Pólya tree ensembles: smoothing and adaptation

Thibault Randrianarisoa
Supervised by Pr. Ismaël Castillo

Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation

2021 ISBA World Meeting

June 11, 2021

Manuscript: `https://arxiv.org/abs/2010.12299`

Available in Poster session P03 (Thursday 6:15-8am) [EDT]

Tree methods

- Tree-based methods (CART, random forests, BCART, BART):
  - build a recursive partition of the sample space,
  - estimate the signal locally.
- Theoretical results so far:
  - Minimax optimality can be established on classes of regularity ≤ 1.
  - Ensemble methods (*aka* forests): results up to regularity 2 for frequentists estimates (Arlot and Genuer, 2014; Mourtada et al., 2020).

What about posterior contraction rates of Bayesian forests on classes of regularity > 1?

Posterior contraction rate

- X are the observation, depending on the model, with sample size n.
- $X \sim P^n_{\theta_0}$, $\theta_0 \in \Theta$ the estimand.
- Prior Π on Θ + likelihood $\longrightarrow$ Posterior $\Pi\left[\cdot \middle| X\right]$.

Definition

A posterior contraction rate for a (semi-)metric d on Θ is a sequence $(\epsilon_n)_{n\geq 1}$ such that, for any $M_n \to \infty$,

$$\Pi\left[\theta : \ d(\theta, \theta_0) \geq M_n \epsilon_n \middle| X\right] \overset{P^n_{\theta_0}}{\to} 0.$$

Posterior contraction rate

- X are the observation, depending on the model, with sample size n.
- $X \sim P^n_{\theta_0}$, $\theta_0 \in \Theta$ the estimand.
- Prior Π on Θ + likelihood $\longrightarrow$ Posterior $\Pi\left[\cdot\middle|X\right]$.

Definition

A posterior contraction rate for a (semi-)metric d on Θ is a sequence $(\epsilon_n)_{n\geq 1}$ such that, for any $M_n \to \infty$,

$$\Pi\left[\theta :\ d(\theta,\theta_0) \geq M_n\epsilon_n \middle| X\right] \overset{P^n_{\theta_0}}{\to} 0.$$

Density estimation: $X_i \overset{\text{i.i.d.}}{\sim} P_{f_0}$, $i = 1,\dots,n$, $dP_{f_0}/d\mu = f_0$, $X_i \in \mathcal{X}$.

(Truncated) Pólya Tree prior

Prior TPT $(L, \mathcal{A})$, $\mathcal{A} = \left\{\nu_\epsilon,\ \epsilon \in \cup_{l=0}^{L} \{0;1\}^l\right\}$ on the set of densities $\mathcal{F} = \left\{f : \mathcal{X} \to \mathbb{R}_+, \int f d\mu = 1\right\}$.

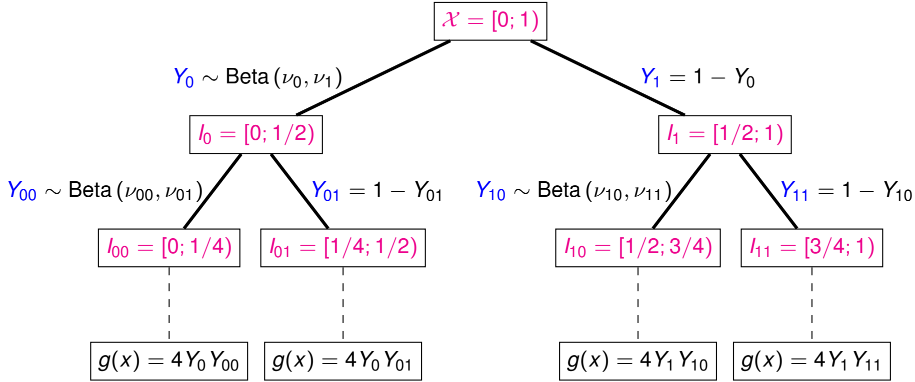


Figure: Truncated Pólya Tree at depth $L = 2$.

Shifted partition

- Focus on the torus $\mathcal{X} = \mathbb{T}$, the addition operation is modulo 1.
- Replace the sets I in the recursive partition with $I + U$, $U > 0$.

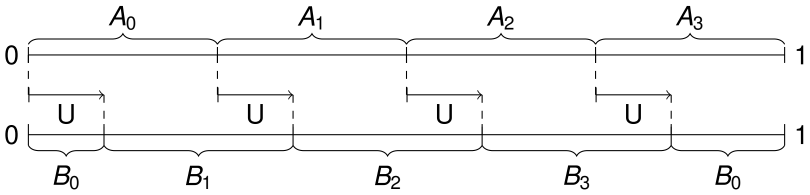


Figure: Shift of a dyadic partition.

