GBHT: Gradient Boosting Histogram Transform for Density Estimation

Jingyi Cui *[1]   Hanyuan Hang *[2]   Yisen Wang [1]   Zhouchen Lin [1,3]

[1]Peking University,
Beijing, China
[2]University of Twente,
Enschede, Netherlands
[3]Pazhou Lab,
Guangzhou, China

July 12, 2021

Outline

Motivation

- **Why boosting?**
  Boosting algorithms are praised as one of the most successful algorithms over two decades.
- **Why employ boosting on density estimation?**
  - With most boosting-based algorithms focus on supervised problems, unsupervised boosting algorithms with solid theoretical guarantees remain to be studied.
  - Density estimation is one of the most imperative topics in unsupervised learning among machine learning community.

Contributions

- We exploit boosting to improve the accuracy in density estimation by taking an **unsupervised loss function**.
- We prove the **fast convergence rates** of GBHT with mild assumptions, which verify the experimental performance.
- We are the first to **explain the strength of boosting** density estimation by comparing the theoretical properties of single estimator and the boosted one.

Outline

Loss Function

- *Average negative log-likelihood* (ANLL)

$$L(x, \hat{f}) := -\log \hat{f}(x).$$

- At the t-th iteration, the minimization of the empirical risk equals to the minimization of $\sum_{i=1}^{n} -\log(F_{t-1}(x_i) + \varepsilon_t f_t(x_i))$.
  Using Taylor expansion, we get

$$\begin{aligned}&\sum_i -\log(F_{t-1}(x_i) + \varepsilon_t f_t(x_i))\\ =&\sum_i -\log(F_{t-1}(x_i)) - \varepsilon_t \cdot \frac{1}{F_{t-1}(x_i)} f_t(x_i) + O(\varepsilon_t^2).\end{aligned}$$

Main Algorithm

Algorithm 1 Gradient Boosting Histogram Transform (GBHT)

Input: Training data $D := \{x_1, \ldots, x_n\}$;
Bandwidth parameters $\underline{h}_0$, $\overline{h}_0$;
Number of iterations T.

Initialization: F_0 is set to be uniformly distributed on cells $A_j \in \pi_H$ satisfying $A_j \cap D \neq \emptyset$.

for $t = 1$ **to** T **do**

Set the sample weight $\omega_{t,i} = 1/F_{t-1}(x_i)$;
For random histogram transformation H_t:
Find $f_t = \arg\max_{f \in \mathcal{F}_t} \sum_{i=1}^{n} \omega_{t,i} f(x_i)$;
Find $\alpha_t := \arg\min_\alpha \sum_{i=1}^{n} -\log\big((1-\alpha)F_{t-1}(x_i) + \alpha f_t(x_i)\big)$;
Update $F_t = (1-\alpha_t)F_{t-1} + \alpha_t f_t$;

end for

return F_T.

Histogram Transform Partition

Histogram Transform for input space:

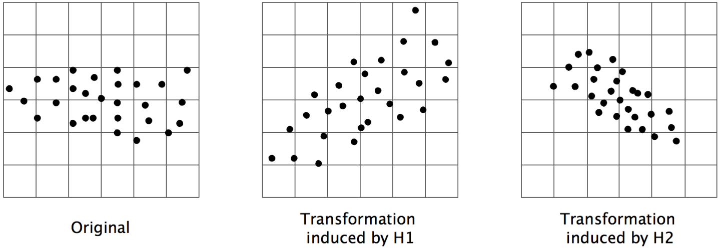

Figure: Two possible histogram transforms in 2-D.

GBHT under the RERM Framework

- Function space:

$$\mathcal{F}_{\mathrm{H}} := \left\{ \sum_{j \in \mathcal{I}_{\mathrm{H}}} c_j \mathbf{1}_{A_j} \;\middle|\; c_j \geq 0, \sum_{j \in \mathcal{I}_{\mathrm{H}}} c_j \mu(A_j) = 1 \right\}.$$

- We calculate the HT density estimator by the RERM over $\mathcal{F}_{\mathrm{H}}$ i.e.

$$(f_{\mathrm{D,H}}, h^*) = \underset{f \in \mathcal{F}_{\mathrm{H}}, h \in \mathbb{R}^d}{\mathit{argmin}} \; \Omega(h) + \mathcal{R}_{\mathrm{L,D}}(f),$$

where $\Omega(h)$ represent the penalty on model complexity.

