Orthogonal Gradient Boosting for Interpretable Additive Rule Ensembles
Supplementary Information

A Proofs

A.1 Proof of Lemma 4.1

Proof. Let $\mathbf{f} = [\mathbf{Q}; \mathbf{g}]\alpha$ and $\tilde{\mathbf{f}} = [\mathbf{Q}; q]\boldsymbol{\beta}$. We can decompose the squared error

$$\begin{aligned}
\|\mathbf{f} - \tilde{\mathbf{f}}\|^2 &= \|[\mathbf{Q}; \mathbf{g}]\boldsymbol{\alpha} - [\mathbf{Q}; \mathbf{q}]\boldsymbol{\beta}\|^2 \\
&= \|[\mathbf{Q}; \mathbf{g}_\parallel + \mathbf{g}_\perp]\boldsymbol{\alpha} - [\mathbf{Q}; \mathbf{q}_\parallel + \mathbf{q}_\perp]\boldsymbol{\beta}\|^2 \\
&= \|[\mathbf{Q}; \mathbf{g}_\parallel]\boldsymbol{\alpha} + \alpha_t\mathbf{g}_\perp - [\mathbf{Q}; \mathbf{q}_\parallel]\boldsymbol{\beta} + \beta_t\mathbf{q}_\perp\|^2 \\
&= \|[\mathbf{Q}; \mathbf{g}_\parallel]\boldsymbol{\alpha} - [\mathbf{Q}; \mathbf{q}_\parallel]\boldsymbol{\beta}\|^2 + \|\alpha_t\mathbf{g}_\perp - \beta_t\mathbf{q}_\perp\|^2
\end{aligned}$$

where the last step follows from the Pythagorean theorem and the fact that $\alpha_t\mathbf{g}_\perp - \beta_t\mathbf{q}_\perp$ is an element from the orthogonal complement of $\text{range}[\mathbf{Q}; \mathbf{g}_\parallel] = \text{range}[\mathbf{Q}; \mathbf{q}_\parallel] = \text{range}\,\mathbf{Q}$. The equality of these ranges also implies that $\beta_1, \ldots, \beta_{t-1}$ can, for all choices of β_t, be chosen such that the left term of the error decomposition is 0. Setting $\gamma = \beta_t/\alpha_t$, it follows that

$$\begin{aligned}
\min_{\boldsymbol{\beta}\in\mathbb{R}^t} \|\mathbf{f} - \tilde{\mathbf{f}}\|^2 &= \min_{\beta\in\mathbb{R}^t} \|\alpha_t\mathbf{g}_\perp - \beta_t\mathbf{h}_\perp\|^2 \\
&= \min_{\gamma\in\mathbb{R}^t} \alpha_t^2\|\mathbf{g}_\perp - \gamma\mathbf{q}_\perp\|^2 \\
&= \min_{\gamma\in\mathbb{R}^t} \alpha_t^2(\|\mathbf{g}_\perp\|^2 - 2\gamma\mathbf{q}_\perp^T\mathbf{g}_\perp + \gamma^2\|\mathbf{q}_\perp\|^2)
\end{aligned}$$

and plugging in the minimizing $\gamma = \mathbf{q}_\perp^T\mathbf{g}_\perp/\|\mathbf{q}_\perp\|^2$

$$= \alpha^2(\|\mathbf{g}_\perp\| - (\mathbf{g}_\perp^T\mathbf{q}_\perp)^2/\|\mathbf{q}_\perp\|^2) \ ,$$

from which, noting that $\mathbf{g}_\perp^T\mathbf{q}_\perp = \mathbf{g}_\perp^T\mathbf{q}$, the claim follows. □

