A toy model of Pólya tree ensemble: smoothing and adaptation

Thibault Randrianarisoa
Supervised by Ismaël Castillo

Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation

Mathematical and Statistical Challenges in Uncertainty Quantification

July 15, 2020

Context: Nonparametric estimation

- **Goal**: Estimate $f \in \mathcal{F}$ (ordinarily a functional space), an infinite dimensional parameter.
  Ex: Regression function, density, c.d.f., etc.

- In regression, CART decision trees (Breiman, 1984) and their ensemble methods, i.e. forests (Breiman, 2001), are a popular class of estimators.

- Single trees for L^2-loss have already been extensively studied, e.g. Donoho (1997), Blanchard, Schäfer & Rozenholc (2004), Gey & Nedelec (2005).

- More recent focus on forest estimators, e.g. Scornet (2016), Scornet, Biau & Vert (2015).

Bayesian tree methods

- Bayesian tree and Bayesian forest algorithms:
  - Bayesian CART (Chipman, George & McCulloch (1998), Denison, Mallick & Smith (1998)), BART (Chipman, George & McCulloch (2010)) in regression.
  - Pólya tree prior (Ferguson (1972-3-4), Mauldin, Sudderth & Williams (1992), Lavine (1992)) in density estimation.

- The work on the theoretical understanding of Bayesian trees (Castillo (2017), Castillo and Ročková (2019)) and forests (Linero and Yang (2018), Ročková and van der Pas (2019)) is just starting.

- Recent interest in Pólya trees and related constructions (Hjort and Walker (2009), Wong and Ma (2010), Nieto-Barajas and Müller (2012), Castillo and Mismer (2019)).

Problem

Tree algorithms build piecewise constant functions on a partition of the sample space.

$\Longrightarrow$ Often sub-optimal convergence rate on smooth functional classes

Arlot and Genuer (2014) develop a toy model where random forests do better than single trees in such situation.

In today's talk: Can we extend their ideas to obtain such smoothing effect with Bayesian forest methods?

Table of contents

Few results on original Breiman random forests. Most results:

- focuses on a particular part of the algorithm.
- make strong assumptions on the parameter to be inferred.
- modify the algorithm: e.g. **Purely random forests** (as seen in this part).

Let us now first present the ideas of Arlot and Genuer (2014) on trees aggregation in the Gaussian white noise model.

Model

1. Gaussian white noise model:

$$dY^{(n)}(t) = f(t)dt + \frac{dW(t)}{\sqrt{n}}, \quad t \in [0;1]$$

with $f \in L^2[0;1]$ and $W(t)$ a standard Brownian motion.

2. Tree estimator: Let $\mathbb{U} \sim \mathcal{U}$ be a random partition of $[0;1]$

$$\widehat{f}\left(x; \mathbb{U}, Y^{(n)}\right) = \sum_{\lambda \in \mathbb{U}} \frac{\mathbb{1}_\lambda(x)}{|\lambda|} \int_\lambda dY^{(n)}(t) \in S_{\mathbb{U}}$$

with $S_{\mathbb{U}}$ the linear space of functions which are constant over each $\lambda \in \mathbb{U}$.

$$\widetilde{f}(x; \mathbb{U}) := \sum_{\lambda \in \mathbb{U}} \frac{\mathbb{1}_\lambda(x)}{|\lambda|} \int_\lambda f(t)dt = \arg\min_{s \in S_{\mathbb{U}}} \|f - s\|_2$$

Forest estimator

Given the family of partitions $\mathbb{V}_q = \{\mathbb{U}_i; 1 \leq i \leq q\}$, $\mathbb{U}_i \overset{i.i.d.}{\sim} \mathcal{U}$,

$$\widehat{f}\left(x; \mathbb{V}_q, Y^{(n)}\right) := \frac{1}{q} \sum_{i=1}^{q} \widehat{f}\left(x; \mathbb{U}_i, Y^{(n)}\right) \qquad \text{(forest estimator)}$$

$$\widetilde{f}\left(x; \mathbb{V}_q\right) := \frac{1}{q} \sum_{i=1}^{q} \widetilde{f}\left(x; \mathbb{U}_i\right) \qquad \text{(Ideal forest)}$$

