on $[0,1]$
- a starting function $h(s) = (s(1-s))^{\beta/2}$
- an iteration/mixing operator $G : [0,1]^2 \times \mathcal{C}[0,1]^4 \to \mathcal{C}[0,1]$

$$G(x,y;f_1,f_2,f_3,f_4)(s) =$$
$$\mathbf{1}_{\{s<x\}}\left[(xy)^\beta f_1\left(\frac{s}{x}\right) + (x(1-y))^\beta f_2\left(\frac{s}{x}\right)\right]$$
$$+\mathbf{1}_{\{s\geq x\}}\left[((1-x)y)^\beta f_3\left(\frac{s-x}{1-x}\right) + ((1-x)(1-y))^\beta f_4\left(\frac{s-x}{1-x}\right)\right]$$

For every node $u \in \mathcal{T}$, let

$$Z_0^u = h$$
$$Z_{n+1}^u = G(U_u, V_u; Z_n^{u1}, Z_n^{u2}, Z_n^{u3}, Z_n^{u4})$$



Lemma

$Z_n = Z_n^\varnothing$, $n \geq 0$, is a non-negative martingale

Uniform convergence of the mean

Proposition

There exists $\varepsilon > 0$ *such that*

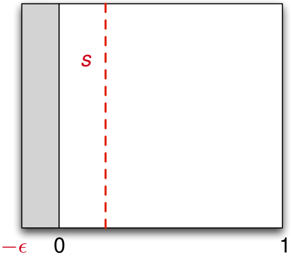
$$\sup_{s\in[0,1]} |t^{-\beta}\mathbf{E}[P_t(s)] - \mu_1(s)| = O(t^{-\varepsilon}).$$

$$\sup_{s\in[0,1]} |t^{-\beta}\mathbf{E}[P_t(s)] - \mu_1(s)| \le \sup_{s\le\delta} \left|t^{-\beta}\mathbf{E}[P_t(s)] - \mu_1(s)\right| + \sup_{s\in(\delta,1/2]} \left|t^{-\beta}\mathbf{E}[P_t(s)] - \mu_1(s)\right|.$$

Proposition (Almost monotonicity)

For any $s < 1/2$ *and* $\varepsilon \in [0, 1-2s)$, *we have*

$$\mathbf{E}[P_t(s)] \le \mathbf{E}\left[P_{t(1+\varepsilon)}\left(\frac{s+\varepsilon}{1+\varepsilon}\right)\right].$$



Thank you!