A.2 Proof of Lemma 4.2

Proof. After the weight correction step $\boldsymbol{\beta}$ is a stationary point of $R(\mathbf{Q}(\cdot))$, i.e., we have for all $j \in [t]$

$$0 = \frac{\partial R(\mathbf{Q}\boldsymbol{\beta})}{\partial\beta_j} = \sum_{i=1}^n \frac{\partial\, l(\tilde{\mathbf{q}}_i^T\boldsymbol{\beta}, y_i)}{\partial\beta_j} = \sum_{i=1}^n q_{ij} \underbrace{\frac{\partial\, l(\tilde{\mathbf{q}}_i^T\boldsymbol{\beta}, y_i)}{\partial\, \tilde{\mathbf{q}}_i^T\boldsymbol{\beta}}}_{g_i} = \mathbf{q}_j^T\mathbf{g} \ .$$

□

A.3 Proof of Theorem 4.3

The condition of Theorem 4.3 states that:

Let $\mathbf{Q} \in \mathbb{R}^{n\times(t-1)}$ be the selected query matrix, $\mathbf{g}$ the corresponding gradient vector after a full weight correction, and $\mathbf{q}^*$ be a maximizer of the **orthogonal gradient boosting objective** function defined by

$$\mathrm{obj}_{\mathrm{ogb}}(q) = \frac{|\mathbf{g}^T\mathbf{q}|}{\|\mathbf{q}_\perp\| + \epsilon}$$

where $\mathbf{q}_\perp$ is the projection of q onto the orthogonal complement of range $\mathbf{Q}$.

A.3.1 Property a

Proposition A.1. For $\epsilon \to 0$, $[\mathbf{q}_1, \ldots, \mathbf{q}_{t-1}, \mathbf{q}^*]$ is the best approximation to $[\mathbf{q}_1, \ldots, \mathbf{q}_{t-1}, \mathbf{g}]$.

Proof. If $\epsilon \to 0$, then $\mathrm{obj}_{\mathrm{ogb}}(q) \to \frac{|g^T q|}{\|q_\perp\|}$. If $\mathbf{q}^*$ is a maximizer of $\mathrm{obj}_{\mathrm{ogb}}$, then as shown in Lemma 4.1, $\mathbf{q}^*$ minimises the minimum distance from all

$$\mathbf{f} \in \mathrm{span}\{\mathbf{q}_1, \cdots, \mathbf{q}_{t-1}, \mathbf{g}\}$$

to the subspace of

$$\mathrm{span}\{\mathbf{q}_1, \cdots, \mathbf{q}_{t-1}, \mathbf{q}^*\}.$$

Therefore, the subspace spanned by $[\mathbf{q}_1, \cdots, \mathbf{q}_{t-1}, \mathbf{q}^*]$ is the best approximation to the subspace spanned by $[\mathbf{q}_1, \cdots, \mathbf{q}_{t-1}, \mathbf{g}]$. □

A.3.2 Property b

Proposition A.2. For $\epsilon \to \infty$, $\mathbf{q}^*$ is also a maximizer of $\mathrm{obj}_{\mathrm{gs}}$ and any maximizer of $\mathrm{obj}_{\mathrm{gs}}$ is also a maximizer of $\mathrm{obj}_{\mathrm{ogb}}$.

Proof. Let q_1 and q_2 be any two queries and denote by $\mathrm{obj}^{(\epsilon)}_{\mathrm{ogb}}(q)$ the $\mathrm{obj}_{\mathrm{ogb}}$-value of q for a specific ϵ. Then

$$\begin{aligned}
&\lim_{\epsilon\to\infty} \epsilon \left(\mathrm{obj}^{(\epsilon)}_{\mathrm{ogb}}(q_1) - \mathrm{obj}^{(\epsilon)}_{\mathrm{ogb}}(q_2)\right) \\
=&\lim_{\epsilon\to\infty} \epsilon \left(\frac{|g^T\mathbf{q}_1|}{\|\mathbf{q}_1^\perp\| + \epsilon} - \frac{|g^T\mathbf{q}_2|}{\|\mathbf{q}_2^\perp\| + \epsilon}\right) \\
=&\lim_{\epsilon\to\infty} \left(\frac{|g^T\mathbf{q}_1|}{\|\mathbf{q}_1^\perp\|/\epsilon + 1} - \frac{|g^T\mathbf{q}_2|}{\|\mathbf{q}_2^\perp\|/\epsilon + 1}\right) \\
=&|g^T\mathbf{q}_1| - |g^T\mathbf{q}_2| \\
=&\mathrm{obj}_{\mathrm{gs}}(q_1) - \mathrm{obj}_{\mathrm{gs}}(q_2)
\end{aligned}$$

