On Uncertainty Estimation by Tree-based Surrogate Models in Sequential Model-based Optimization

Jungtaek Kim

jtkim@postech.ac.kr

POSTECH
Pohang 37673, Republic of Korea
https://jungtaek.github.io

A joint work with Seungjin Choi

AISTATS 2022

Table of Contents

Introduction

- In sequential model-based optimization, GP regression [Rasmussen and Williams, 2006] is a popular choice as a surrogate model, because of its capability of calculating uncertainties analytically.
- On the other hand, an ensemble of randomized trees is another option and has practical merits over GPs due to its scalability and easiness of handling continuous/discrete mixed variables [Hutter et al., 2011].
- We revisit various ensembles of randomized trees to investigate their behavior in the perspective of prediction uncertainty estimation.
- Then, we propose our own tree-based model, referred to as *BwO forest*.

Uncertainty Estimation by Tree-based Models

- **Sum-of-Trees Model**: Posterior mean and variance functions are

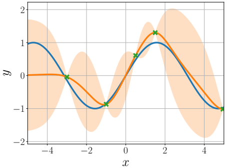
$$\mu\left(\mathbf{x}; \{\mathcal{T}_b\}_{b=1}^{B}, \mathbf{X}, \mathbf{y}\right) = \frac{1}{B} \sum_{b=1}^{B} \sum_{\tau \in \boldsymbol{\tau}_{b,l}} \mu_\tau 1_{\mathbf{x} \in \tau}, \tag{1}$$

$$\begin{aligned} &\sigma^2\left(\mathbf{x}; \{\mathcal{T}_b\}_{b=1}^{B}, \mathbf{X}, \mathbf{y}\right) \\ &= \frac{1}{B} \sum_{b=1}^{B} \left(\Big(\sum_{\tau \in \boldsymbol{\tau}_{b,l}} \sigma_\tau 1_{\mathbf{x} \in \tau} \Big)^2 + \Big(\sum_{\tau \in \boldsymbol{\tau}_{b,l}} \mu_\tau 1_{\mathbf{x} \in \tau} \Big)^2 \right) - \mu\left(\mathbf{x}; \{\mathcal{T}_b\}_{b=1}^{B}, \mathbf{X}, \mathbf{y}\right)^2, \end{aligned} \tag{2}$$

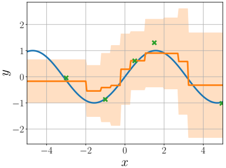
  by the law of total variance, as described in [Hutter et al., 2014].

- **Gradient Boosting Models**: It updates parameters $\boldsymbol{\theta}$ using their gradients in terms of the objective of parametric distribution, e.g., likelihood function.

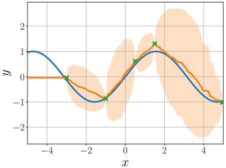
Uncertainty Estimation by Tree-based Models



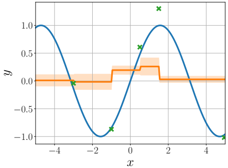
(a) Gaussian process



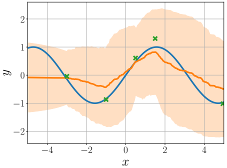
(b) Random forest



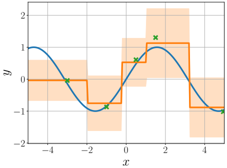
(c) Extremely randomized trees



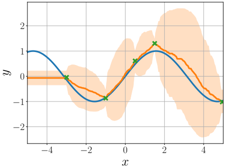
(d) BART



(e) Mondrian forest



(f) NGBoost



(g) BwO forest

Figure 1: Results with GP regression and tree-based models such as random forest, extremely randomized trees, BART, Mondrian forest, NGBoost, and BwO forest (ours).

Tree Construction

- The uncertainty of an ensemble is derived from the randomness of individual trees:
  (i) **bagging** [Breiman, 1996]: it samples a bootstrap sample with replacement and then aggregates base estimators;
  (ii) **random feature selection**: this technique randomly selects a feature from a set of dimensions;
  (iii) **random split location**: it randomly selects a split location between lower and upper bounds of the selected dimension;
  (iv) **random tree sampling**: this strategy randomly samples a tree under the assumption on a prior distribution.

- Random forest [Breiman, 2001] employs (i) and (ii), extremely randomized trees [Geurts et al., 2006] employs (ii) and (iii), BART [Chipman et al., 2010] employs (i), (ii), and (iv), and Mondrian forest [Lakshminarayanan et al., 2016] employs (i) and (iii).

- On the contrary, NGBoost [Duan et al., 2020] is defined as a gradient boosting model.