Outline

Main Theoretical Results

- **Fast convergence rates.**
  - Rate $O(n^{-\frac{2\alpha}{4\alpha+d}})$ in $C^{0,\alpha}$ with high probability.
- **Boosting helps improve rates.**
  - Rate $O(n^{-\frac{2(1+\alpha)}{4(1+\alpha)+d}})$ in $C^{1,\alpha}$ by choosing $T_n \gtrsim n^{\frac{2\alpha}{4(1+\alpha)+d}}$.
- **Deficiency of base estimators.**
  - Lower bound of excess risk $O(n^{-\frac{2}{2+d}})$ in $C^{1,\alpha}$.
  - When $d \geq 2(1+\alpha)/\alpha$, the upper bound of boosting estimator is smaller than this lower bound for base estimators.

Outline

Experiments

Table: Average *ANLL* and *MAE* over simulated datasets

d	Method	Type I		Type II		Type III		Type IV	
		ANLL	*MAE*	*ANLL*	*MAE*	*ANLL*	*MAE*	*ANLL*	*MAE*
5	GBHT (Ours)	**6.26**	**2.41e-3**	**−0.80**	**10.31**	**8.23**	**6.61e-4**	**3.85**	**0.14**
	KDE	6.33	2.36e-3	−0.32	12.40	8.65	8.27e-4	3.86	0.15
	MIX	6.53	3.08e-3	1.82	13.91	9.64	9.54e-4	5.35	0.14
	HDE	9.33	4.86e-3	10.17	19.70	10.77	1.33e-3	6.09	0.17
7	GBHT (Ours)	**8.36**	**4.33e-4**	**−0.45**	**34.91**	**10.81**	**5.30e-5**	**5.10**	**0.18**
	KDE	8.77	5.13e-4	0.03	40.74	12.48	6.05e-5	5.16	0.18
	MIX	8.65	5.38e-4	2.61	42.13	11.34	6.32e-5	7.02	0.19
	HDE	11.35	1.45e-3	11.48	73.97	11.49	1.05e-4	9.88	0.20

* The best results are marked in **bold**.

Experiments

Table: Average *ANLL* over real data sets

Datasets	d'	GBHT	KDE	MIX	Datasets	d'	GBHT	KDE	MIX
Adult	2	−**1.2371**	−0.7402	1.3572	Diabetes	1	−**0.7057**	−0.2627	0.7131
		(0.0312)	(0.0027)	(0.0050)			(0.1253)	(0.0111)	(0.0186)
	4	−**1.9312**	−0.3075	1.7609		3	−**1.5982**	−0.4042	0.5193
		(0.0667)	(0.0032)	(0.0059)			(0.1011)	(0.0403)	(0.0600)
	8	−**5.5922**	−2.2970	0.8562		4	−**1.8605**	−0.8353	0.0403
		(0.1097)	(0.0108)	(0.3183)			(0.1424)	(0.0773)	(0.0771)
	10	−**6.0740**	−3.4372	−0.8975		6	−**2.6134**	−1.9693	−1.2393
		(0.1044)	(0.0110)	(0.0982)			(0.2310)	(0.1550)	(0.1087)
Australian	2	−**0.7966**	1.3155	1.8577	Ionosphere	3	**2.8681**	2.9544	3.4988
		(0.0904)	(0.0234)	(0.0263)			(0.0917)	(0.0423)	(0.0776)
	4	−**5.8510**	0.8518	3.0147		10	**4.1625**	4.6447	–
		(0.2947)	(0.0291)	(0.0370)			(0.2150)	(0.4448)	–
	8	−**3.7957**	0.6879	2.6446		17	**3.8920**	5.3236	–
		(0.5823)	(0.1056)	(0.6659)			(0.4198)	(0.9654)	–
	10	−**1.3659**	0.4995	2.2421		24	**2.1412**	4.5570	–
		(0.4382)	(0.1748)	(0.4280)			(0.6710)	(1.3684)	–
Breast-cancer	1	**0.3580**	0.6907	1.3141	Parkinsons	2	−**0.9465**	−0.0847	1.0913
		(0.0561)	(0.0394)	(0.0246)			(0.0402)	(0.0094)	(0.0172)
	3	−**0.5446**	0.1743	0.7889		7	−**5.7700**	−2.1513	0.1867
		(0.1887)	(0.1268)	(0.0626)			(0.1439)	(0.0189)	(0.0538)
	6	−**3.2099**	−1.1397	−0.7526		11	−**10.0932**	−7.8291	−5.6844
		(0.6068)	(0.2788)	(0.4959)			(0.1492)	(0.0340)	(0.0906)
	8	−**6.4362**	−2.1110	−3.1482		15	−**16.9316**	−16.8767	−15.6404
		(0.8144)	(0.3906)	(0.6501)			(0.2151)	(0.1025)	(0.1163)

* The best results are marked in **bold**, and the standard deviation is reported in the parenthesis. The results of MIX on Ionosphere with $d' = 10, 17, 24$ is corrupted due to numerical problems.


thank
you