Thus for large enough ϵ, the signs of $\mathrm{obj}^{(\epsilon)}_{\mathrm{ogb}}(q_1) - \mathrm{obj}^{(\epsilon)}_{\mathrm{ogb}}(q_2)$ and $\mathrm{obj}_{\mathrm{gs}}(q_1) - \mathrm{obj}_{\mathrm{gs}}(q_2)$ agree. Therefore, a query q is a $\mathrm{obj}_{\mathrm{gs}}$-maximizer, i.e., $\mathrm{obj}_{\mathrm{gs}}(q) \geq \mathrm{obj}_{\mathrm{gs}}(q')$ for all $q' \in \mathcal{Q}$, if and only if q is a $\mathrm{obj}_{\mathrm{ogb}}$-maximizer, i.e., $\mathrm{obj}_{\mathrm{ogb}}(q) \geq \mathrm{obj}_{\mathrm{ogb}}(q')$ for all $q' \in \mathcal{Q}$. □

A.3.3 Property c

Proposition A.3. For $\epsilon = 0$ and $\|\mathbf{q}_\perp\| > 0$, the ratio $(\frac{\mathrm{obj}_{\mathrm{ogb}}(q)}{\mathrm{obj}_{\mathrm{gb}}(q)})^2$ is equal to $1 + (\frac{\|\mathbf{q}_\parallel\|}{\|\mathbf{q}_\perp\|})^2$.

Proof. If $\epsilon = 0$ and $\|q_\perp\| > 0$, then

$$\left(\frac{\mathrm{obj}_{\mathrm{ogb}}(q)}{\mathrm{obj}_{\mathrm{gb}}(q)}\right)^2 = \frac{\frac{|\mathbf{g}^T\mathbf{q}|^2}{\|\mathbf{q}_\perp\|^2}}{\frac{|\mathbf{g}^T\mathbf{q}|^2}{\|\mathbf{q}\|^2}} = \frac{\|\mathbf{q}\|^2}{\|\mathbf{q}_\perp\|^2}$$
$$= \frac{\|\mathbf{q}_\parallel\|^2 + \|\mathbf{q}_\perp\|^2}{\|\mathbf{q}_\perp\|^2}$$
$$= 1 + \left(\frac{\|\mathbf{q}_\parallel\|}{\|\mathbf{q}_\perp\|}\right)^2$$

□

A.3.4 Property d

Proposition A.4. The objective value $\mathrm{obj}_{\mathrm{ogb}}(q)$ is upper bounded by $\|\mathbf{g}\|$.

Proof. If we divide the numerator and denominator of $\mathrm{obj}_{\mathrm{ogb}}(\mathbf{q})$ with $\|\mathbf{q}_bot\|$, then we can get

$$\mathrm{obj}_{\mathrm{ogb}}(\mathbf{q}) = \frac{|\mathbf{g}^T\mathbf{q}|}{\|q_\perp\| + \epsilon}$$
$$= \frac{\frac{|\mathbf{g}^T\mathbf{q}_\perp|}{\|\mathbf{q}_\perp\|}}{1 + \frac{\epsilon}{\|\mathbf{q}_\perp\|}}$$

according to the Cauchy–Schwarz inequality, $\frac{|\mathbf{g}^T\mathbf{q}|}{\|\mathbf{q}_\perp\|} \leq \frac{\|\mathbf{g}\|\|\mathbf{q}_\perp\|}{\|\mathbf{q}_\perp\|} = \|\mathbf{g}\|$, so,

$$\mathrm{obj}_{\mathrm{ogb}}(\mathbf{q}) \leq \frac{\|\mathbf{g}\|}{1 + \frac{\epsilon}{\|\mathbf{q}_\perp\|}}$$

as $\|\mathbf{q}_\perp\|$ is upper bounded by the number of data points n,

$$\mathrm{obj}_{\mathrm{ogb}}(\mathbf{q}) \leq \frac{\|\mathbf{g}\|}{1 + \frac{\epsilon}{n}}$$
$$\mathrm{obj}_{\mathrm{ogb}}(\mathbf{q}) \leq \|\mathbf{g}\|.$$