Elaborating Uncertainty Estimation by Tree-based Models

- As pointed out in the work [Mendelson et al., 2016], the expectation and variance of an indicator for the existence of $\mathbf{x}_i$ in a bootstrap sample $\mathbf{B}$ are expressed as

$$\mathbb{E}[1_{\mathbf{x}_i \in \mathbf{B}}] = 1 - \left(1 - \frac{1}{N}\right)^{M}, \tag{3}$$

$$\mathrm{Var}[1_{\mathbf{x}_i \in \mathbf{B}}] = \left(1 - \frac{1}{N}\right)^{M} - \left(1 - \frac{1}{N}\right)^{2M}, \tag{4}$$

where N is the size of $\mathbf{X}$ and M is the size of a bootstrap sample.

- The distribution of unique original elements in a bootstrap sample $\mathbf{B}$ is:

$$\mathbb{E}[|\mathrm{unique}(\mathbf{B})|] = N - \frac{(N-1)^M}{N^{M-1}}, \tag{5}$$

$$\mathrm{Var}[|\mathrm{unique}(\mathbf{B})|] = (N-1)\frac{(N-2)^M}{N^{M-1}} + \frac{(N-1)^M}{N^{M-1}} - \frac{(N-1)^{2M}}{N^{2M-2}}, \tag{6}$$

where $\mathrm{unique}(\mathbf{B})$ filters duplicates.

Elaborating Uncertainty Estimation by Tree-based Models

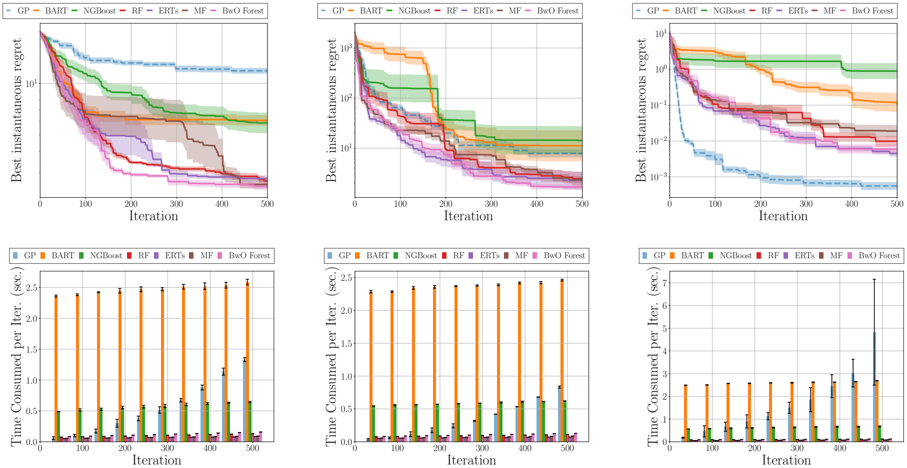
Algorithm 1 Training BwO Forest

Input: Ensemble size B, training data and evaluations $\mathbf{X} \in \mathbb{R}^{N \times d}$ and $\mathbf{y} \in \mathbb{R}^{N}$, size of bootstrap sample $M = \alpha N$ for an oversampling rate $\alpha > 1$.

Output: Set of decision trees $\{\mathcal{T}_b\}_{b=1}^{B}$

Initialize a set of decision trees.
for $b = 1, \ldots, B$ **do**
    Sample a bootstrap sample $\mathbf{B}_b \in \mathbb{R}^{M \times d}$ from $\mathbf{X}$.
    Set a root node $\boldsymbol{\tau}_r$ that contains all the elements in $\mathbf{B}_b$, and $\boldsymbol{\tau}_r$ is set as the current split node.
    Train a decision tree using random feature selection and random split location.
end for
return A set of decision trees $\{\mathcal{T}_b\}_{b=1}^{B}$

Experimental Results



(a) Ackley (4D) (b) Bohachevsky (c) Branin

Figure 2: Experimental results on various benchmark functions.

Thank you!

arXiv

GitHub

References I

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

T. Duan, A. Avati, D. Y. Ding, K. K. Thai, S. Basu, A. Ng, and A. Schuler. NGBoost: Natural gradient boosting for probabilistic prediction. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2690–2700, Virtual, 2020.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *Proceedings of the International Conference on Learning and Intelligent Optimization (LION)*, pages 507–523, Rome, Italy, 2011.

F. Hutter, L. Xu, H. H. Hoos, and K. Leyton-Brown. Algorithm runtime prediction: Methods & evaluation. *Artificial Intelligence*, 206:79–111, 2014.

B. Lakshminarayanan, D. M. Roy, and Y. W. Teh. Mondrian forests for large-scale regression when uncertainty matters. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1478–1487, Cadiz, Spain, 2016.

A. F. Mendelson, M. A. Zuluaga, B. F. Hutton, and S. Ourselin. What is the distribution of the number of unique original items in a bootstrap sample? *arXiv preprint arXiv:1602.05822*, 2016.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.