Single Tree vs. Infinite Forest [Arlot & Genuer, 2014]

Toy model $\mathbb{U} \sim \mathcal{U}_{toy}$: for $k \in \mathbb{N}^*$ and $T \sim \mathcal{U}[0,1)$,

$$\mathbb{U} = \left[0, \frac{1-T}{k}\right), \dots, \left[\frac{i-T}{k}, \frac{i+1-T}{k}\right), \dots, \left[\frac{k-T}{k}, 1\right)$$

For f twice continuously differentiable:

1. MISE of the infinite forest $\widehat{f}_\infty(x; Y^{(n)}) := \lim_{q \to +\infty} \widehat{f}(x; \mathbb{V}_q, Y^{(n)})$

$$\inf_{1/\epsilon \leq k \leq n} \int_\epsilon^{1-\epsilon} \mathbb{E}\left[(\widehat{f}_\infty(x; Y^{(n)}) - f(x))^2\right] dx = \mathcal{O}(n^{-4/5})$$

2. Single tree MISE:

$$\inf_{1/\epsilon \leq k \leq n} \int_\epsilon^{1-\epsilon} \mathbb{E}\left[(\widehat{f}(x; \mathbb{U}, Y^{(n)}) - f(x))^2\right] dx \gtrsim n^{-2/3}$$

$\rightarrow$ Up to C^2 regularity, the forest estimator attains optimal rates of convergence (but not the tree estimator!).

Density estimation

Model: $X^{(n)} \sim \mathbb{P}_f^{\otimes n}$, with $f = \frac{d\mathbb{P}_f}{d\lambda}$, a density w.r.t. to the Lebesgue measure λ and supported on $I = [0; 1)$.

From a prior Π on the space

$$\mathcal{F} := \left\{ f : I \mapsto \mathbb{R} \quad \middle| \quad f \geq 0, \int f d\lambda = 1 \right\}$$

we define the posterior distribution $\Pi[\cdot|X^{(n)}]$.

Frequentist analysis of the Bayesian methods: Assume $X^{(n)} \sim \mathbb{P}_{f_0}^{\otimes n}$, how does the posterior behave (asymptotically)?

Common assumption: Hölder regularity

$$\Sigma(\alpha, K, I) = \left\{ f : I \mapsto \mathbb{R} \quad \middle| \quad \|f\|_{C^\alpha} := \sup_{x \neq y} \frac{|f^{(\lfloor \alpha \rfloor)}(x) - f^{(\lfloor \alpha \rfloor)}(y)|}{|x - y|^{\alpha - \lfloor \alpha \rfloor}} \leq K \right\}$$

Tree-based prior

Let's write $\mathcal{E}^* = \bigcup_{l \geq 0} \{0; 1\}^l$. Consider a sequence of partitions of I:

$$\mathcal{T}_0 = \{I_\varnothing = I\},\ \mathcal{T}_1 = \{I_0, I_1\}, \mathcal{T}_2 = \{I_{00}, I_{01}, I_{10}, I_{11}\} \dots$$

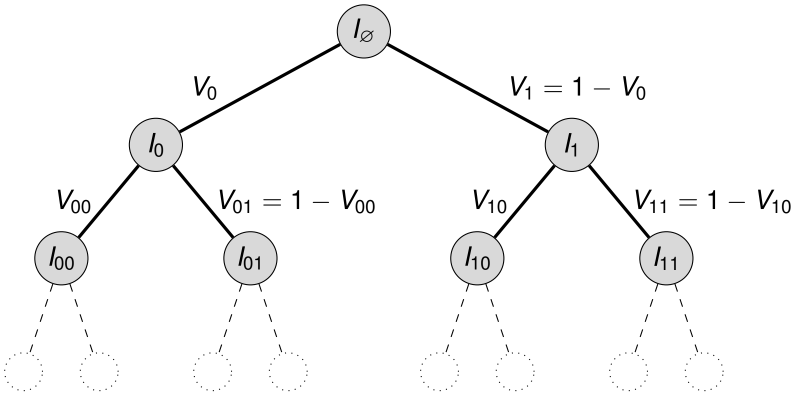
such that $I_\epsilon = I_{\epsilon 0} \bigcup I_{\epsilon 1}$ and the set $\{I_\epsilon | \epsilon \in \mathcal{E}^*\}$ generates the Borel σ-field.

Also, define the random variables $V_\epsilon \in [0; 1]$ and for $\epsilon = \epsilon_1 \dots \epsilon_l$ (i.e. $|\epsilon| = l$),

$$P(I_\epsilon) = \prod_{i=1}^{l} V_{\epsilon_1 \dots \epsilon_i}$$

P extends to a probability measure on Borelians under mild conditions on V_ϵ's.

Tree-based prior



Pólya tree prior

Definition

A random probability measure P is said to follow a Pólya tree process $PT(\mathcal{A}, \{\mathcal{T}_i\})$ with parameters $\mathcal{A} = \{a_\epsilon | \epsilon \in \mathcal{E}^*\}$ on the sequence $\{\mathcal{T}_i\}$ of partitions if the r.v.'s $V_{\epsilon 0}$, for $\epsilon \in \mathcal{E}^*$, are independent, $V_{\epsilon 0} \sim Beta(a_{\epsilon 0}, a_{\epsilon 1})$ and $V_{\epsilon 1} = 1 - V_{\epsilon 0}$.

Popular prior with nice properties:

- With good parameters $\mathcal{A}$, it is a prior on densities with good asymptotic properties (see Barron, Schervish & Wasserman (1999), Lavine (1992) for consistency, Castillo (2017) for rates of convergence).
- Conjugate prior

N.B.: It is customary to take $a_\epsilon = \widetilde{a}_{|\epsilon|}$.

Truncated Pólya tree

Simplified prior on densities: $TPT_L\left(\mathcal{A}\right)$

Take the sequence of partitions given by

$$\mathcal{T}_l = \{I_{lk} := [k2^{-l}; (k+1)2^{-l}), 0 \leq k \leq 2^l - 1\}$$

Note: for any $\epsilon \in \{0; 1\}^l$, the sequence can be seen as the expression in base 2^{-1} of some dyadic number $k2^{-l} = \sum_{i=1}^{l} \epsilon_i 2^{-i}$: we can then identify $I_\epsilon = I_{lk}$.

We stop the process at depth L and associate to a draw of V_ϵ's the distribution that evenly spreads its mass inside the elements of $\mathcal{T}_L$.

Induced distribution on densities:

$$f \sim TPT_L\left(\mathcal{A}\right) \implies \forall x \in [0; 1), f(x) = \sum_{|\epsilon|=L} 2^L \mathbb{1}_{I_\epsilon}(x) \prod_{i=1}^{L} V_{\epsilon_1 \ldots \epsilon_i}$$

Truncated Pólya tree

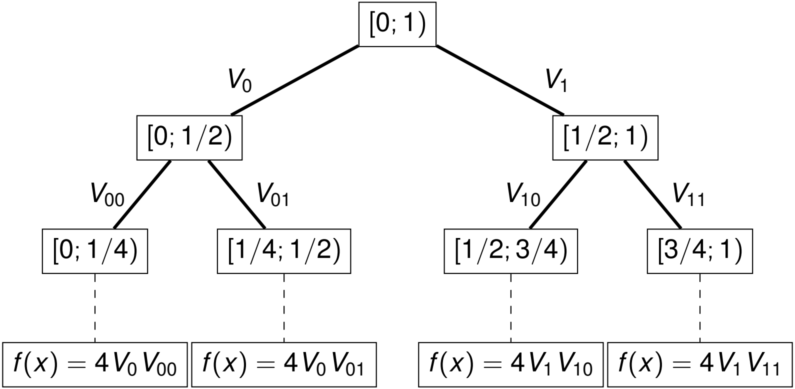


Figure: Truncated Pólya Tree at depth $L = 2$

Contraction rate

$$B_{KL}(f_0, \epsilon) := \left\{ f \in \mathcal{F} \quad | \quad KL(f_0; f) \vee V(f_0; f) \leq \epsilon^2 \right\}$$

Theorem (Ghosal, Ghosh, van der Vaart, 2000)

Suppose that there exists a sequence $(\epsilon_n)_{n \geq 0}$ *and subsets* $\mathcal{F}_n$ *verifying* $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$ *and*

1. $\Pi[B_{KL}(f_0, \epsilon_n)] \geq e^{-cn\epsilon_n^2}$;
2. $\log N(\epsilon_n, \mathcal{F}_n, d) \leq Dn\epsilon_n^2$ *(bound on the metric entropy);*
3. $\Pi[\mathcal{F}_n^c] \leq e^{-(c+4)n\epsilon_n^2}$.

for some $c > 0$, $D > 0$. *Then, for a constant* M *sufficiently large, as* $n \to \infty$,

$$\mathbb{E}_{f_0} \Pi[d(f_0, f) > M\epsilon_n | X^{(n)}] \to 0$$

with d *the Hellinger or* $L1$ *distance.*

Contraction rate: one tree

Theorem 0 (Fixed regularity)

Let $f_0 \in \Sigma(\alpha, K, [0,1))$, $0 < \alpha \leq 1$ *and* $f_0 \geq \rho$ *for some* $\rho > 0$. *Also, let* Π_n *be the* $TPT_{L_n}(\mathcal{A})$ *distribution with* $2^{L_n} \asymp \left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}}$ *and for some* $b > 0$,

$$\forall\, \widetilde{a}_{|\epsilon|} \in \mathcal{A},\; b \leq \widetilde{a}_{|\epsilon|} \leq 2^{2\alpha|\epsilon|}$$

If we endow f *with the* Π_n *prior, then, for* M *large enough, as* $n \to \infty$,

$$\mathbb{E}_{f_0}\Pi_n\left[h(f_0, f) > M\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}} \middle| X^{(n)}\right] \to 0$$

Remark: An adaptive version also exists, with the addition of a prior on the depth L of the tree.

Forest of Pólya trees

Modified aggregation scheme from Arlot et al. (2014):

The prior distribution $FPT_{L,q}(\mathcal{A})$ is the image measure of $TPT_L(\mathcal{A})$ by

$$\begin{aligned} \phi_{L,q} \colon \mathcal{F} &\to \mathcal{F} \\ f &\mapsto \tilde{f}_q^L := \frac{1}{q} \sum_{i=0}^{q-1} f\left(\cdot - \frac{i}{q} 2^{-L}\right) \end{aligned}$$

with the shift being congruent modulo 1.