□

A.4 Proof of Theorem 4.4

Proof. To see the claim, we first rewrite the objective value for the i-th prefix as

$$\frac{\mathbf{g}^T\mathbf{q}^{(i)}}{\|\mathbf{q}_\perp^{(i)}\| + \epsilon} = \frac{\mathbf{g}^T\mathbf{q}^{(i)}}{\|\mathbf{q}^{(i)}\| - \|\mathbf{q}_\parallel^{(i)}\| + \epsilon} \ .$$

The value of $\|\mathbf{q}^{(i)}\|$ is trivially given as $\sqrt{i}$, and $\mathbf{g}^T\mathbf{q}^i$ can be easily computed for all $i \in [l]$ in time $O(n)$ via cumulative summation. Finally we can reduce the problem of computing the (squared) norms of the l projected

prefixes to computing the t (squared) norms of the prefixes on the subspaces given by the individual orthonormal basis vectors via

$$\|\mathbf{q}_{\|}^{(i)}\|^2 = \left\|\sum_{k=1}^{t} \mathbf{o}_k \mathbf{o}_k^T \mathbf{q}^{(i)}\right\|^2 = \sum_{k=1}^{t} \|\mathbf{o}_k \mathbf{o}_k^T \mathbf{q}^{(i)}\|^2 \ .$$

Each of these t sequences of (squared) norms can be computed in time $O(n)$ by rewriting

$$\begin{aligned}\|\mathbf{o}_k \mathbf{o}_k^T \mathbf{q}^{(i)}\| &= \left\|\mathbf{o}_k \mathbf{o}_k^T \left(\sum_{j=1}^{i} \mathbf{e}_{\sigma(j)}\right)\right\| \\ &= \|\mathbf{o}_k\| \sum_{j=1}^{i} \mathbf{o}_k^T \mathbf{e}_{\sigma(j)} \\ &= \sum_{j=1}^{i} o_{k,\sigma(j)}\end{aligned}$$

where the last equality shows how an $O(n)$-computation is achieved via cumulative summation of the k-th basis vector elements in the order given by σ. □

