# Bayesian Dyadic Trees and Histograms for Regression

Stéphanie van der Pas and Veronika Ročková
svdpas@math.leidenuniv.nl; veronika.rockova@chicagobooth.edu

## Key insights

For the nonparametric regression problem

$$Y_i = f_0(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0,1), \quad i = 1, \dots, n,$$

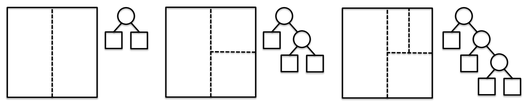
Bayesian dyadic trees and histograms recover piecewise constant regression surfaces $f_0$

- ✓ at (near) optimal speed;
- ✓ while automatically learning the optimal number of leaves;
- ✓ and the optimal locations of the splits;

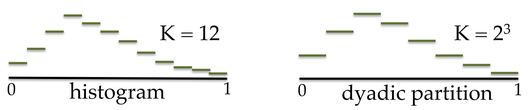
provided that the prior encourages **aggressive pruning**.

## Dyadic trees & histograms

Dyadic trees: nested parallel-axis midpoint splits.



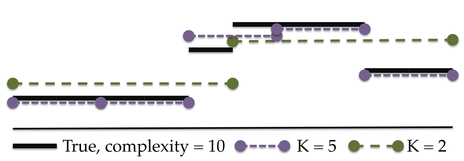
Our results are for a one-dimensional predictor, using histograms, of which dyadic trees are a special case.




Our results can be extended to more flexible, **data-dependent partitioning schemes**.

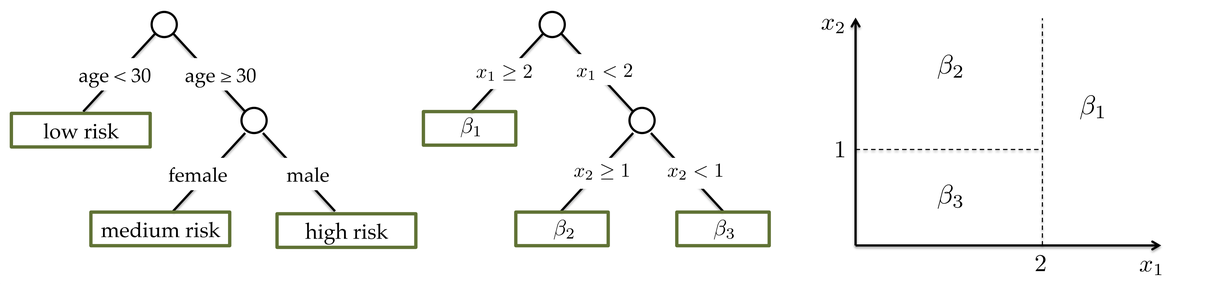
## Complexity $K_{f_0}$ of $f_0$

The **complexity** $K_{f_0}$ of $f_0$ with respect to a class of partitions $\mathbf{\Omega}$ is the smallest size of a partition in $\mathbf{\Omega}$ needed to completely cover $f_0$ without any overlap. Restricted cell count is zero.




Depends on the true number of jumps in $f_0$, as well as the interval lengths.

## Regression trees




## Contribution - posterior concentration

We consider the class $\mathcal{F}$ of step functions supported on equivalent blocks $\{\Omega_k\}_{k=1}^K$ (i.e. all cells have the same number of data points), equipped with independent standard normal priors on the step heights. If

- ▷ the prior on the number of steps $K$ is **at least exponential**; bounded from above by $ae^{-bK}$ for constants $a, b > 0$, recommendation: $\pi_K(k) \propto e^{-ck \log k}$;
- ▷ and if $f_0 : [0,1] \to \mathbb{R}$ given by $f_0 = \sum_{k=1}^{K_0} \beta^0 \mathbf{1}_{\Omega_k^0}$ is a step function with $K_0$ steps, where
  - $K_0$ and the partition $\{\Omega_k^0\}_{k=1}^{K_0}$ may be **unknown**;
  - $\|\beta^0\|_\infty^2 \lesssim \log n$,

