universität freiburg

ICML 2024, Vienna

In-context Freeze-Thaw Bayesian Optimization for Hyperparameter Optimization

Herilalaina Rakotoarison*, Steven Adriaensen*, Neeratyoy Mallik*,
Samir Garibov, Edward Bergman, Frank Hutter

* Equal Contribution

Efficient Hyperparameter Optimization

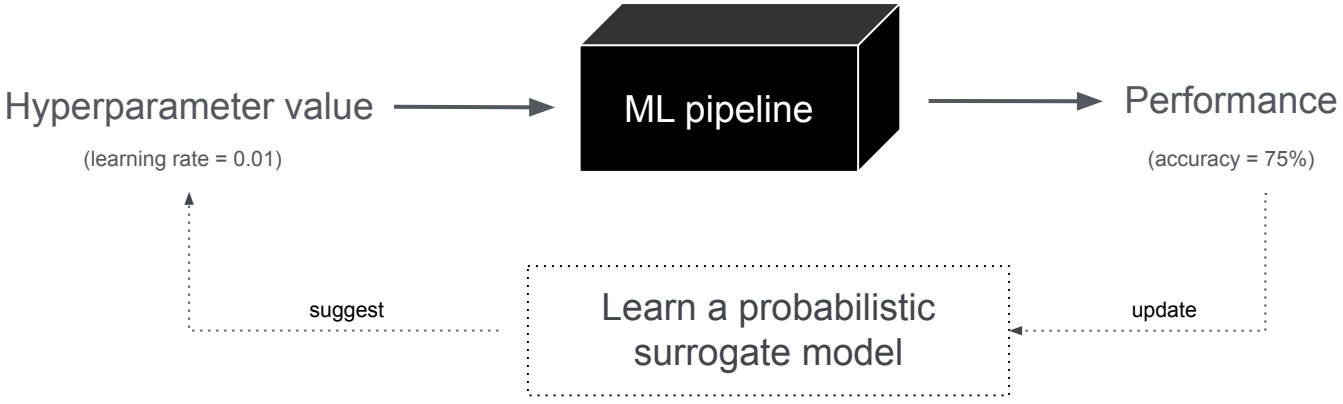
Hyperparameter Optimization: Find the best hyperparameter value for your Machine Learning (ML) pipeline.

Manual tuning or grid search: costly, time-consuming, error-prone.

Efficient Hyperparameter Optimization

Hyperparameter Optimization: Find the best hyperparameter value for your Machine Learning (ML) pipeline.

Bayesian Optimization



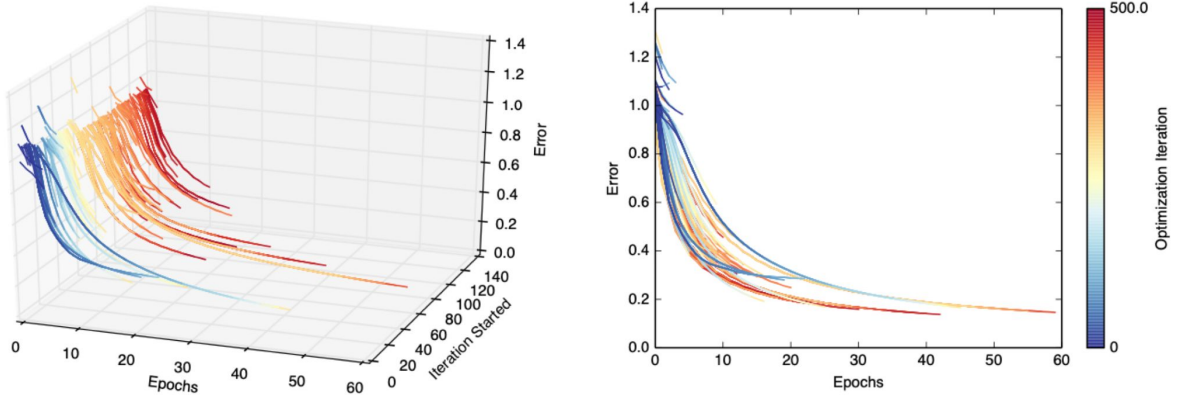
Not feasible for expensive applications (e.g. large models)

Efficient Hyperparameter Optimization

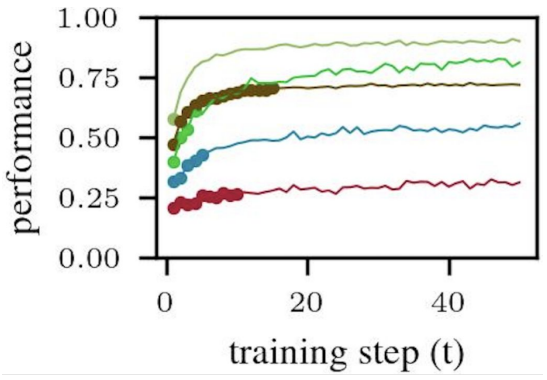
Hyperparameter Optimization: Find the best hyperparameter value for your Machine Learning (ML) pipeline.

