HyperBO+: Pre-training a universal hierarchical Gaussian process prior for Bayesian optimization

Zhou Fan
Harvard University
zfan@g.harvard.edu

Xinran Han
Harvard University
xinranhan@g.harvard.edu

Zi Wang
Google Research, Brain Team
wangzi@google.com

Harvard John A. Paulson School of Engineering and Applied Sciences

Google Research

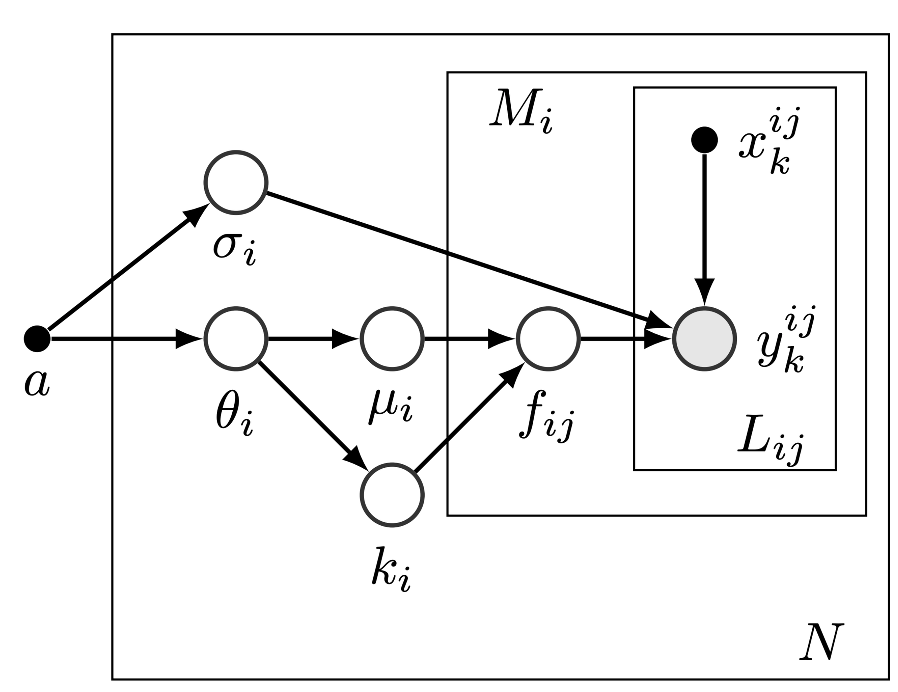
Introduction

- BayesOpt requires expert knowledge to select priors.
- Popular solutions: learn the prior from multi-task data, e.g. multi-task BO (Swersky et al., 2013), few-shot BO (Wistuba and Grabocka, 2021) and HyperBO (Wang et al., 2022).
    - Limitations: input domains must be the same for all tasks.
- We present **HyperBO+**: a pre-training approach for hierarchical Gaussian processes that **enables the same prior to work universally for Bayesian optimization on functions with different domains**.
- Our contributions:
    - two-step pre-training method that **learns a universal hierarchical GP prior**.
    - transfer learning BO framework that **generalizes to unseen search spaces**.
    - analyses on **empirical and theoretical advantages**.

Problem Formulation

Goal: to optimize unseen black-box functions by pre-training on existing data from functions in **multiple search spaces**, which can have **different numbers of dimensions** for their respective domains.

- **Definitions**:
    - ***super-dataset*** D*:* all datapoints collected across multiple search spaces. $D = \{D_i\}_{i=1}^N$.
    - ***dataset*** D_i*:* the data from a single search space, consists of observations on a collection of black-box functions $F_i = \{f_{ij}: \mathcal{X}_i \to \mathbb{R}\}_{j=1}^{M_i}$ where functions in F_i share the same compact search space $\mathcal{X}_i \in \mathbb{R}^{d_i}$. $D_i = \{D_{ij}\}_{j=1}^{M_i}$.
    - ***sub-dataset*** D_{ij}*:* the collection of datapoints from a single function within a search space. $D_{ij} = \{(x_k^{ij}, y_k^{ij})\}_{k=1}^{L_{ij}}$. $L_{ij}$is the number of observations on function f_{ij}.
    - ***observation*** (x_k^{ij}, y_k^{ij})*:* function value perturbed by *i.i.d.* additive Gaussian noise, i.e. $y_k^{ij} \sim \mathcal{N}(f_{ij}(x_k^{ij}), \sigma_i^2)$.



