Estimation and Inference in Modern Nonparametric Statistics

William G. Underwood

Final Public Oral Examination | May 7th, 2024

Department of Operations Research and Financial Engineering
Princeton University

Research overview

Some of my recent work consists of

- Inference and estimation with Mondrian random forests
- Uniform inference for dyadic kernel density estimators
- Yurinskii's coupling for martingales

Why random forests?

Why do tree-based models still outperform deep learning on tabular data?

Léo Grinsztajn
Soda, Inria Saclay
leo.grinsztajn@inria.fr

Edouard Oyallon
ISIR, CNRS, Sorbonne University

Gaël Varoquaux
Soda, Inria Saclay

- Ensemble methods for regression and classification, with good performance, flexibility, robustness and efficiency
- Many variants including the popular "Random Forest"
- Estimation theory has developed rapidly in recent years but applicability to statistical inference is less well understood
- In joint work with Matias D. Cattaneo and Jason M. Klusowski, I develop valid feasible inference procedures and minimax optimal estimation results for Mondrian random forests

Setup

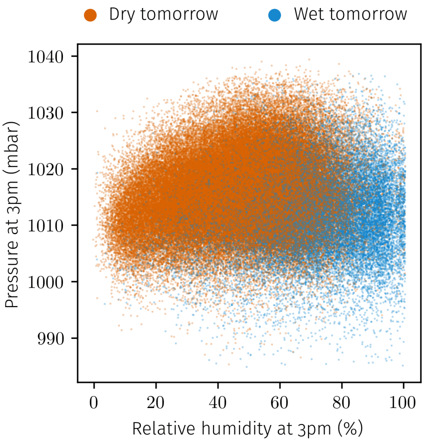
Nonparametric regression setting

- Data (X_i, Y_i) in $[0,1]^d \times \mathbb{R}$ i.i.d. for $1 \leq i \leq n$
- $Y_i = \mu(X_i) + \varepsilon_i$ with $\mathbb{E}[\varepsilon_i \mid X_i] = 0$
- Aim is to estimate and perform inference on unknown $\mu(x)$

Random forest regression estimators

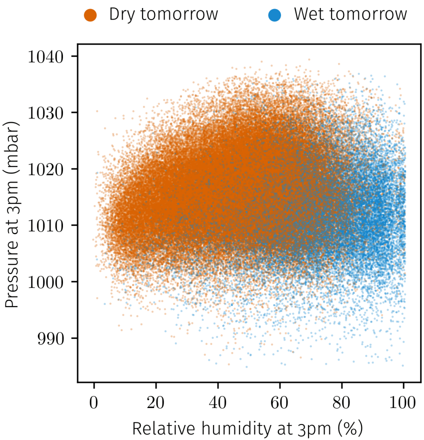
1) Form a partition of $[0,1]^d$, usually using a tree structure
2) Fit constant estimates of μ on each cell in the partition
3) Repeat with different partitions and average the estimates

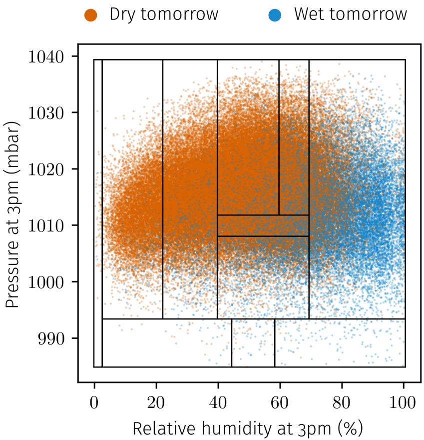
Motivating example: weather forecasting in Australia



- Weather data from Australian Bureau of Meteorology
- Rainfall from 2007–2017 at 49 locations with 125 927 samples
- Predict dry or wet tomorrow with humidity and pressure today
- Random forest classification

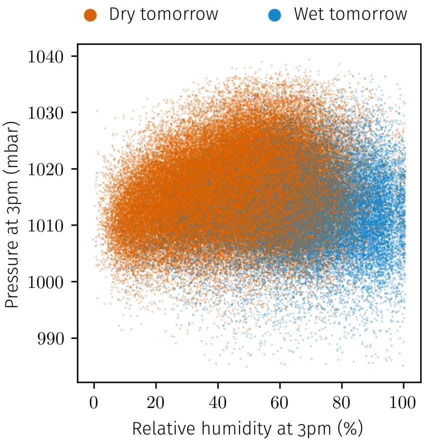
Motivating example: weather forecasting in Australia

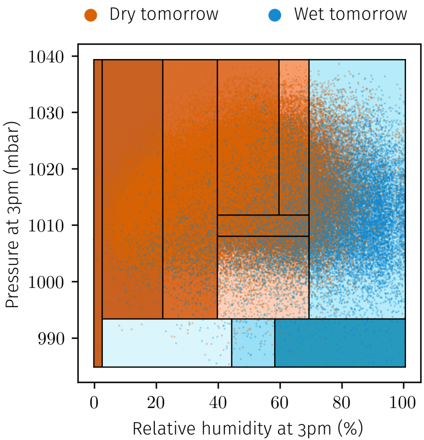




- First generate a partition of the predictor space

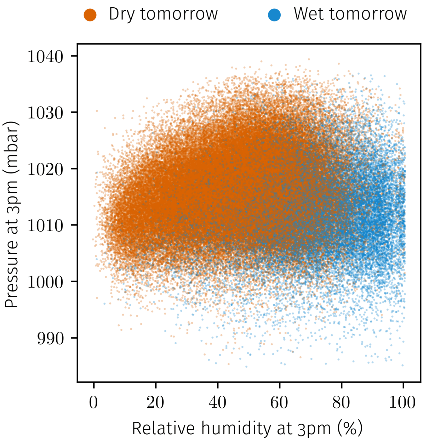
Motivating example: weather forecasting in Australia