B Greedy approximation to bounding function

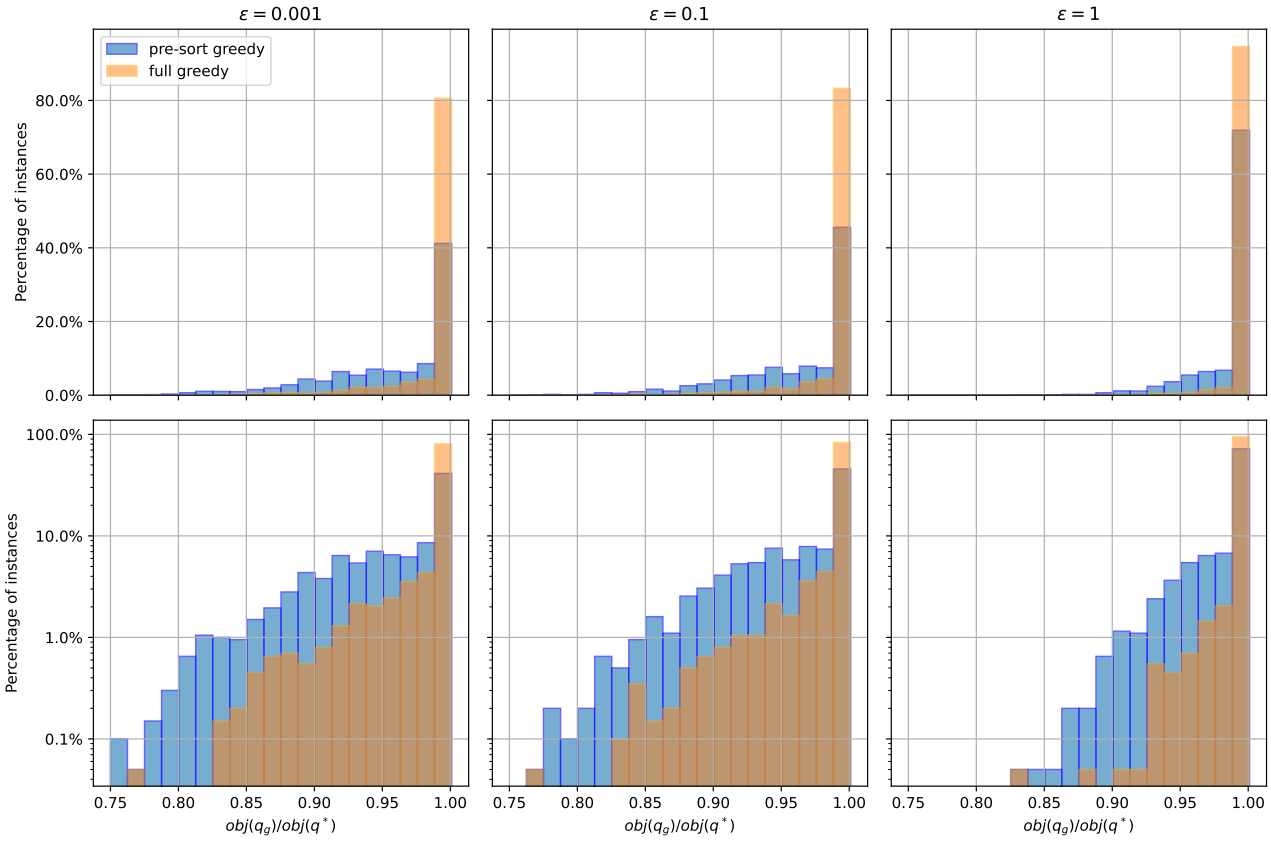

Figure 6: The number of instances of ratios between the best objective values obtained from the greedy search and the true optimal objective value. The upper figures are in linear scales and the lower figures are in log scales. The total variation distances for these three values of ϵ are 0.394, 0.377 and 0.227.

The branch-and-bound search described in Section 3.3 requires an efficient way of calculating the value of $\text{bnd}(\mathbf{q}) = \max\{obj(\mathbf{q}') : \mathbf{q}' \leq \mathbf{q}, \mathbf{q}' \in \{0,1\}^n\}$. It is too expensive to enumerate all possible $\mathbf{q}'$s as there are 2^n cases in the worst case. One solution to this problem is that we can relax the constraint $\mathbf{q}' \in \{0,1\}^n$ to $\mathbf{q}' \in [0,1]^n$ and it can be solved by quadratic programming. However, this relaxation is too loose and inefficient. Instead, we consider relaxing the admission constraint and solve the problem using greedy algorithms.

A full greedy approach can be used to approximate the maximum objective value of the subset of data points selected by $\mathbf{q}$, which is the bounding value $\text{bnd}(\mathbf{q})$. Given a query $\mathbf{q}'^{(t-1)} \leq \mathbf{q}$, we need to find the data point selected by $\mathbf{q}$ which maximise the objective function, and use it with $\mathbf{q}'^{(t-1)}$ to form a $\mathbf{q}'(t)$.