then, if the complexity $K_{f_0}$ of $f_0$ increases at most at rate $\sqrt{n}$:

$$\Pi\left(f \in \mathcal{F} : \|f - f_0\|_n \geq M_n n^{-1/2} \sqrt{K_{f_0} \log(n/K_{f_0})} \mid Y^n\right) \to 0$$

in $P_{f_0}^n$-probability, for every $M_n \to \infty$ as $n \to \infty$, meaning that **the posterior distribution concentrates around $f_0$ at rate** $n^{-1/2}\sqrt{K_{f_0} \log(n/K_{f_0})}$. Benchmark rate: $n^{-1/2}\sqrt{K_0 \log(n/K_0)}$ (Gao et al. (2017)).

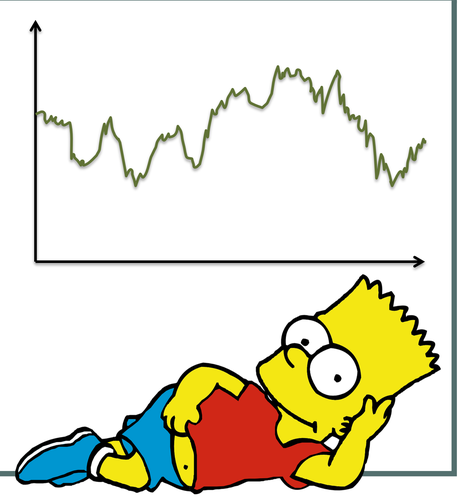
## More good news - extensions

**Q: What if $f_0$ is not a step function?**
✓ The result holds for Hölder-continuous $f_0$, and the smoothness is automatically learned from the data.

**Q: What about high-dimensional predictors and variable selection?**
✓ Automatic dimension reduction is performed if $f_0$ is high-dimensional and depends only on a subset of the features.

**Q: How about sum-of-tree models like BART?**
✓ The result can be extended to ensembles of regression trees like BART.

See **arXiv:1708.08734**, Ročková and Van der Pas, Posterior Concentration for Bayesian Regression Trees and Their Ensembles.

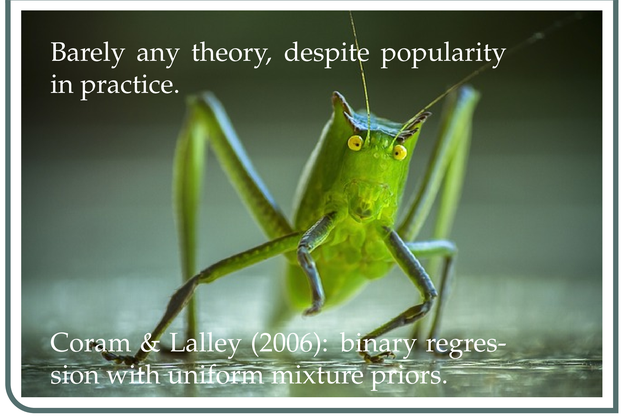


## Applications

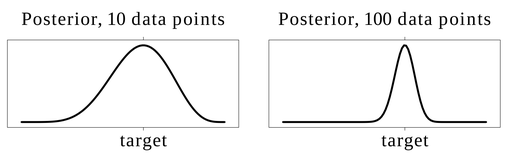
A sampling platter of applications of regression tree methods like CART and BART:

- ▷ phishing detection; Abu-Nimeh et al. (2007)
- ▷ public health: identifying high-risk subgroups; Lemon et al. (2003)
- ▷ imputing missing data; Burgette & Reiter (2010)
- ▷ credit risk modelling. Zhang & Härdle (2010)

## Previous theory




## Posterior concentration




Posterior concentration is a **performance measure** of both location and spread of Bayesian algorithms. Theoretical study often leads to actionable insights such as:

- ▷ **guidelines** for the choice of **tuning parameters**; e.g. topic modelling with the latent Dirichlet allocation model (Tang et al. (2014)).
- ▷ **recommendations** of optimal **priors**;
- ▷ **characterization of problems** for which the method is suitable.