In the following, $\mathcal{I}_{L,U}$ is the regular partition with step 2^{-L} shifted by U.

Discrete Pólya aggregate prior

We introduce the Discrete Pólya aggregate prior, denoted $\mathrm{DPA}\,(L, \mathcal{A}, q, m)$, that is the distribution of

$$f = \frac{1}{q^m} \sum_{i=0}^{(q-1)m} \#\left\{0 \leq j_1, \ldots, j_m < q,\ j_1 + \cdots + j_m = i\right\} g_i$$

where the g_i are defined like a TPT on $\mathcal{I}_{L,iq^{-1}2^{-L}}$ and share the same Beta variables along the trees.

⚠ Only the underlying partition changes between trees g_i, the Y variables are the same.

Posterior contraction rate

Let $\Sigma(\alpha, \mathbb{T})$ be the Hölder class of densities on the torus.

Theorem (Contraction rate for arbitrary, fixed regularities)

Let $f_0 \in \Sigma(\alpha, \mathbb{T})$, $\alpha > 0$ and suppose $f_0 \geq \rho$ for some $\rho > 0$.
Under mild conditions on $\mathcal{A}$ and for $2^{L_n} \asymp \left(n \log^{-1} n\right)^{\frac{1}{2\alpha+1}}$, the prior $\Pi_n = DPA(L_n, \mathcal{A}, 2^{\alpha L_n}, \lfloor \alpha \rfloor)$ gives, as $n \to \infty$, M large enough and d the Hellinger or L^1 distance,

$$E_{f_0} \Pi_n \left[d(f_0, f) > M \left(n^{-1} \log n\right)^{\frac{\alpha}{2\alpha+1}} \middle| X \right] \to 0.$$

Adaptive posterior contraction rate

Theorem (Adaptive version)

Under the same assumptions of f_0 and conditions on $\mathcal{A}$ as in the preceding theorem, for some map ϕ, ψ, if we endow f with the hierarchical prior

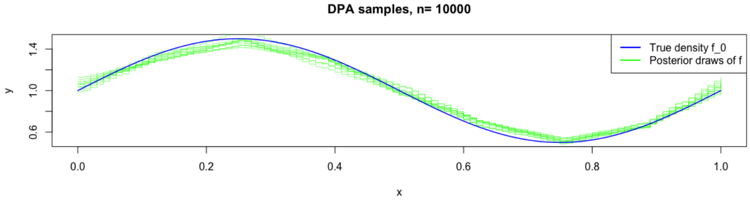
$$l \sim \Pi_L\left[\{l\}\right] \propto 2^{-l2^l}$$
$$f \Big| l \sim DPA(l, \mathcal{A}, \phi(l,n), \psi(l,n)),$$

then, as $n \to \infty$, for M large enough and d the Hellinger or L^1 distance,

$$E_{f_0}\Pi\left[d(f_0, f) > M\left(n^{-1}\log n\right)^{\frac{\alpha}{2\alpha+1}} \Big| X\right] \to 0.$$

Simulation

$f_0 : x \to 1 + sin(2\pi x) \in \Sigma(2.5, \mathbb{T})$.



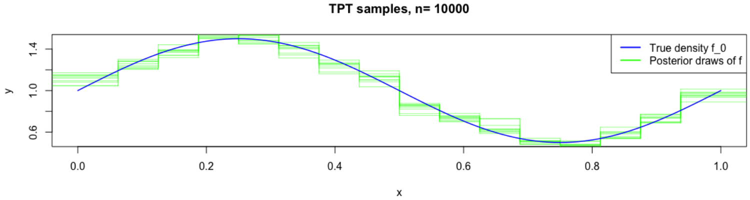


Figure: Posterior draws from fixed-depth version of DPA and TPT, with parameters tuned for regularity 2.5.

Conclusion

- In the paper: handles boundaries of $[0;1)$, instead of $\mathbb{T}$ (addition of a stochastic layer to redefine the samples near the boundary).
- Ongoing work: allowing for random shifts of the partitions and loosening of the almost sure equality of Betas.
- Should extend to higher dimensions and different models (nonparametric regression, etc...).

Take-home messages:

Bayesian histogram forest estimators can achieve :

- optimal contraction rate for any Hölder regularity $\alpha > 0$ of the true density,
- adaptation.