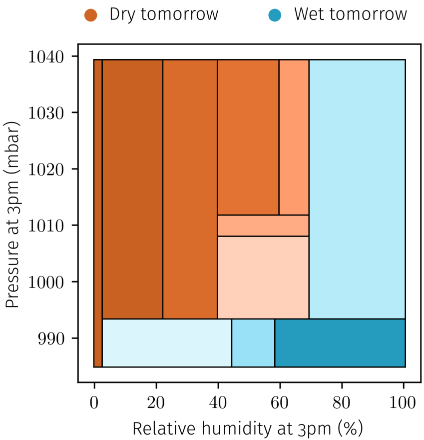




- Compute average response in each cell with dry $= 0$, wet $= 1$

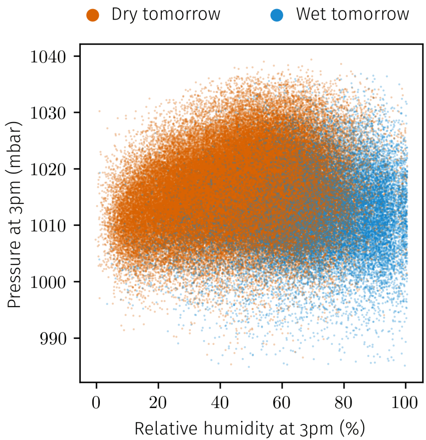
Motivating example: weather forecasting in Australia

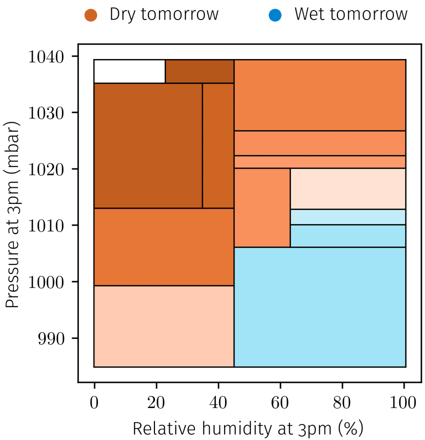




- This gives a single tree estimator of $\mu(x)$

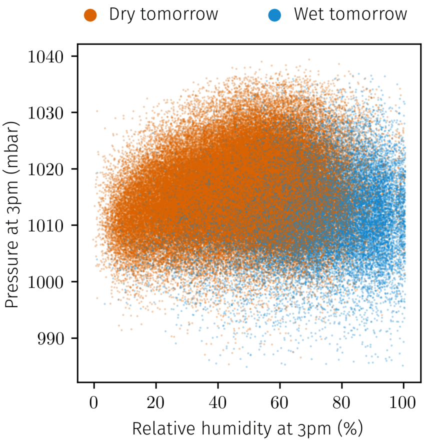
Motivating example: weather forecasting in Australia

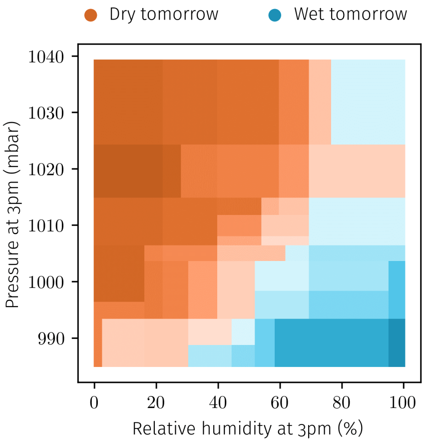




- Repeat with a different partition

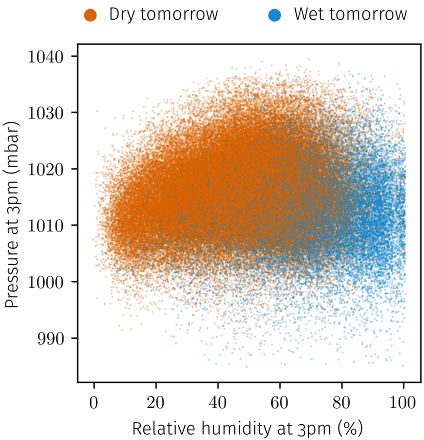
Motivating example: weather forecasting in Australia

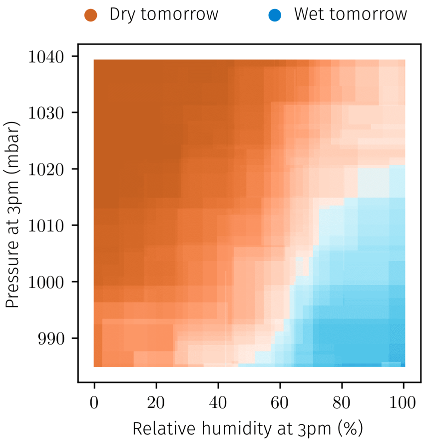




- Average predictions across 2 partitions

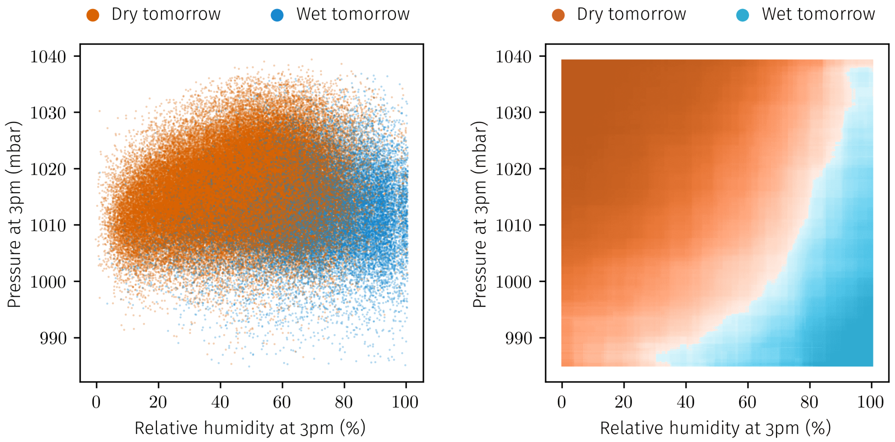
Motivating example: weather forecasting in Australia