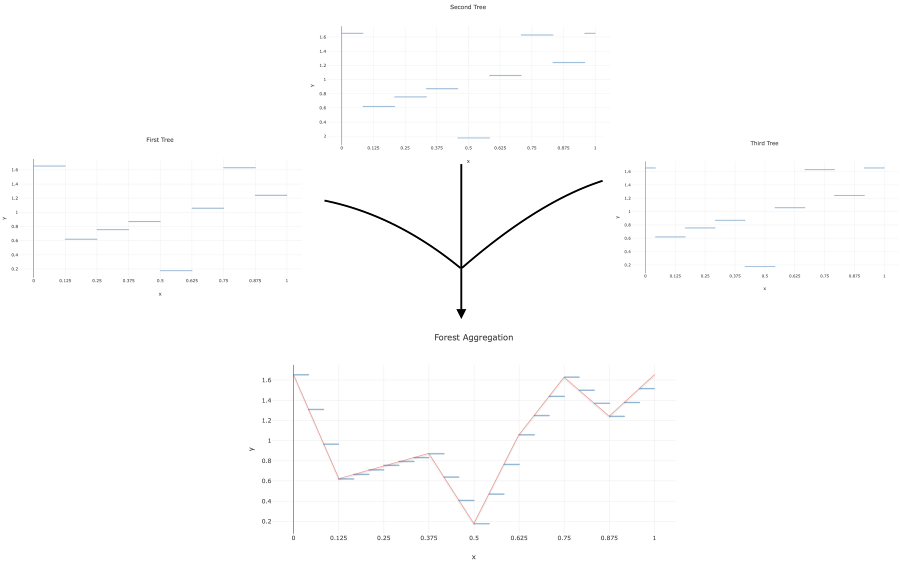


Figure: The $\phi_{3,3}$ operation. The function in red is $\tilde{f}^L_\infty$, obtained with $q \to +\infty$.

Contraction rate

$$\Sigma_p(\alpha, K, [0,1)) := \left\{ f\big|_{[0;1)} \middle| f \text{ 1-periodic }, f \in \Sigma(\alpha, K, \mathbb{R}) \right\}$$

Theorem 1 (Higher regularities)

Let $f_0 \in \Sigma_p(\alpha, K, [0,1))$, $0 < \alpha \leq 2$ and $f_0 \geq \rho$ for some $\rho > 0$.
Also let $\Pi_n = FPT_{L_n, q_n}(\mathcal{A})$ be the prior on f with the same conditions on $\mathcal{A}$ as before and such that

- $2^{L_n} \asymp \left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}}$
- $q_n \geq 2^{\alpha L_n}$

Then, for M large enough, as $n \to \infty$,

$$\mathbb{E}_{f_0} \Pi_n \left[h(f_0, f) > M \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}} \middle| X^{(n)} \right] \to 0$$

Also, adaptation is possible via appropriate priors on L, q

Various aggregations

For $f\colon \mathbb{R} \longrightarrow \mathbb{R}$, one can iterate the aggregation operation

- a 1-step discrete aggregation:

$$\begin{aligned} f^1_{q,s}\colon \mathbb{R} &\to \mathbb{R} \\ x &\mapsto \frac{1}{q}\sum_{i=0}^{q-1} f\left(x - \frac{is}{q}\right) \end{aligned}$$

- an m-step discrete aggregation:

$$f^{m+1}_{q,s} = \left(f^m_{q,s}\right)^1_{q,s}$$

Higher aggregation prior:

$$f \sim DPA(m, L, q, \mathcal{A}) \iff f = \widetilde{g}^m_{q,2^{-L}}\Big|_{[0;1)}$$

with $\widetilde{g}$ the 1-periodic extension of $g \sim TPT_L\left(\mathcal{A}\right)$

Contraction rate

Theorem 2 (contraction rate for arbitrary, fixed regularities)

*Let's $f_0 \in \Sigma_p(\alpha, K, [0,1))$, $\alpha > 0$ and $f_0 \geq \rho$ for some $\rho > 0$.
Let $\Pi_n = DPA(\lfloor \alpha \rfloor, L_n, 2^{\alpha L_n}, \mathcal{A})$ prior with constant tree parameters $\mathcal{A}$ and L_n as before, then, as $n \to \infty$, for M large enough,*

$$\mathbb{E}_{f_0} \Pi_n \left[h(f_0, f) > M \left(\frac{\log n}{n} \right)^{\frac{\alpha}{2\alpha+1}} \middle| X^{(n)} \right] \to 0$$

Adaptive version

$$\xi(l,n) = \left\lfloor \frac{1}{2}\left[\frac{1}{l}\log_2\left(\frac{n}{\log n}\right) - 1\right]\right\rfloor$$

Theorem 3 (Adaptive version)

Let f_0 and $\mathcal{A}$ be as before. If we endow f with the hierarchical prior

$$l \sim \Pi_L\left[\{l\}\right] \propto 2^{-l2^l}$$
$$f \Big| l \sim DPA(\xi(l,n), l, 2^{\xi(l,n)l}, \mathcal{A})$$

then, as $n \to \infty$, for M large enough,

$$\mathbb{E}_{f_0}\Pi\left[h(f_0,f) > M\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}} \Big| X^{(n)}\right] \to 0$$

Link with Spline functions.

