Universiteit
Leiden
The Netherlands

Bachelor Computer Science & Datascience and Artificial Intelligence

A study of TaBPFN as Surrogate Model in Bayesian Optimization

Ocean Wang

Supervised by: Professor E. Raponi

RESEARCH PROPOSAL BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

19/02/2025

1 Background and theory

Bayesian optimization (BO) is a machine-learning-based approach for optimizing expensive black-box functions that requires a lot of time to evaluate [Fra18]. It has become popular for tuning hyperparameters in machine learning algorithms [SLA12] and has applications across domains like experimental design, simulator calibration, and materials discovery [LRW24].

The basic BO framework uses Gaussian process (GP) regression as a surrogate model, which provides both predictions and uncertainty estimates [RW06]. For any finite collection of points, the function values follow a multivariate normal distribution with a mean function μ_0 and a covariance kernel Σ_0. However, GPs have limitations: they struggle with non-stationary functions, scale poorly to high dimensions, and need careful kernel design [LRW24].

The Tabular Prior-Data Fitted Network (TabPFN) represents a breakthrough approach for tabular learning that could potentially be adapted for surrogate modeling. Traditional methods require training from scratch for each dataset. However, TabPFN uses in-context learning through a single forward pass of a pre-trained transformer [HMEH23, HMP+25]. The original TabPFN showed promise for small classification tasks, but the recently improved version is capable of handling data sets with up to 10,000 samples and 500 features and also supports regression tasks [HMP+25]. TabPFN is able to effectively capture complex feature dependencies because it approximates Bayesian inference by using an advanced prior that combines principles from Structural Causal Models (SCMs) [Pea09] and Bayesian Neural Networks. SCMs are formal frameworks that can represent causal relationships underlying data. The TabPFN model architecture uses a dual-attention mechanism that processes tabular structures efficiently, where one handles feature interactions and the other manages sample relationships. With such a design, TabPFN is able to learn patterns without extensive training, while also outperforming state-of-the-art models tuned for hours [HMP+25]. TabPFN has not yet been specifically applied to surrogate modeling in BO, where computational efficiency and accuracy on limited data are crucial. However, it is a potential candidate, as it has the ability to rapidly generate accurate predictions for regression tasks on small datasets.

2 Research question

The goal of this thesis is to answer the main question: "How effectively does TabPFN perform as a surrogate model for Bayesian Optimization in regression tasks when evaluated on standardized benchmarks?" This research will focus on benchmarking TabPFN within the BBOB test suite of the COCO environment, a standardized framework for gradient-free optimization. The study will evaluate TabPFN's performance across different objective function landscapes to determine if it facilitates efficient optimization. Performance metrics will be compared against existing benchmark data available in IOHprofiler. The research will also analyze the computational efficiency of TabPFN and how it scales with increasing problem dimensionality to provide a comprehensive assessment of its practical use for surrogate modeling in Bayesian Optimization contexts.

3 Study design/method

The research will be done in three phases: implementation, experimentation, and analysis. During implementation, the TaBPFN model will be set up as a surrogate within the Bayesian optimization framework with evaluation metrics. The experimentation phase will use the BBOB test suite within the COCO environment to evaluate the performance of TaBPFN and to measure key metrics. Finally, we will use statistical methods to compare their performance. With such study design, we hope to provide a sufficient evaluation of TabPFN's capabilities as a surrogate model for BO.

4 Global Planning

The thesis should be completed within a timeline of 8 weeks. During the first two weeks, relevant literature will be reviewed to gain a basic understanding of Bayesian optimization and the TaBPFN model we will be benchmarking. The development environment will be setup and the initial implementation of the taBPFN model as surrogate will be executed.

Week 3-4 will be dedicated to implementing the COCO benchmarking environment and statistical metrics necessary for comparison, including testing to ensure that TaBPFN works properly within the BO framework. Preliminary experiments will be conducted using simple test functions like Sphere, Rosenbrock, and Rastrigin functions in low dimensions. Performance metrics will be validated on these functions, tracking optimization progress, computational time and model accuracy. These experiments will serve as a smoke test for the research setup before proceeding to full-scale experiments.

During week 5-6, the research will move into full-scale experimental runs using the complete BBOB test suite within the COCO environment. The TaBPFN surrogate model will be evaluated across optimization problems, collecting automated data on convergence rates, model prediction accuracy, and computational resource usage.

The last 2 weeks will be dedicated to comprehensive analysis of all experimental data and thesis writing. This phase will involve documenting the research process, results, and implications while addressing the original research question. The final days will focus on reviewing and revising the thesis, ensuring accurate details, clearly presented results, and well-supported conclusions.

References

[Fra18] Peter I Frazier. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[HMEH23] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023.

[HMP+25] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025.

[LRW24] Yucen Lily Li, Tim GJ Rudner, and Andrew Gordon Wilson. A study of bayesian neural network surrogates for bayesian optimization. *ICLR*, 2024.

[Pea09] Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

[RW06] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[SLA12] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 25:2951–2959, 2012.