- Average predictions across 10 partitions

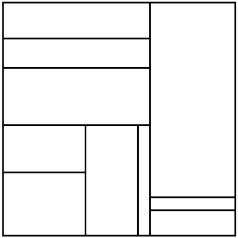
Motivating example: weather forecasting in Australia



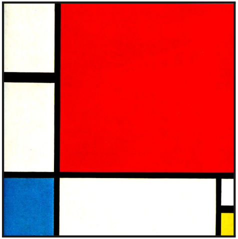
- Average across 30 partitions to get a random forest $\hat{\mu}(x)$

The Mondrian process

- Rectangular partitions sampled from a Mondrian process (Roy and Teh, 2008), write $T \sim \mathcal{MP}([0,1]^d, \lambda)$
- Tree complexity is controlled by the lifetime parameter $\lambda > 0$
- The expected number of cells in T is $(1+\lambda)^d$
- Mondrian random forests popular recently (Mourtada, Gaïffas and Scornet, NeurIPS 2017, AoS 2020, JRSSSB 2021)

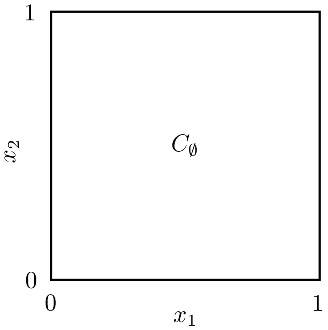


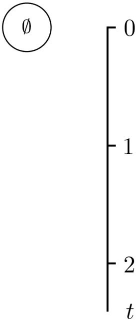
A typical two-dimensional Mondrian partition with $\lambda = 4$



Composition II in Red, Blue, and Yellow, Piet Mondrian, 1930

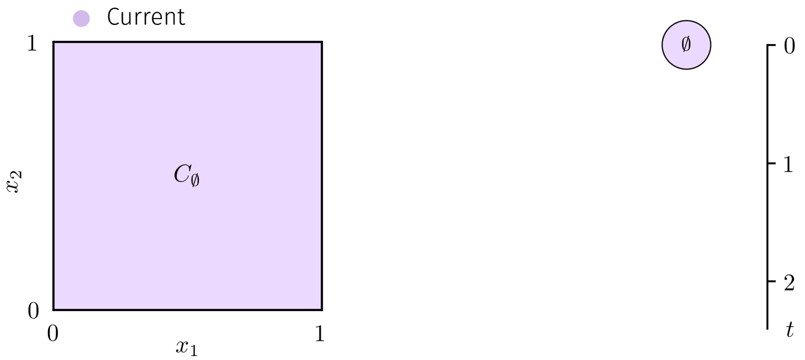
Sampling a partition from the Mondrian process





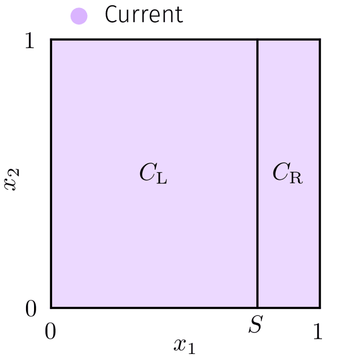
- Fix $\lambda = 2$ and set $t = 0$. The root cell is $C_\emptyset = [0,1]^d$ with $d = 2$
- We make recursive axis-aligned splits to generate a partition
- The lifetime parameter λ determines when to stop splitting
- For any cell C, let $|C|_1 = \sum_{j=1}^{d} |C_j|$ be the half-perimeter

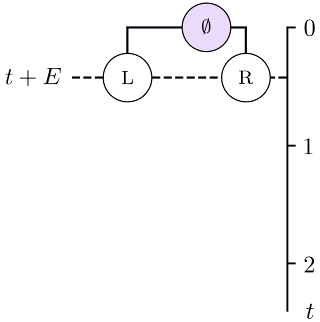
Sampling a partition from the Mondrian process



- Decide whether to split cell $C_\emptyset$
- Sample $E \sim \mathrm{Exp}(|C_\emptyset|_1)$, so $\mathbb{E}[E] = 1/|C_\emptyset|$
- Get $t + E \leq \lambda$ so $C_\emptyset$ is split

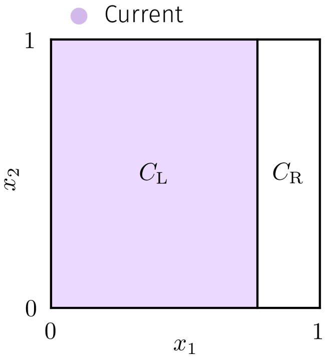
Sampling a partition from the Mondrian process

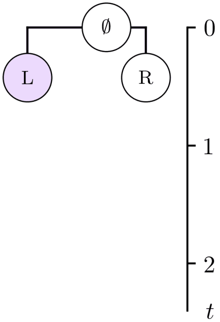




- Choose split axis by $\mathbb{P}(J = j) = \frac{|C_{\emptyset j}|}{|C_\emptyset|_1}$, get $J = 1$
- Select split location by $S \sim \mathrm{Unif}(C_{\emptyset J})$
- Replace $C_\emptyset$ by $C_{\mathrm{L}} = \{x \in C : x_J \leq S\}$ and $C_{\mathrm{R}} = C \setminus C_{\mathrm{L}}$

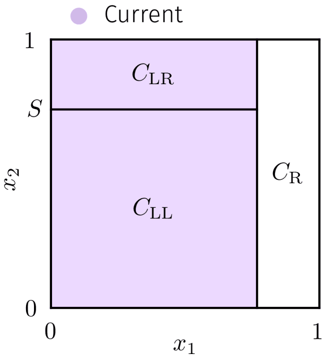
Sampling a partition from the Mondrian process

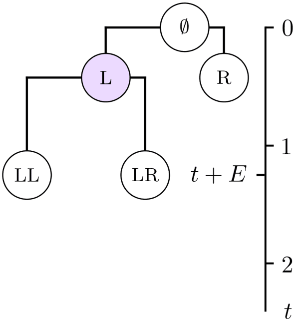




- Decide whether to split cell C_{L}
- Sample $E \sim \mathrm{Exp}(|C_{\mathrm{L}}|_1)$
- Get $t + E \leq \lambda$ so C_{L} is split

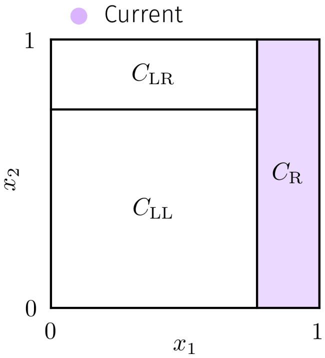
Sampling a partition from the Mondrian process

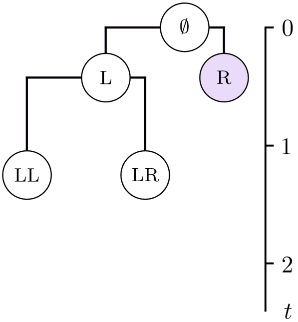




- Choose split axis by $\mathbb{P}(J = j) = \frac{|C_{\mathrm{L}j}|}{|C_{\mathrm{L}}|_1}$, get $J = 2$
- Select split location by $S \sim \mathrm{Unif}(C_{\mathrm{L}J})$
- Replace C_{L} by $C_{\mathrm{LL}} = \{x \in C_{\mathrm{L}} : x_J \leq S\}$ and $C_{\mathrm{LR}} = C_{\mathrm{L}} \setminus C_{\mathrm{LL}}$

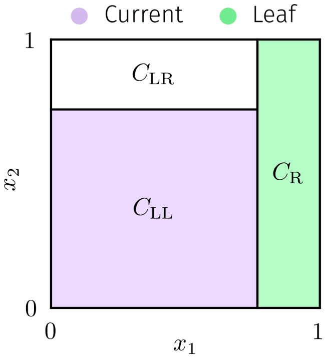
Sampling a partition from the Mondrian process

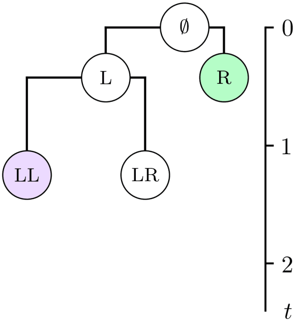




- Decide whether to split cell C_{R}
- Sample $E \sim \mathrm{Exp}(|C_{\mathrm{R}}|_1)$
- Get $t + E > \lambda$ so C_{R} is not split and becomes a leaf

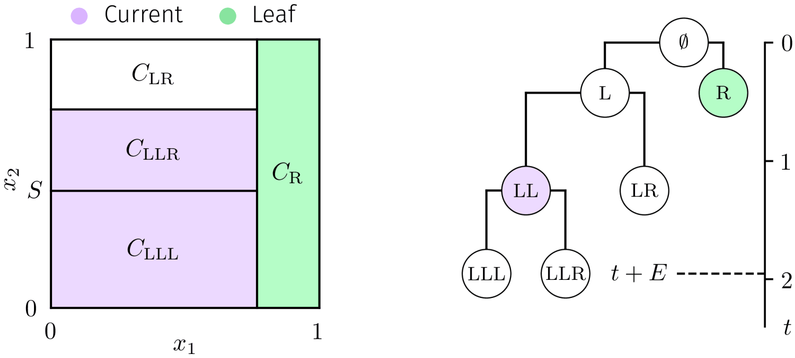
Sampling a partition from the Mondrian process





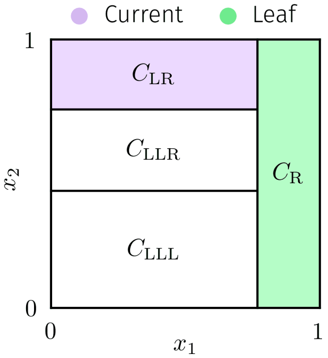
- We continue this process

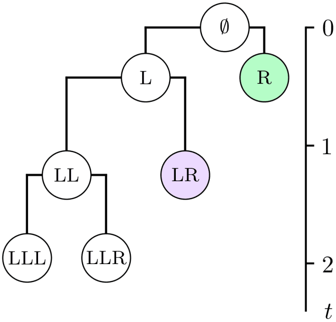
Sampling a partition from the Mondrian process



- C_{LL} is split on axis 2

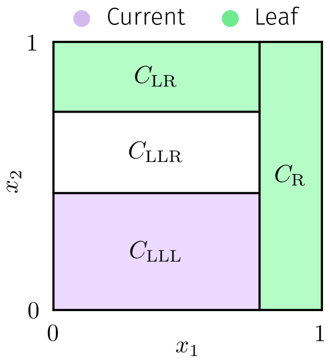
Sampling a partition from the Mondrian process

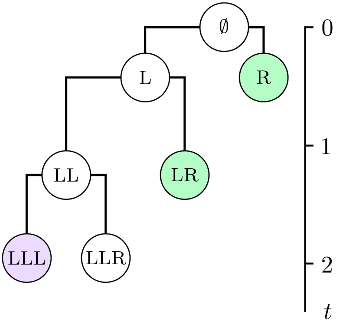




- C_{LR} is not split and becomes a leaf

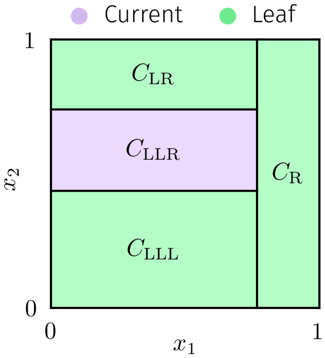
Sampling a partition from the Mondrian process