- For each $i = 1, \dots, N$, we assume all functions in F_i are *i.i.d.* function samples from the same GP: $GP_i = GP(\mu_i, k_i)$. For each function set F_i and its corresponding GP_i with mean function $\mu_i: \mathcal{X}_i \to \mathbb{R}$ and kernel $k_i: \mathcal{X}_i \times \mathcal{X}_i \to \mathbb{R}$, we denote the parameters of μ_i, k_i by θ_i.
- **Our major assumption**: The GP parameters and noise standard deviation $\{(\theta_i, \sigma_i)\}_{i=1}^N$ are *i.i.d.* samples from a distribution, as in $(\theta_i, \sigma_i) \sim p((\theta, \sigma); a)$. The distribution $p((\theta, \sigma); a)$ is a ***universal prior*** for all search spaces.

Methodology

The HyperBO+ framework consists of mainly two phases: (1) **Training**: estimate the universal prior a from the super-dataset $D = \{D_i\}_{i=1}^N$ with a two-step approach. (2) **Optimization**: running BO with the hierarchical GP parameterized by the learned a on testing functions.

Two-step pre-training

- **Estimating GP parameters of each search space.** For each function collection F_i with domain $\mathcal{X}_i$, we can infer its GP parameters θ_i and noise standard deviation σ_i by **minimizing the negative log-likelihood** of the dataset as in the original HyperBO:

$$(\hat{\theta}_i, \hat{\sigma}_i)_{ML} = \operatorname{argmin}_{\theta,\sigma} - \sum_{j=1}^{M_i} \log p\left(D_{ij} \mid \theta, \sigma\right).$$

- **Estimate the universal prior.** Using the estimated $\{(\hat{\theta}_i, \hat{\sigma}_i)\}_{i=1}^N$ from all datasets, we can use the the **maximum likelihood estimator** for the universal prior parameter a as $\hat{a} = \operatorname{argmax}_a p(\{(\hat{\theta}_i, \hat{\sigma}_i)\}_{i=1}^N; a)$.

Generalization: Bayesian optimization on an unseen function with new search spaces

HyperBO+ models functions via a hierarchical GP with pre-trained universal prior parameter a.

- At step t of Bayesian optimization, select x to optimize acquisition function:

$$ac_t(x; \hat{a}) = \sum_{r=1}^{R} \left[ac_t(x; \theta_r, \sigma_r) p((x_k, y_k)_{k=1}^t | \theta_r, \sigma_r)\right]$$

- $(\theta_1, \sigma_1), \dots, (\theta_R, \sigma_R)$ are *i.i.d.* samples from the prior distribution $p((\theta, \sigma); \hat{a})$.

