Random Forest Density Estimation

Hongwei Wen[1], Hanyuan Hang[1]

Department of Applied Mathematics
University of Twente

June 20, 2022

Outline

Motivation

- **Task:**
  Density estimation is one of the most imperative topics in unsupervised learning among machine learning community.

- **Popular Methods:**
  Kernel density estimation (KDE): Lack of adaptivity.
  Histogram density estimation (HDE): Low computational efficiency.
  Tree-based methods: Boundary Discontinuity.

- **Random Forest Density Estimation (RFDE):**
  Local adaptivity;
  Alleviate boundary discontinuity;
  High efficient partition.

Contributions

- We propose a tree-based density estimation algorithm called random forest density estimation (RFDE).
- From a learning theory point of view, we prove the fast convergence rates of RFDE with assumptions that the underlying density functions lie in the Hölder space.
- We are the first to explain the benefits of ensemble for density estimation from the perspective of the convergence rates.
- In experiments, we validate the theoretical results and evaluate our RFDE through comparisons on both synthetic and real data.

Outline

Random Tree Partition

Mid-point random tree partitions suggested by Biau (2012) and Breiman (2004).

- Each dimension has the equal probability $1/d$ to be chosen.
- The split is at the midpoint of the chosen dimension.

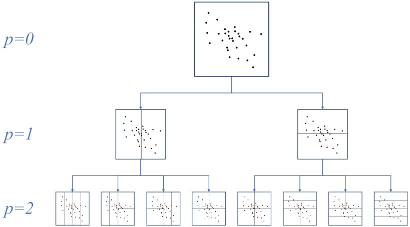

Figure: Random tree partitions with depth p for the dimension $d = 2$.

Note: In each partition with depth p, there are 2^p cells with equal volume 2^{-p}.

Algorithm

- For a certain partition with depth p, we denote $A_p(x)$ as the cell containing the point x.
- Based on the i.i.d observations $\{x_i\}_{i=1}^n \sim P$, the random tree density estimator is defined by

$$f_D^p(x) := \frac{D(A_p(x))}{\mu(A_p(x))} = \frac{n^{-1}\mathbf{1}\{x_i \in A_p(x)\}}{2^{-p}}$$

  where $D(A_p(x)) \to P(X \in A_p(x))$ as $n \to \infty$.
- RFDE with T base learners is given by

$$f_{D,E}(x) = \frac{1}{T}\sum_{t=1}^{T} f_{D,t}^p(x).$$

Outline

Main Theoretical Results

- **Fast convergence rates.**
  - Rate $O\big(n^{-\frac{1-4^{-\alpha}}{d\ln 2+1-4^{-\alpha}}}\big)$ in $C^{0,\alpha}$ with high probability.
- **Ensemble estimators achieve the asymptotic smoothness.**
  - Rate $O\big(n^{-\frac{1}{1+d\ln 2}}\big)$ in $C^{1,\alpha}$ by choosing $T_n \gtrsim n^{\frac{1}{4+4d\ln 2}}$.
- **Benefits of ensemble.**
  - Lower bound for random tree density estimators $O\big(n^{-\frac{1-4^{-\alpha}}{d\ln 2+1-4^{-\alpha}}}\big)$ in $C^{1,\alpha}$.
  - When $d \geq 2$, the upper bound for RFDE is strictly smaller than this lower bound for tree estimators.

Outline

Empirical Comparison

Table 1. Average *ANLL* and *MAE* over simulated datasets

d	Method	Type I		Type II		Type III	
		ANLL	*MAE*	*ANLL*	*MAE*	*ANLL*	*MAE*
2	RFDE (Ours)	**−0.57***	**0.65**	**3.14***	**1.64e-2***	**1.97***	**3.29e-2***
	KDE	−0.37	1.06	3.27	2.31e-2	2.14	5.32e-2
	HDE	−0.52	0.66	3.21	1.81e-2	2.01	3.82e-2
5	RFDE (Ours)	**−1.18***	**7.77***	**8.17***	**6.78e-4***	**3.12***	**0.09***
	KDE	−0.32	12.40	8.65	8.27e-4	3.86	0.15
	HDE	10.17	19.70	10.77	1.33e-3	6.09	0.17
7	RFDE (Ours)	**−1.48***	**30.60***	**10.89***	**5.54e-5***	**3.96***	**0.13***
	KDE	0.03	40.74	12.48	6.05e-5	5.16	0.18
	HDE	11.48	73.97	11.49	1.05e-4	9.88	0.20

* The best results are marked in **bold**. We use * to represent that the best method is significantly better than the other compared methods.

Table 2. Average *ANLL* over real data sets

Datasets	d'	RFDE	KDE	HDE	Datasets	d'	RFDE	KDE	HDE
Adult	2	**−1.5226** (0.0113)	−0.7402 (0.0027)	−0.9838 (0.0143)	Diabetes	1	**−0.8073** (0.0576)	−0.2627 (0.0111)	−0.6067 (0.0676)
	4	**−1.8374** (0.0141)	−0.3075 (0.0032)	−0.7789 (0.0303)		3	**−1.5378** (0.0953)	−0.4042 (0.0403)	−0.3142 (0.3422)
	8	**−5.7832** (0.0557)	−2.2970 (0.0108)	–		4	**−1.8387** (0.1433)	−0.8353 (0.0773)	2.9933 (0.6034)
	10	**−6.6704** (0.0475)	−3.4372 (0.0110)	–		6	**−2.3838** (0.1912)	−1.9693 (0.1550)	9.1732 (0.3902)
Australian	2	**−0.5836** (0.1796)	1.3155 (0.0234)	0.3898 (0.1494)	Credit	2	**1.2659** (0.1142)	1.5435 (0.0183)	1.6649 (0.1968)
	4	**−5.2131** (0.3508)	0.8518 (0.0291)	−2.2163 (0.2507)		5	**−1.3479** (0.2889)	1.4844 (0.0516)	1.3455 (0.5457)
	8	**−3.6821** (0.3678)	0.6879 (0.1056)	–		8	**2.1191** (0.2905)	3.0453 (0.1067)	–
	10	**−1.8187** (0.3474)	0.4995 (0.1748)	–		11	**3.1343** (0.3182)	3.5221 (0.2292)	–
Breast-cancer	1	**−0.0323** (0.2059)	0.6907 (0.0394)	0.3697 (0.1011)	Abalone	1	0.5664 (0.0144)	**0.5458** (0.0103)	0.5609 (0.0140)
	3	**−3.3262** (0.5219)	0.1743 (0.1268)	1.3773 (0.3432)		3	**−2.6793** (0.0818)	−0.9493 (0.0282)	−1.2716 (0.0594)
	6	**−7.5657** (0.9746)	−1.1397 (0.2788)	1.8392 (0.5542)		4	**−4.0743** (0.0619)	−2.6572 (0.0309)	−2.2145 (0.1534)
	8	**−5.1952** (1.2260)	−2.1110 (0.3906)	–		6	**−7.1922** (0.0722)	−6.4804 (0.0445)	0.3270 (0.3553)

* The best results are marked in **bold**, and the standard deviation is reported in the parenthesis. The results of HDE with $d' > 7$ is corrupted due to numerical problems.