Bayesian Optimization: Not feasible for expensive applications (e.g. large models)

Freeze-thaw Bayesian Optimization: pause and resume runs (*Swersky et al, 2014*)



Freeze-Thaw Bayesian Optimization



Challenges

- Predict performance for higher budget (surrogate model)
- Find the most promising run to continue/start (acquisition function)

(Recent) prior works	surrogate model	acquisition function
DyHPO (Wistuba et al, 2022)	Deep Gaussian Process	Expected Improvement at the next training step
DPL (Kadra et al, 2023)	Ensemble of Deep Power Laws	Expected Improvement at the max training step

require online training of the surrogate
- ➔ **computational overhead**
- ➔ **additional hyper-hyperparameters**
- ➔ **training instabilities**

can yield suboptimal decisions
- ➔ **optimal setting depends on the shape of the learning curves (e.g crossing curves)**

ifBO: In-context Freeze-Thaw Bayesian Optimization

In-context surrogate model **(FT-PFN)**

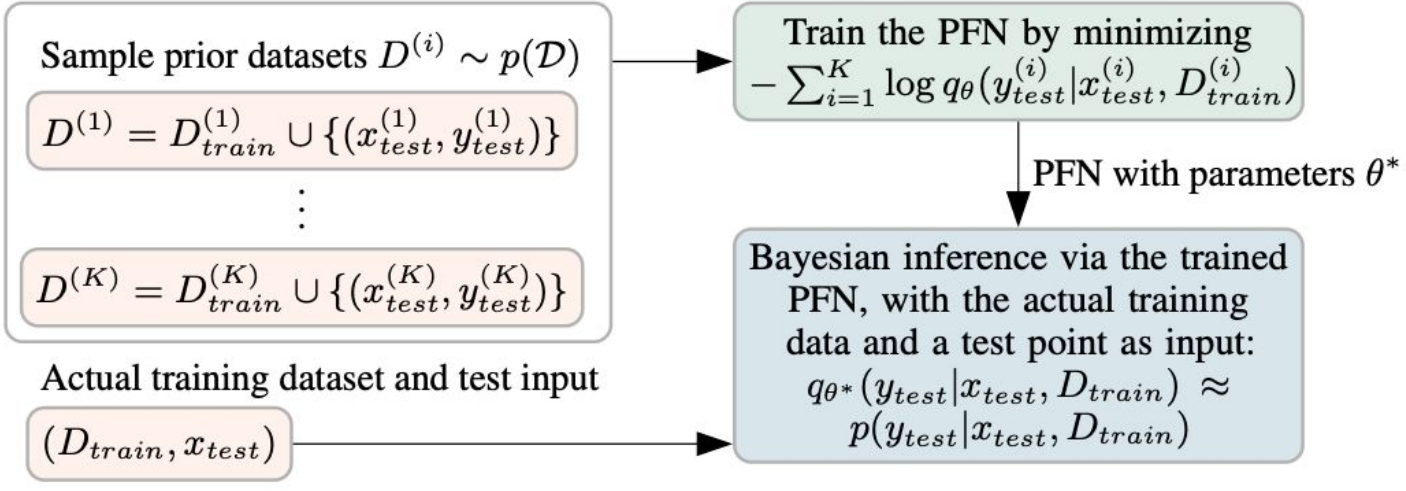
- ➔ **no need for retraining → fast**
- ➔ **no hyper-hyperparameters → easy to use**

Acquisition function **(MFPI-Random)**

- ➔ **adaptive setting → robust**

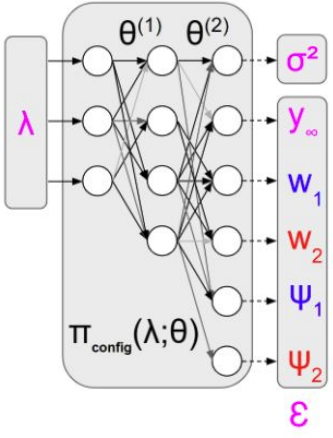
ifBO / FT-FPN - Background on in-context learning with Prior-fitted Networks (PFNs)

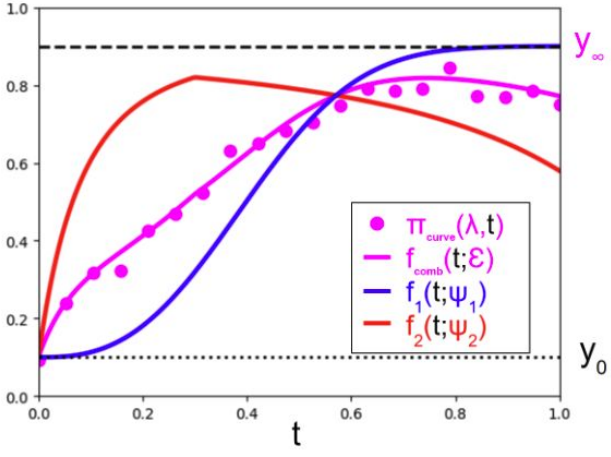
PFNs idea (*Müller et al., 2022*): train a Transformer on datasets generated from a **predefined prior**



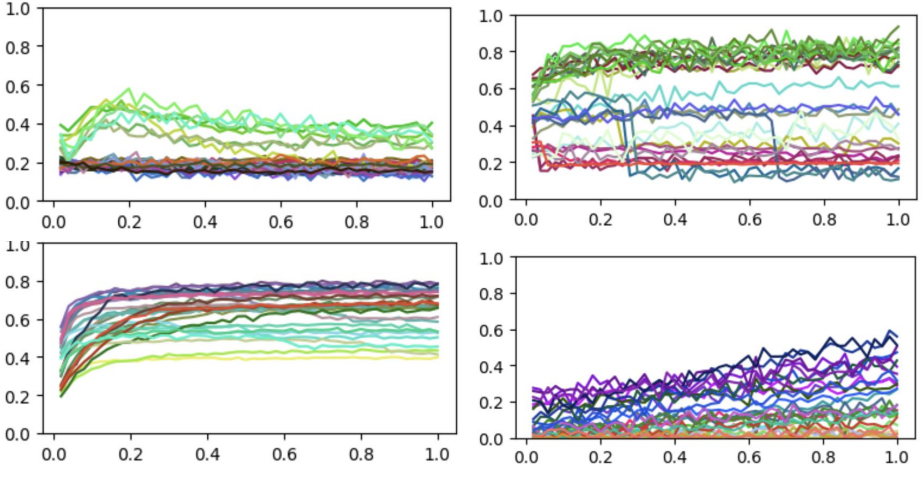
ifBO / FT-FPN - Our joint learning curves and hyperparameters prior

Inspired from LCNet *(Klein et al, 2017)*





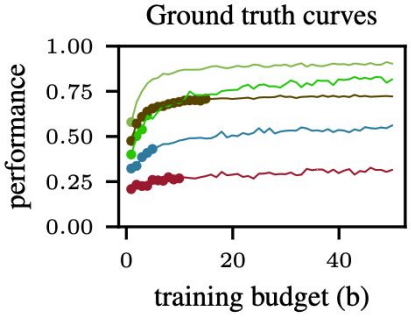
λ	hyperparameter
f_1, f_2	basis curves
Ψ_1, Ψ_2	parameters of the basis curves
W_1, W_2	weights of the basis curves
Π_{curve}	weighted combination of basis curves
σ^2	noise of the curve



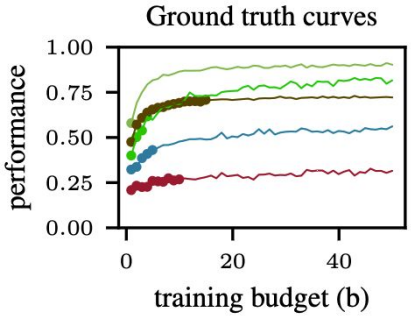
PFN trained on the generated datasets:

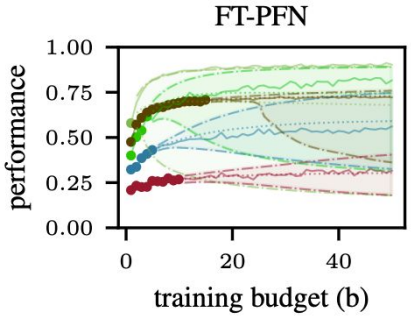
- sample NN (architecture and weights) and sample uniformly a set of configurations (s) from a unit cube.