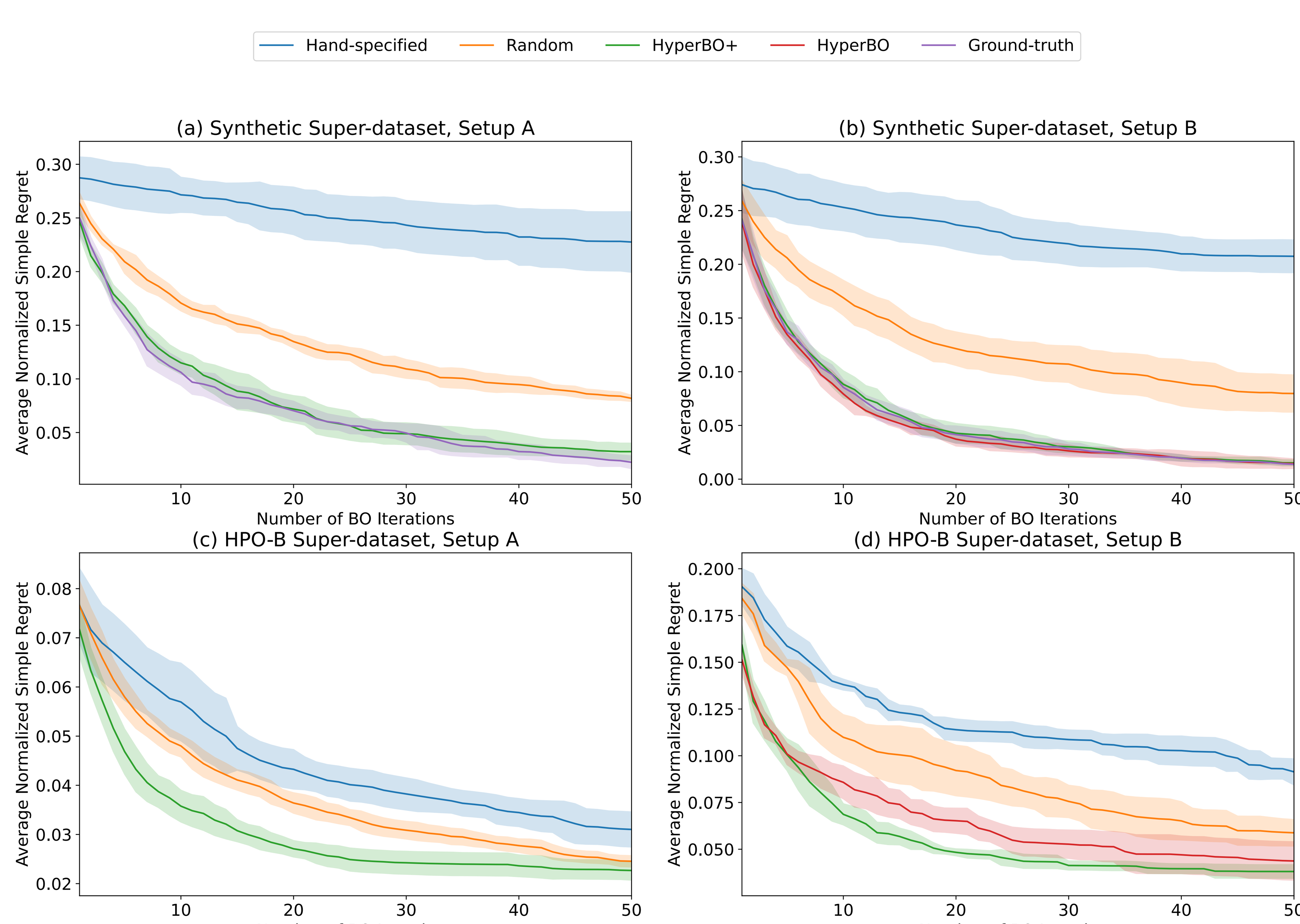
Experiments

We demonstrate the strong performance of HyperBO+ compared to baseline methods on two datasets: a **Synthetic Super-dataset** and **HPO-B Super-dataset** (Pineda-Arango et al., 2021), a collection of real-world hyperparameter tuning tasks that involves multiple search spaces.

Two experiment setups:

- **Setup A** is designed to demonstrate the ability of HyperBO+ to **generalize its learned prior to unseen search spaces.** We split the super-dataset into training datasets and testing datasets.
- **Setup B** aims to to test the ability of HyperBO+ to **generalize to new functions in seen search spaces** and compare its performance with HyperBO. We split each dataset in the super-dataset into training sub-datasets and testing sub-datasets.

Baselines: (1) A hand-specified (and potentially misspecified) hierarchical GP prior, fixed over all search spaces. (2) Random sampling for optimization. (3) HyperBO, only applicable to Setup B. (4) The ground-truth hierarchical GP prior, only available for Synthetic Super-dataset.



References

- Wang, Z., Dahl, G. E., Swersky, K., Lee, C., Mariet, Z., Nado, Z., Gilmer, J., Snoek, J., and Ghahramani, Z. (2022). Pre-trained Gaussian processes for Bayesian optimization. arXiv preprint arXiv:2109.08215.
- Pineda-Arango, S., Jomaa, H. S., Wistuba, M., and Grabocka, J. (2021). HPO-B: A large-scale reproducible benchmark for black-box HPO based on openml. Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks.