Orthogonal Gradient Boosting for Interpretable Additive Rule Ensembles Supplementary Information

Anonymous Author(s)
Affiliation
Address
email

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

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

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 $\mathrm{range}[\mathbf{Q}; \mathbf{g}_{\|}] = \mathrm{range}[\mathbf{Q}; g_{\|}] = \mathrm{range}\, \mathbf{Q}$. 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 $\mathrm{range}\,\mathbf{Q}$.

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 $\mathrm{obj}_{\mathrm{ogb}}(q) \to \dfrac{|g^T q|}{\|q_\perp\|}$. 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

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

to the subspace of

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

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

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 $q' \in \mathcal{Q}$. $\square$

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

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

□

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 $\|\mathbf{q}_b ot\|$, 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, $\frac{|\mathbf{g}^T\mathbf{q}|}{\|\mathbf{q}_\perp\|} \le \frac{\|\mathbf{g}\|\|\mathbf{q}_\perp\|}{\|\mathbf{q}_\perp\|} = \|\mathbf{g}\|$, so,

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

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

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

□

A.4 Proof of Theorem 4.4

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

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

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. $\square$

B Greedy approximation to bounding function

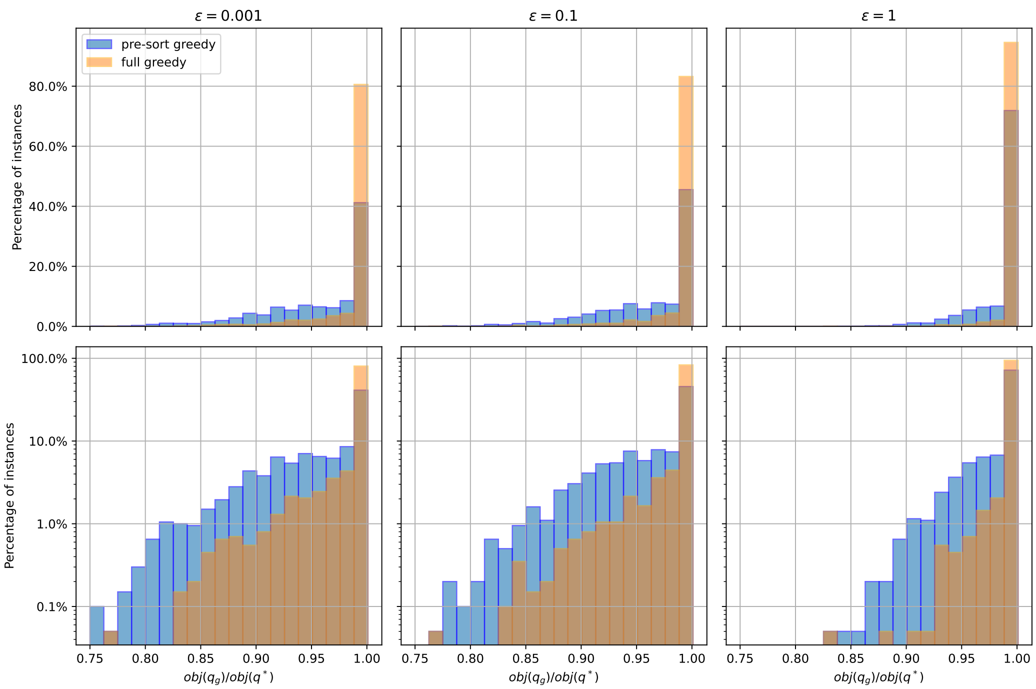

Figure 6: The number of instances of ratios between the best objective values obtained from the greedy search and the true optimal objective value. The upper figures are in linear scales and the lower figures are in $\log$ scales. The total variation distances for these three values of ϵ are 0.394, 0.377 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 $\mathbf{q}'(t)$.

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

where $I(\mathbf{q}) = \{i : \mathbf{q}(x_i) = 1, 1 \le i \le n\}$, $0 \le t \le |I(\mathbf{q})|$, $\mathbf{q}'^{(0)} = \mathbf{0}$ and $\mathbf{q}'^{(t)} = \mathbf{q}'^{(t-1)} + \mathbf{e}_{i_*^{(t)}}$. We use the maximum value of $\mathrm{obj}(\mathbf{q}'^{(t)})$ 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 $\mathrm{bnd}(\mathbf{q})$ 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 $\mathrm{bnd}(\mathbf{q})$ 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 $\mathrm{bnd}(\mathbf{q})$ 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.

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.