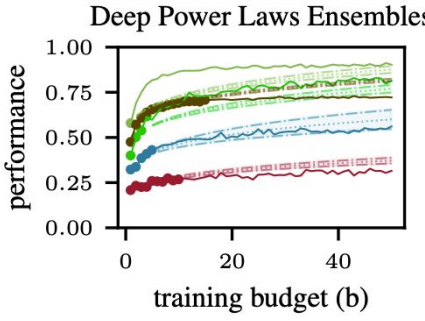
Quality of the surrogate model - Visual comparison of the predictions

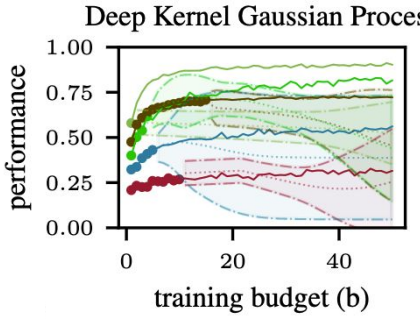


Quality of the surrogate model - Visual comparison of the predictions







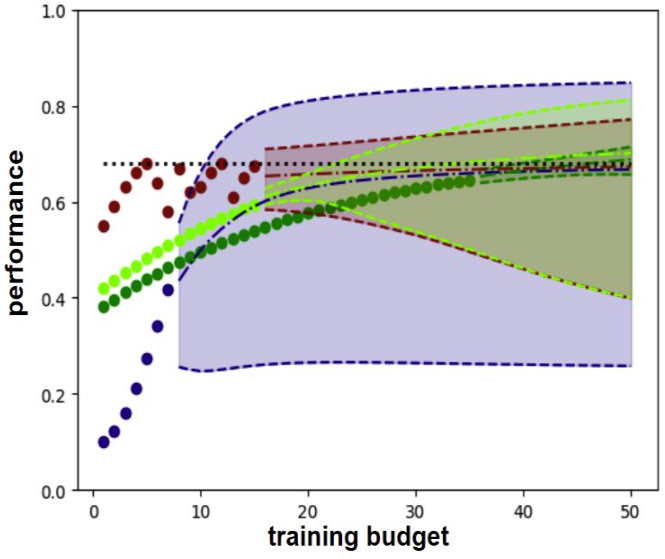


Results generalize to real learning curves benchmarks

ifBO / MFPI-Random - challenges

- How far to look at?
  - DyHPO: at one step ahead
  - DPL: at the maximum step

- How do we measure improvement?
  - DyHPO: compared to the best at the next step if it exists
  - DPL: compared to the best seen so far



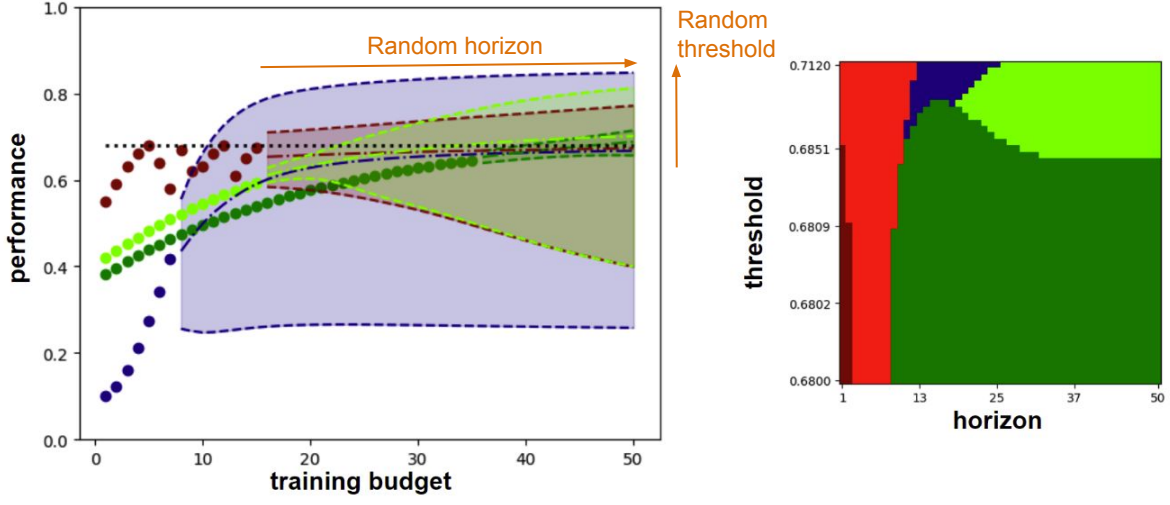
ifBO / MFPI-Random

Probability of Improvement (PI)