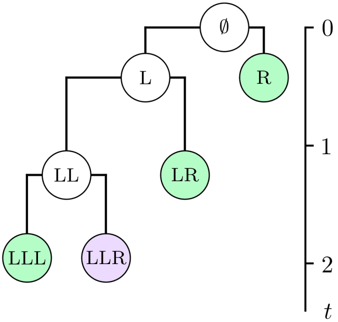




- C_{LLL} becomes a leaf

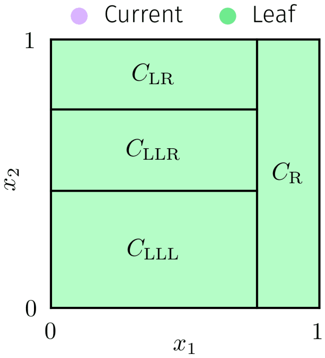
Sampling a partition from the Mondrian process

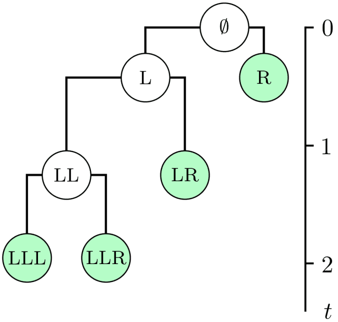




- C_{LLR} becomes a leaf

Sampling a partition from the Mondrian process





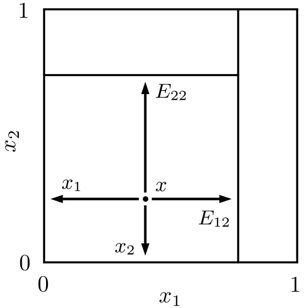
- All cells are now leaves, and the sampling is complete
- To increase λ we continue this process, allowing online fitting
- Australian weather data: rescaled to $[0, 1]^2$ and set $\lambda = 5$

Properties of the Mondrian process

Lemma (Cell shape distribution)

Let $T \sim \mathcal{MP}\big([0,1]^d, \lambda\big)$, take $x \in [0,1]^d$ and write $T(x)$ for the cell containing x. With E_{j1} and E_{j2} independent $\mathrm{Exp}(\lambda)$,

$$T(x) = [0,1]^d \cap \prod_{j=1}^{d} \left[x_j - E_{j1}, x_j + E_{j2}\right]$$



- Roy and Teh (NeurIPS 2008); Mourtada, Gaïffas and Scornet (NeurIPS 2017, AoS 2020, JRSSSB 2021)
- With $d = 1$, have a Poisson process on $[0,1]$ with intensity λ
- The smallest cell is much smaller than the average cell

Mondrian random forests

- Let B be the desired number of trees in the forest
- Sample $T_1, \ldots, T_B \sim \mathcal{MP}\big([0,1]^d, \lambda\big)$ independently
- For each cell in T_b, compute the average Y_i value
- Finally average across all the trees
- Writing $N_b(x) = \sum_{i=1}^n \mathbb{I}\big\{X_i \in T_b(x)\big\}$ for the number of data points in the same cell as x, and with $0/0 = 0$, we have

Definition (Mondrian random forest estimator)

$$\hat{\mu}(x) = \underbrace{\frac{1}{B} \sum_{b=1}^{B}}_{\text{Forest}} \underbrace{\frac{1}{N_b(x)} \sum_{i=1}^{n} Y_i \, \mathbb{I}\big\{X_i \in T_b(x)\big\}}_{\text{Mean of } Y_i \text{ in cell containing } x}$$

Bias–variance decomposition

With $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{T} = (T_1, \ldots, T_B)$,

$$\hat{\mu}(x) - \mu(x) = \underbrace{\hat{\mu}(x) - \mathbb{E}\big[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\big]}_{\text{Variance}} + \underbrace{\mathbb{E}\big[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\big] - \mu(x)}_{\text{Bias}}$$

1) Derive a central limit theorem for the variance term
2) Approximate the bias term in probability
3) Perform inference by ensuring the bias is negligible
4) Minimax optimal estimation with debiasing

Assumptions on data and estimator

- Recall (X_i, Y_i) in $[0,1]^d \times \mathbb{R}$ i.i.d. with $Y_i = \mu(X_i) + \varepsilon_i$
- X_i has Lebesgue density f, bounded away from zero
- A version of $\sigma^2(X_i) = \mathbb{E}\left[\varepsilon_i^2 \mid X_i\right]$ is Lipschitz
- $\mathbb{E}\left[\varepsilon_i^4 \mid X_i\right]$ is bounded almost surely
- Both μ and f are β-Hölder continuous for some $\beta \geq 1$
- $x \in (0,1)^d$ is an interior evaluation point
- $\frac{\lambda^d \log n}{n} \to 0$ and $\log \lambda \asymp \log B \asymp \log n$, so $\lambda \to \infty$ and $B \to \infty$

Definition (β-Hölder continuity)

With $\underline{\beta}$ the largest integer less than β, for all $x, x' \in [0,1]^d$,

$$\max_{|\nu| = \underline{\beta}} \left| \partial^\nu g(x) - \partial^\nu g(x') \right| \lesssim \|x - x'\|_2^{\beta - \underline{\beta}}$$

Central limit theorem for Mondrian random forests

Theorem (Central limit theorem for Mondrian random forests)

$$\sqrt{\frac{n}{\lambda^d}}\Big(\hat{\mu}(x) - \mathbb{E}\big[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\big]\Big) \rightsquigarrow \mathcal{N}\big(0, \Sigma(x)\big)$$

where

$$\Sigma(x) = \frac{\sigma^2(x)}{f(x)} \left(\frac{4 - 4\log 2}{3}\right)^d$$

$$\hat{\mu}(x) - \mathbb{E}\big[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\big] = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{N_b(x)} \sum_{i=1}^{n} \varepsilon_i \, \mathbb{I}\big\{X_i \in T_b(x)\big\}$$

- Essential that $B \to \infty$, or randomness persists in the limit
- No conditional independence as $N_b(x)$ depends on all X_i
- Replacing $N_b(x)$ by $nf(x)|T_b(x)|$ fails as $\mathbb{E}\left[\frac{1}{|T_b(x)|^2}\right] = \infty$
- Central limit theorems based on $2 + \delta$ moments inadequate

Central limit theorem for Mondrian random forests

$$\hat{\mu}(x) - \mathbb{E}\big[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\big] = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{N_b(x)} \sum_{i=1}^{n} \varepsilon_i \, \mathbb{I}\big\{X_i \in T_b(x)\big\}$$

- Use a martingale central limit theorem (Hall and Heyde, 1980)
- Take the filtration $\mathcal{F}_{ni} = \sigma\left(\mathbf{X}, \mathbf{T}, \varepsilon_1, \ldots, \varepsilon_i\right)$ and consider $\sum_{i=1}^{n} M_{ni}(x)$ with the martingale differences

$$M_{ni}(x) = \sqrt{\frac{n}{\lambda^d}} \frac{1}{B} \sum_{b=1}^{B} \frac{1}{N_b(x)} \varepsilon_i \, \mathbb{I}\{X_i \in T_b(x)\}$$

- Verify $\mathbb{E}\left[\max_{1 \leq i \leq n} M_{ni}(x)^2\right] \lesssim 1$ and $\sum_{i=1}^{n} M_{ni}(x)^2 \to_{\mathbb{P}} \Sigma(x)$
- Nonlinear structure handled by the Efron–Stein inequality

Bias of Mondrian random forests

Theorem (Bias of Mondrian random forests)

There exist $B_r(x)$ depending only on f and μ such that

$$\left| \mathbb{E}\big[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\big] - \mu(x) - \sum_{r=1}^{\lfloor \beta/2 \rfloor} \frac{B_r(x)}{\lambda^{2r}} \right| \lesssim_{\mathbb{P}} \frac{1}{\lambda^\beta} + \frac{1}{\lambda\sqrt{B}} + \frac{\log n}{\lambda}\sqrt{\frac{\lambda^d}{n}}$$

- We approximate the bias with a Taylor polynomial in $1/\lambda^2$
- If B does not diverge there is a first-order bias of size $1/\lambda$
- In large forests and with $\beta \geq 2$, leading bias is of size $1/\lambda^2$
- Setting $\lambda \asymp n^{\frac{1}{d+4}}$ and $B \gg n^{\frac{2}{d+4}}$ gives for $\beta \geq 2$

$$\left|\hat{\mu}(x) - \mu(x)\right| \lesssim_{\mathbb{P}} \underbrace{\sqrt{\frac{\lambda^d}{n}}}_{\text{Variance}} + \underbrace{\frac{1}{\lambda^2} + \frac{1}{\lambda\sqrt{B}}}_{\text{Bias}} \lesssim n^{-\frac{2}{d+4}}$$

Inference with Mondrian random forests

- Combine central limit theorem and bias bound for inference
- Bias is negligible if $\beta \geq 2$ and $\frac{1}{\lambda^2} + \frac{1}{\lambda\sqrt{B}} \ll \sqrt{\frac{\lambda^d}{n}}$
- We construct a variance estimator $\hat{\Sigma}(x) \to_{\mathbb{P}} \Sigma(x)$
- Let q_α be the $1 - \frac{\alpha}{2}$ quantile of $\mathcal{N}(0,1)$

Theorem (Feasible confidence intervals)

With $\beta \geq 2$, if $\lambda \gg n^{\frac{1}{d+4}}$ and $B \gg n^{\frac{2}{d+4}}$ then

$$\mathbb{P}\left(\mu(x) \in \left[\hat{\mu}(x) \pm \sqrt{\frac{\lambda^d}{n}} \hat{\Sigma}(x)^{1/2} q_\alpha\right]\right) \to 1 - \alpha$$

Debiased Mondrian random forests

- Bias approximation with $\beta > 2$ for lifetimes λ and 2λ gives

$$\mathbb{E}\left[\hat{\mu}(x;\lambda) \mid \mathbf{X}, \mathbf{T}\right] \approx \mu(x) + \frac{B_1(x)}{\lambda^2} \tag{1}$$

$$\mathbb{E}\left[\hat{\mu}(x;2\lambda) \mid \mathbf{X}, \mathbf{T}\right] \approx \mu(x) + \frac{B_1(x)}{4\lambda^2} \tag{2}$$

- Take a linear combination to annihilate the leading bias

$$\mathbb{E}\left[-\frac{1}{3}\hat{\mu}(x;\lambda) + \frac{4}{3}\hat{\mu}(x;2\lambda) \mid \mathbf{X}, \mathbf{T}\right] \approx \mu(x) + 0$$

- Cancel all $J = \lfloor \underline{\beta}/2 \rfloor$ bias terms to get the debiased estimator

$$\hat{\mu}_{\mathrm{d}}(x) = \sum_{s=0}^{J} \omega_s \hat{\mu}(x; a_s \lambda)$$

- Here a_s are fixed, and ω_s solve the linear equations $\sum_{s=0}^{J} \omega_s = 1$ and $\sum_{s=0}^{J} \omega_s a_s{}^{-2r} = 0$ for $1 \leq r \leq J$

Results for debiased Mondrian random forests

Theorem (Improved bias bound)

$$\left|\mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x) \mid \mathbf{X}, \mathbf{T}\right] - \mu(x)\right| \lesssim_{\mathbb{P}} \frac{1}{\lambda^{\beta}} + \frac{1}{\lambda\sqrt{B}} + \frac{\log n}{\lambda}\sqrt{\frac{\lambda^d}{n}}$$

Theorem (Central limit theorem with debiasing)

$$\sqrt{\frac{n}{\lambda^d}}\Big(\hat{\mu}_{\mathrm{d}}(x) - \mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x) \mid \mathbf{X}, \mathbf{T}\right]\Big) \rightsquigarrow \mathcal{N}\big(0, \Sigma_{\mathrm{d}}(x)\big)$$