$$i_*^{(t)} = \underset{i \in I(q) - I(q'^{(t-1)})}{\arg\max} \frac{\mathbf{g}^T \left(\mathbf{q}'^{(t-1)} + \mathbf{e}_i\right)}{\left\| \left(\mathbf{q}'^{(t-1)} + \mathbf{e}_i\right)_\perp \right\| + \epsilon}.$$

where $I(\mathbf{q}) = \{i : \mathbf{q}(x_i) = 1, 1 \leq i \leq n\}$, $0 \leq t \leq |I(\mathbf{q})|$, $\mathbf{q}'^{(0)} = \mathbf{0}$ and $\mathbf{q}'^{(t)} = \mathbf{q}'^{(t-1)} + \mathbf{e}_{i_*^{(t)}}$. We use the maximum value of $\text{obj}(\mathbf{q}'^{(t)})$ as the bounding value for query $\mathbf{q}$. The computation time complexity level of this approach is $O(n^2)$ for each query, which is not as efficient as the presorting greedy approach described in Section 4.3.

The presorting greedy approach of solving the prefix optimization problem described in Section 4.3 leads to another approximation to the optimal objective function value for the queries which cover subsets of data points covered by $\mathbf{q}$. As proved in Theorem 4.4, this approach has a time complexity of $O(tn)$.

We test 2000 instances for different initial queries and initial gradient values to see the difference between the approximation of $\text{bnd}(\mathbf{q})$ obtained by the full greedy approach, the pre-sorting greedy approach, and the actual optimal objective values (obtained by a brute-force approach). We choose three different values of ϵ: 0.001, 0.1 and 1.

Figure 6 compares the ratio between the approximations to $\text{bnd}(\mathbf{q})$ obtained by the two greedy approaches and the true optimal objective value. The Y axis of Figure 6 represents the percentage of instances of different ratios.

According to the comparison, the full greedy approach can approximate the true bounding function better than the presorting greedy approach. For smaller ϵ values ($\epsilon = 0.001$), there are 90% instances whose approximation values are more than 90% of the true bounding function values, while 96% of instances approximate more than 90% of the value of $\text{bnd}(\mathbf{q})$ using the full greedy approach. For $\epsilon = 0.1$, both algorithms have slight better (both 1% promotion) approximation than $\epsilon = 0.001$. It can be observed that for $\epsilon = 1$, both algorithms have more instances where the approximations are closed to the true bounding values. However, if the value of ϵ is too large, then the calculated objective values cannot be accurate according to Theorem 4.3. Comparing the statistical distances of these two greedy approaches, it is reasonable to use the presorting greedy approach to approximate the bounding values.

To approximate the true bounding function more efficiently and more accurate, we adopt the presorting greedy approach in this research.

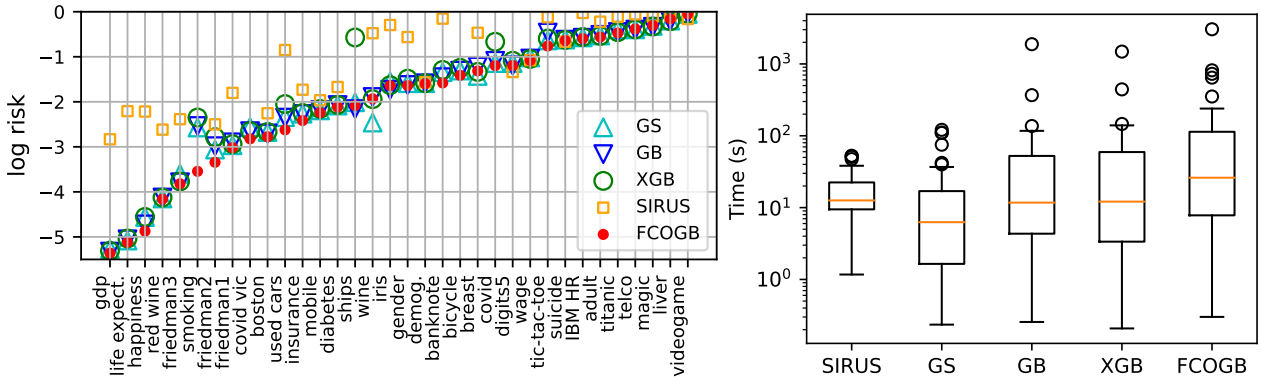

Figure 7: Left: comparison of log training risks over different datasets. Right: run times of different algorithms.

Table 2: Comparison of normalised risks and computation times for rule ensembles, averaged over cognitive complexities between 1 and 50, using SIRUS(SRS), Gradient Sum(GS), Gradient boosting (GB), XGBoost (XGB) and FCOGB (CB), for benchmark datasets of classification (upper), regression (middle) and Poisson regression problems (lower).

Dataset	d	n	Train risks						Test risks						Computation times				
			SRS	GS	GB	XGB	CB_{Gr}	CB_{BB}	SRS	GS	GB	XGB	CB_{Gr}	CB_{BB}	SRS	GS	GB	XGB	FCOGB
titanic	7	1043	.895	.653	.656	.646	.639	**.621**	.894	.695	.707	.711	.713	**.695**	7.077	2.624	9.858	10.21	25.71
tic-tac-toe	27	958	.892	.555	.641	.577	.650	**.506**	.885	.596	.682	.629	.721	**.577**	12.59	3.971	10.34	6.09	13.99
iris	4	150	.685	.261	.261	.332	**.192**	.251	.745	.521	.440	.459	.531	**.424**	11.02	0.775	1.099	1.453	2.487
breast	30	569	.569	**.277**	.310	.314	.290	.304	.627	**.269**	.362	.338	.383	.349	11.48	6.744	74.43	74.83	239.2
wine	13	178	.578	.216	.250	.192	.191	**.183**	.621	.346	.431	.409	.483	**.250**	9.456	1.530	4.432	2.154	55.183
ibm hr	32	1470	.980	.567	.560	.571	.558	**.553**	.974	.640	.645	.636	.652	**.617**	11.15	17.24	10.99	12.92	12.03
telco churn	18	7043	.944	.679	.682	.678	**.664**	**.667**	.945	.663	.677	.665	**.650**	.653	50.83	40.01	1883	1485	3039
gender	20	3168	.566	.230	.230	.249	.224	**.224**	.570	**.243**	.247	.263	.246	.246	22.42	22.73	25.49	24.27	32.95
banknote	4	1372	.854	.304	.267	.290	.253	**.228**	.858	.311	.268	.299	.264	**.229**	8.933	6.298	5.648	7.060	8.444
liver	6	345	.908	.815	.834	.814	**.802**	.854	.917	.879	.927	**.873**	.940	.924	9.734	1.997	99.72	124.1	193.9
magic	10	19020	.906	.718	.708	.710	.707	**.707**	.903	.698	.693	.693	.688	**.687**	1.364	75.14	89.18	101.9	352.2
adult	11	30162	.804	.594	.599	.588	**.575**	.576	.802	.603	.615	.601	**.589**	.590	2.169	121.0	136.7	146.0	728.3
digits5	64	3915	**.248**	.332	.312	.344	.329	.315	**.262**	.329	.314	.341	.320	.317	52.60	110.8	72.74	101.5	97.4
insurance	6	1338	.169	.130	.142	.144	**.120**	.127	.177	.132	.145	.147	**.127**	.131	14.06	7.507	15.94	12.98	39.53
friedman1	10	2000	.180	.089	.074	.068	.067	**.065**	.165	.091	.077	.075	**.070**	.072	16.79	2.514	4.302	3.171	6.915
friedman2	4	10000	.082	.133	.119	.115	**.077**	.099	.082	.135	.120	.115	**.077**	.102	47.33	11.79	17.56	13.18	28.4
friedman3	4	5000	.093	.045	.042	.042	.041	**.041**	.092	.048	.047	.046	.045	**.045**	29.86	6.243	10.61	8.559	17.65
wage	5	1379	.427	.370	.362	.359	.352	**.348**	**.341**	.358	.405	.411	.368	.365	14.18	5.605	12.12	13.17	25.19
demographics	13	6876	.219	.214	.214	.214	.212	**.212**	**.209**	.216	.217	.217	.214	.215	38.24	36.80	29.40	33.04	72.42
gdp	1	35	.063	.020	.020	.020	.024	**.020**	.059	.020	.020	.020	.027	**.020**	7.974	.261	.351	.282	.488
used cars	4	1770	.373	.130	.204	.153	**.113**	.132	.427	.101	.157	.100	.116	**.082**	15.00	8.371	12.10	9.484	20.27
diabetes	10	442	.156	.138	.142	.139	.132	**.132**	.188	.141	.141	.147	**.136**	.148	10.50	2.204	3.574	3.920	7.591
boston	13	506	.101	.086	.087	.086	**.080**	.081	.105	**.079**	.087	.089	.088	.087	10.96	3.055	6.731	5.285	10.44
happiness	8	315	.109	.031	.031	.031	.029	**.029**	.109	.033	.039	.039	.035	**.033**	6.344	1.160	11.37	11.31	26.43
life expect.	21	1649	.109	.026	.026	.026	.026	**.026**	.110	.027	.027	.027	.026	**.026**	21.44	16.16	58.43	63.82	131.2
mobile prices	20	2000	.148	.131	.137	.137	**.122**	.134	.140	.134	.143	.143	**.126**	.139	33.81	15.03	367.7	442.5	815.4
suicide rate	5	27820	.547	.543	.540	.540	.540	**.531**	.514	.521	.521	.521	.519	**.509**	52.35	109.6	117.1	139.6	644.6
videogame	6	16327	**.895**	.953	.953	.953	.953	.953	.850	**.720**	**.720**	**.720**	**.720**	**.720**	1.171	41.91	34.38	45.90	119.1
red wine	11	1599	.072	.034	.035	.034	**.034**	**.034**	.073	.035	.036	.036	.035	**.035**	19.94	9.149	15.32	21.99	35.34
covid vic	4	85	NA	.153	.121	.132	.105	**.105**	NA	.182	**.100**	.133	.104	.104	NA	.523	.600	.628	.854
covid	2	225	NA	.344	.371	.891	.343	**.335**	NA	.459	.411	.741	.417	**.407**	NA	.701	.690	.682	1.143
bicycle	4	122	NA	.317	.324	.337	.296	**.295**	NA	.366	.478	.457	.529	**.328**	NA	.695	1.103	1.105	2.124
ships	4	34	NA	.174	.181	.146	**.125**	.186	NA	.197	.203	.199	.464	**.172**	NA	.235	.296	.311	.448
smoking	2	36	NA	.127	.128	.163	**.078**	.091	NA	.136	.250	.322	.121	**.103**	NA	.266	.256	.208	.301