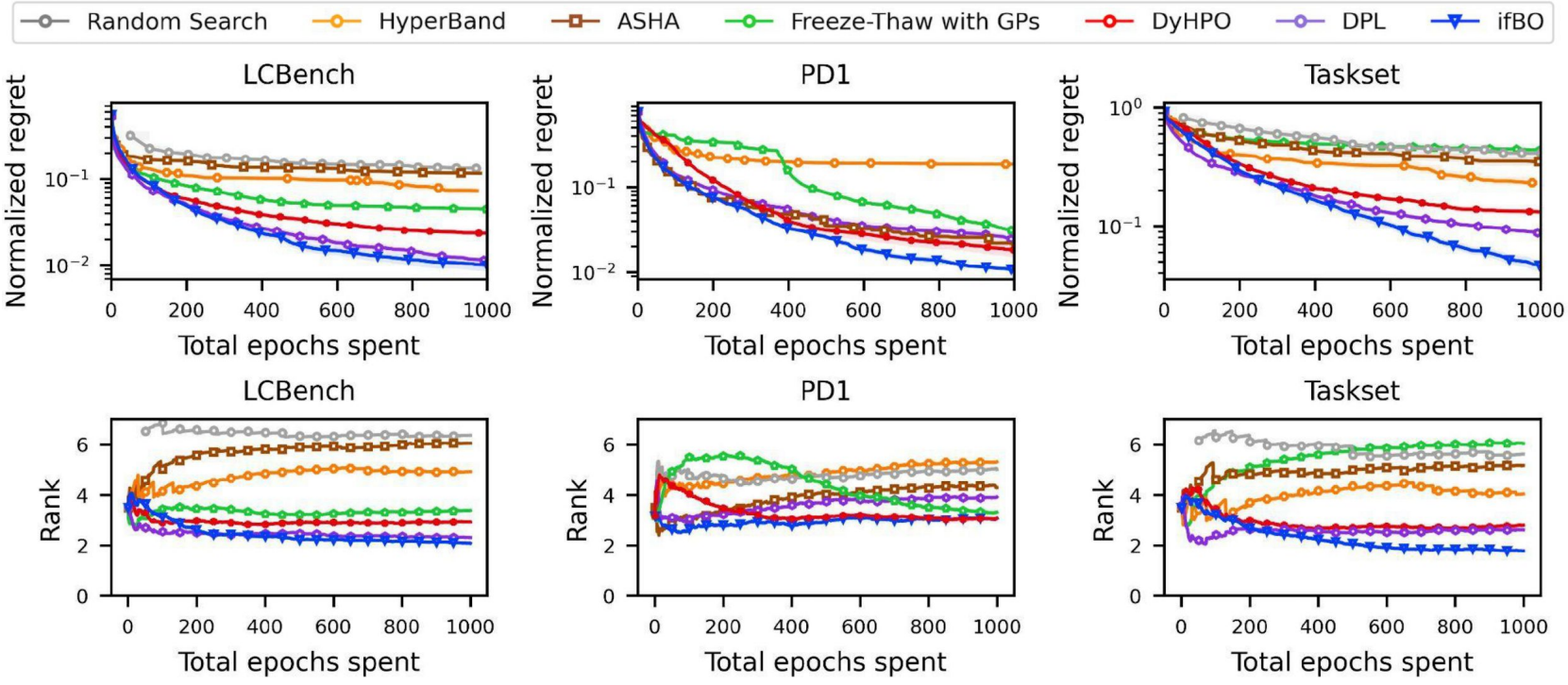
Randomize:

(i) the extrapolation length

(ii) scaling factor for threshold of improvement



Empirical comparison / HPO tasks



Summary

Hall C 4-9 #3003
Tue 23 Jul 11:30 a.m. CEST — 1 p.m. CEST

- **ifBO**: In-context Freeze-thaw Bayesian Optimization
  - FT-PFN: better extrapolation performances and uncertainty estimates
  - Combined with a simple acquisition function → robust HPO performances

- Further research directions
  - Consider different fidelities such as model size; incorporate user priors
  - Improving acquisition function: cost-aware

paper

code