Theorem (Feasible confidence intervals with debiasing)

If $\lambda \gg n^{\frac{1}{d+2\beta}}$ and $B \gg n^{\frac{2\beta-2}{d+2\beta}}$, with $\hat{\Sigma}_{\mathrm{d}}(x)$ a variance estimator,

$$\mathbb{P}\left(\mu(x) \in \left[\hat{\mu}_{\mathrm{d}}(x) \pm \sqrt{\frac{\lambda^d}{n}}\hat{\Sigma}_{\mathrm{d}}(x)^{1/2} q_{\alpha}\right]\right) \to 1 - \alpha$$

Minimax optimality

Theorem (Minimaxity of debiased Mondrian random forests)

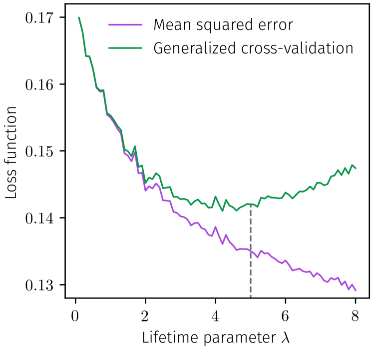
If $\lambda \asymp n^{\frac{1}{d+2\beta}}$ and $B \gtrsim n^{\frac{2\beta-2}{d+2\beta}}$, then

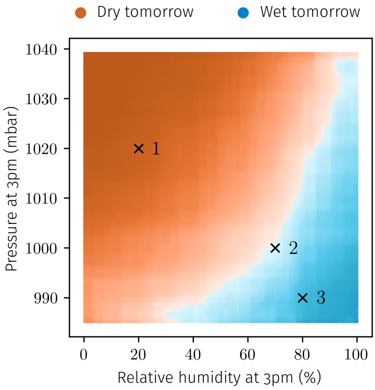
$$\mathbb{E}\left[\left(\hat{\mu}_{\mathrm{d}}(x) - \mu(x)\right)^2\right]^{1/2} \lesssim \underbrace{\sqrt{\frac{\lambda^d}{n}}}_{\text{Variance}} + \underbrace{\frac{1}{\lambda^\beta} + \frac{1}{\lambda\sqrt{B}}}_{\text{Bias}} \lesssim n^{-\frac{\beta}{d+2\beta}}$$

Estimator	Minimax condition
Mondrian tree*	$\beta \in (0, 1]$
Mondrian random forest*	$\beta \in (0, 2]$
Debiased Mondrian random forest	$\beta \in (0, \infty)$

*Established by Mourtada et al. (2020)

Example: weather forecasting in Australia





Point	Humidity	Pressure	Chance of rain	95% confidence interval
1	20%	1020 mbar	4.3%	4.1% – 4.6%
2	70%	1000 mbar	53.0%	52.0% – 54.0%
3	80%	990 mbar	77.5%	74.4% – 80.6%

Conclusion and ongoing work

Contributions to studying the Mondrian random forest estimator

- Provided a novel central limit theorem allowing fully feasible statistical inference via variance estimation
- Presented a new debiasing procedure allowing for inference under milder conditions
- Demonstrated minimax optimality for arbitrary dimension and smoothness, the first result for any forest estimator

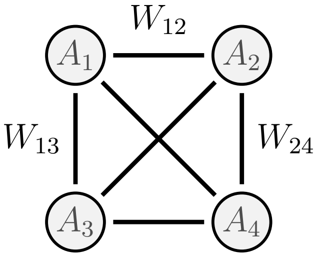
Ongoing and future work

- Heterogeneous and data-dependent lifetimes $\hat{\lambda}_j$ or $\hat{\lambda}(x)$
- Improved estimation with additive models or local regression
- Uniform inference via strong approximation

Uniform Inference for Kernel Density Estimators with Dyadic Data

With Matias D. Cattaneo and Yingjie Feng

Dyadic data



Example of dyadic data

- A_i is GDP of country i
- W_{ij} is value of trade $i \leftrightarrow j$

- W_{ij} random variables associated with edges of a network
- Write $W_{ij} = W(A_i, A_j, V_{ij})$ by Aldous–Hoover with A_i latent node variables and V_{ij} latent idiosyncratic shocks
- Unknown Lebesgue density $f(w)$ estimated by $\hat{f}(w)$ on $\mathcal{W}$
- We provide the minimax-optimal estimation rate for $\hat{f}(w)$
- Uniform inference on $f(w)$ by strong approximation

Dyadic kernel density estimation

Dyadic kernel density estimator

$$\hat{f}(w) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{h} K\left(\frac{W_{ij} - w}{h}\right)$$

- Bandwidth h controls bias–variance tradeoff
- Higher-order boundary kernels K improve bias properties
- We analyze the U-statistic Hoeffding-type decomposition

$$\hat{f}(w) - f(w) = \underbrace{B(w)}_{\text{smoothing bias}} + \underbrace{L(w)}_{\text{i.i.d. average}} + \underbrace{E(w)}_{\text{conditional i.n.i.d. average}} + \underbrace{Q(w)}_{\text{U-statistic}}$$

- $L(w)$, $E(w)$ and $Q(w)$ are mean-zero and orthogonal

Minimax-optimal uniform dyadic estimation

- Using an order p boundary kernel, if f is β-Hölder then

$$\sup_{w \in \mathcal{W}} \big|B(w)\big| \lesssim h^{p \wedge \beta} \qquad \mathbb{E}\left[\sup_{w \in \mathcal{W}} |L(w)|\right] \lesssim \frac{D}{\sqrt{n}}$$

$$\mathbb{E}\left[\sup_{w \in \mathcal{W}} |E(w)|\right] \lesssim \sqrt{\frac{\log n}{n^2 h}} \qquad \mathbb{E}\left[\sup_{w \in \mathcal{W}} |Q(w)|\right] \lesssim \frac{1}{n}$$