C Experiments configurations

The experiments are conducted on a computer with CPU 'Intel(R) Core(TM) i5-10300H CPU @ 2.50GHz' and memory of 24G.

D Additional experiment results

Table 2 shows the normalised average training risk, test risk and computation time of the rule ensembles generated by SIRUS, Gradient Sum, Gradient Boosting, XGBoost FCOGB using greedy search and FCOGB using branch-and-bound search over complexity levels from 1 to 50 for the 34 datasets used in the experiments of this paper. In the table, we bold the lowest training and test risks for each dataset. The red colours indicate the FCOGB approach using greedy search or branch-and-bound search have lower risks than all the other methods.

For the FCOGB with greedy search, there are 29 out of 34 datasets whose training risks are lower than the other methods. However, it has only 15 out of 34 datasets whose test risks are lower than other methods. Therefore, using branch-and-bound search generates better rule ensembles than greedy search. The One-sided T-test at significance level 0.05 also shows the same results.

Left of Figure 7 compares the normalised average logged training risks over complexity levels from 1 to 50 for different datasets. According to Fig. 7 and Table 2, FCOGB generates lower training risks than the other algorithms for 26 out of 34 datasets.

The right part of Figure 7 shows the box plot of running time of different algorithms on the 34 datasets. Although the overall running time of FCOGB is higher than the other methods, they are still at the same scale.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]