For q large enough, $\widetilde{g}^m_{q,h} \approx h^{-1}\chi^{*m}(\cdot/h) * \widetilde{g}$ with $\chi = \mathbb{1}_{[0;1]}$. Also,

$$h^{-1}\chi^{*m}(\cdot/h) * \left(\sum_{j\in\mathbb{Z}} \theta_j h^{-1} \mathbb{1}_{[jh;(j+1)h[}(\cdot) \right) = \sum_{j\in\mathbb{Z}} \theta_j h^{-1} \chi^{*(m+1)} \left(\frac{\cdot}{h} - j \right)$$

But, $\chi^{*(m+1)}$ and its translation are the cardinal splines of order $m+1$ on the knot sequence $\mathbb{Z}$.

$\implies$ Use of the approximation properties of spline to control the "bias".

If $h^{-1} \in \mathbb{N}^*$ and $(\theta_i)_{i\in\mathbb{Z}}$ is h^{-1}-periodic:

$$\sum_{i=0}^{h^{-1}-1} \theta_i h^{-1} \sum_{p\in\mathbb{Z}} \chi^{*(m+1)} \left(\frac{\cdot}{h} - (j + ph^{-1}) \right) := \sum_{i=0}^{h^{-1}-1} \theta_i S_{i,h^{-1},m+1}$$

Link with Spline functions.

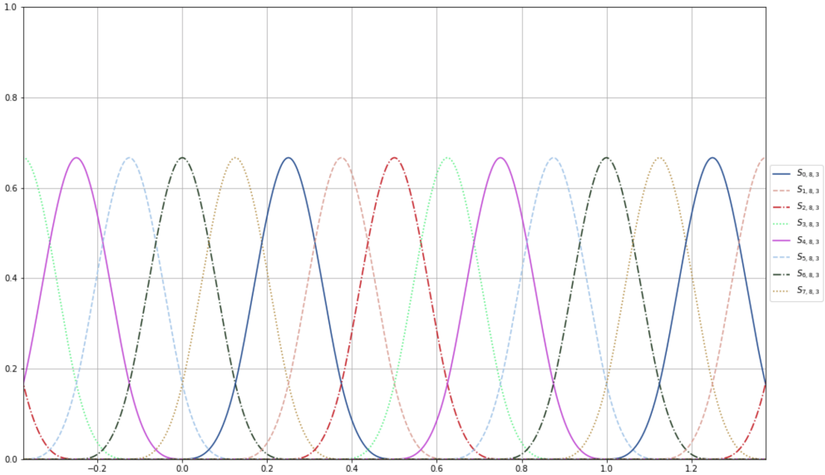


Figure: Periodic rescaled Cardinal splines

Prior behaviour on the edges: handling the boundaries

How to relax the periodicity on f_0 and its behaviour on the edges of $[0;1)$?

Additional treatment of the prior to break the periodicity:

Periodicity of spline functions comes from periodicity of coordinates in the Cardinal splines basis. One needs to "decouple" the coordinates of splines covering the edges of the interval $[0;1)$.

For splines of order m, only $m-1$ of such pairs of coordinates to handle:

$\rightarrow$ we can draw random uniform variables to perform this without increasing the complexity of the prior too much.

A similar theorem as above holds for $f_0 \in \Sigma(\alpha, K, [0;1])$ for modified prior.

Conclusion

Take-home messages:

- Bayesian histogram forest estimators can achieve optimal contraction rate for any Hölder regularity of the true density.
- Such methods are also **adaptive**.

Further work:

- Working on more general constructions (e.g. with a prior on the split points of the partition underlying the Pólya tree distribution).
- Extension to other models (nonparametric regression...)