- Optimize the bound with $p \geq \beta$ and $h \asymp \left(\frac{\log n}{n^2}\right)^{\frac{1}{2\beta+1}}$
- Then we attain the minimax dyadic estimation rate

Theorem (Minimax-optimal uniform dyadic estimation)

$$\sup_{w \in \mathcal{W}} \big|\hat{f}(w) - f(w)\big| \lesssim_{\mathbb{P}} \underbrace{h^{p \wedge \beta}}_{B(w)} + \underbrace{\frac{D}{\sqrt{n}}}_{L(w)} + \underbrace{\sqrt{\frac{\log n}{n^2 h}}}_{E(w)} \lesssim \frac{D}{\sqrt{n}} + \left(\frac{\log n}{n^2}\right)^{\frac{\beta}{2\beta+1}}$$

Dyadic strong approximation construction

- Need distributional approximations for both $L(w)$ and $E(w)$
- No uniform central limit theorem as $E(w)$ is not tight
- For the i.i.d. sum $L(w)$, use KMT coupling (Komlós et al., 1975)

$$\sup_{w \in \mathcal{W}} \left| \sqrt{n} L(w) - Z_L(w) \right| \lesssim_{\mathbb{P}} \frac{D \log n}{\sqrt{n}}$$

- $E(w)$ is a sum of $\binom{n}{2}$ conditionally independent but not i.i.d. terms so use a version of Yurinskii's coupling (Yurinskii, 1978)

$$\sup_{w \in \mathcal{W}} \left| \sqrt{n^2 h} E(w) - Z_E(w) \right| \lesssim_{\mathbb{P}} \frac{(\log n)^{3/8}}{n^{1/4} h^{3/8}}$$

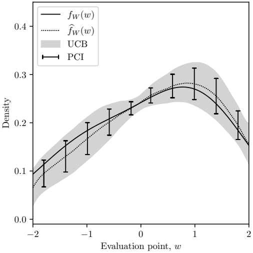
- Combine these with the uniform bounds on $B(w)$ and $Q(w)$

Dyadic uniform inference via strong approximation

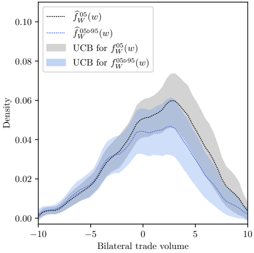
Theorem (Strong approximation and uniform confidence bands)

$$\sup_{w \in \mathcal{W}} \left| \frac{\hat{f}(w) - f(w)}{\sqrt{\operatorname{Var}\big[\hat{f}(w)\big]}} - Z(w) \right| \to_{\mathbb{P}} 0, \qquad Z(w) \text{ Gaussian process}$$

$$\mathbb{P}\left(f(w) \in \left[\hat{f}(w) \pm \hat{q}_{1-\alpha} \sqrt{\widehat{\operatorname{Var}}\big[\hat{f}(w)\big]} \right] \ \forall w \in \mathcal{W} \right) \to 1 - \alpha$$



(a) Synthetic data with degeneracy



(b) Counterfactual trade analysis

Questions

Cattaneo, M. D., Klusowski, J. M., and Underwood, W. G. (2023)
Inference with Mondrian random forests
`arXiv:2310.09702`
`github.com/wgunderwood/MondrianForests.jl`

Cattaneo, M. D., Feng, Y., and Underwood, W. G. (2024).
Uniform inference for kernel density estimators with dyadic data
`arXiv:2201.05967`
`github.com/wgunderwood/DyadicKDE.jl`

References I

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Breiman, L. (2004). Consistency for a simple model of random forests. *University of California at Berkeley. Technical Report*, 670.

Bureau of Meteorology, Australian Government (2017). Daily weather observations. `http://www.bom.gov.au/climate/data/`.

Cattaneo, M. D., Feng, Y., and Underwood, W. G. (2024). Uniform inference for kernel density estimators with dyadic data. *Journal of the American Statistical Association*, forthcoming.

Cattaneo, M. D., Klusowski, J. M., and Underwood, W. G. (2023). Inference with Mondrian random forests. *Preprint*. `arXiv:2310.09702`.

Cattaneo, M. D., Masini, R. P., and Underwood, W. G. (2022). Yurinskii's coupling for martingales. *Preprint*. `arXiv:2210.00362`.

Cutler, A. and Zhao, G. (2001). PERT: perfect random tree ensembles. *Computing Science and Statistics*, 33(4):90–4.

References II

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63:3–42.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and its Application*. Academic Press, New York, NY.

Komlós, J., Major, P., and Tusnády, G. (1975). An approximation of partial sums of independent RVs, and the sample DF. I. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32(1-2):111–131.

Lakshminarayanan, B., Roy, D. M., and Teh, Y. W. (2014). Mondrian forests: Efficient online random forests. *Advances in Neural Information Processing Systems*, 27.

Menze, B. H., Kelm, B. M., Splitthoff, D. N., Koethe, U., and Hamprecht, F. A. (2011). On oblique random forests. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Proceedings, Part II 22*, pages 453–469. Springer.

References III

Mourtada, J., Gaïffas, S., and Scornet, E. (2017). Universal consistency and minimax rates for online Mondrian forests. *Advances in Neural Information Processing Systems*, 30.

Mourtada, J., Gaïffas, S., and Scornet, E. (2020). Minimax optimal rates for Mondrian trees and forests. *Annals of Statistics*, 48(4):2253–2276.

Mourtada, J., Gaïffas, S., and Scornet, E. (2021). AMF: Aggregated Mondrian forests for online learning. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(3):505–533.

Roy, D. M. and Teh, Y. W. (2008). The Mondrian process. In *Neural Information Processing Systems*, volume 21.

Yurinskii, V. V. (1978). On the error of the Gaussian approximation for convolutions. *Theory of Probability & its Applications*, 22(2):236–247.