Accelerate and Improve SMBO with Existing Data

Philipp Bordne*[1], Zainab Sultan*[1]

*Equal Contribution [1]University of Freiburg, Germany

UNI FREIBURG

Abstract

- We **present** and **compare** different approaches to enhance SMBO of a black box function through SMAC3 utilizing an **existing warmstart dataset** from which best configurations are hidden.
- Accordingly we found approaches that do **not assume optimality** of the warmstart data to yield the **best** results.
- Rather surprisingly **warmstarting** SMAC's surrogate model did **not** result in **improved** final performance.

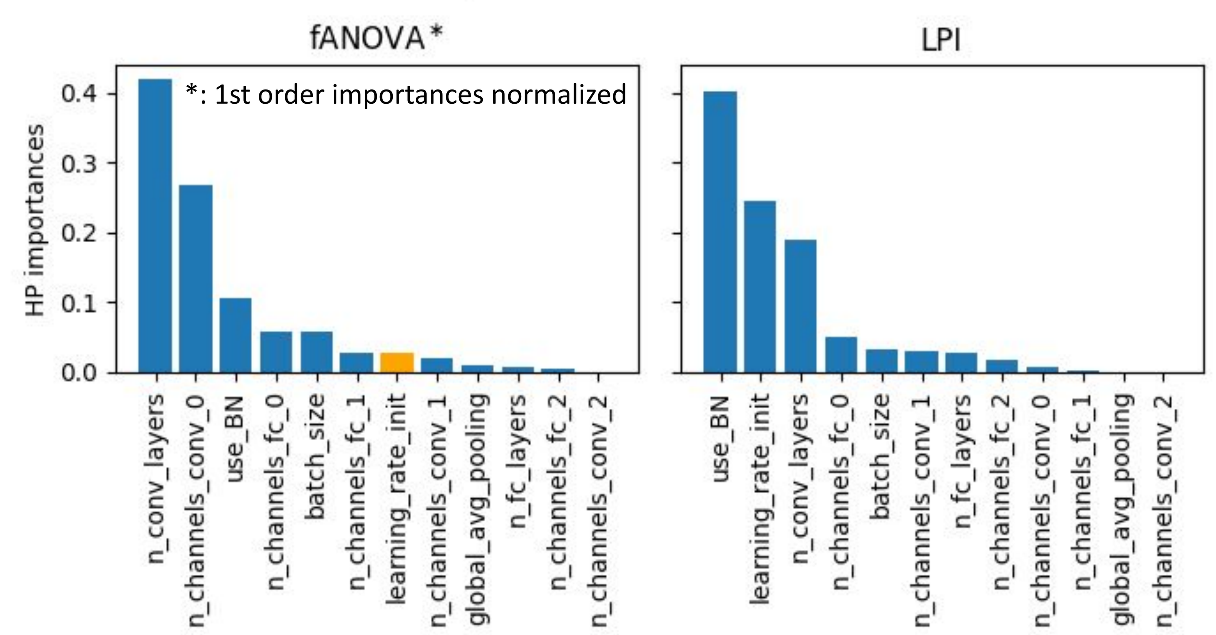
Problem setting

- **Objective** (minimize) $f(\lambda)$: misclassification rate in image classification task
  f: CNN training pipeline and evaluation on test set
- **Inputs** λ: hyperparameters (HPs) from configuration space Λ
- **Warmstart dataset** *D* of size *K*, given**:**
  - Configurations at highest budget from *n* seeded multi-fidelity runs
  - **Best 20%** of configurations are **hidden**:
    $D = \{(\lambda_1, f(\lambda_1)), \dots (\lambda_1, f(\lambda_K))\} \quad \forall\, d \in \{1, \dots, K\}: f(\lambda_d) > f(\lambda^*)$
  - Accuracy of best configurations in *D* < 60%

Approach 1: Reducing search space dimension

- Identify most **important hyperparameters using LPI** from DeepCAVE.
- Apply Pareto principle (explained importance >80%), keep at least 4 tunable HPs.
- Set constant HPs to values of best configurations in *D*.

1st order HP importances from warmstart data



Why LPI over fANOVA: Low importance of `learning_rate_init` under fANOVA seemed unrealistic (~ *expert prior*), see future work for better approaches.
Limitation: fANOVA and LPI do not consider conditionality, but our search space is highly conditional (`n_channels_<layer>` conditions on `n_layers`)

Approach 2: Reducing search space size

Constrain configuration space to **box** around best configurations:

- $\Lambda^* = \{\lambda_1^*, \dots, \lambda_n^*\}$ values with highest accuracy known for *n* seeds in metadata.
- Limit numerical hyperparameters of reduced configspace Λ' to **new bounds** $[l', u']$, s.t.: $\forall i \in \{1, \dots, n\}$: $l' + m \le \lambda_i^* \le u' - m$.
- *m* is margin accounting for fact that true optimizers of problem have been excluded. We set *m=0.2.*

Alternative approach: Adapt **only upper bounds** of new search space.
IDEA: Overfitting likely cause for suboptimal configurations at highest fidelity. *D* was collected on multi-fidelity runs.

Approach 3: Warmstarting

Warmstart surrogate with *D* through SMAC's `tell`-interface (2 approaches):

greedy: Fit surrogate to data from seed with most given values and sample from it right away (i.e. no further configuration evals based on SMACs initial design).
IDEA: Does *D* **alone** suffice to guide SMAC to good regions?

all-in: Assume data to be coming all from the same seed and fit the model to it. Sample initial design configurations at beginning of SMBO as in default SMAC.
IDEA: Can *D* provide an **information advantage** to SMAC over the baseline?

Side contribution: We fixed a bug in racing implementation of SMAC that lead to `rejected_configs` not being updated correctly. Pull request was created. :-)

Approach 4: Gentle Pruning[1]

This approach prunes the search space by evaluating the potential of a hyperparameter configuration λ. The potential of λ is defined as:

$$\text{Potential}(\lambda) = \hat{y}(\lambda) - \max_{\lambda' \in \Lambda_t} \hat{y}(\lambda')$$

Where $\hat{y}$ is a **gaussian process** fit on *D*, and Λ_t is the set of all hyperparameters evaluated in the current SMAC run. This metric is used by the **acquisition function** throughout the optimization to **rank** and keep configurations in the **top N^{th} percentile** where *N* allows control of pruning aggressiveness. We test for *N*=0.2 and *N*=0.8 and report the better performer (*N*=0.8)

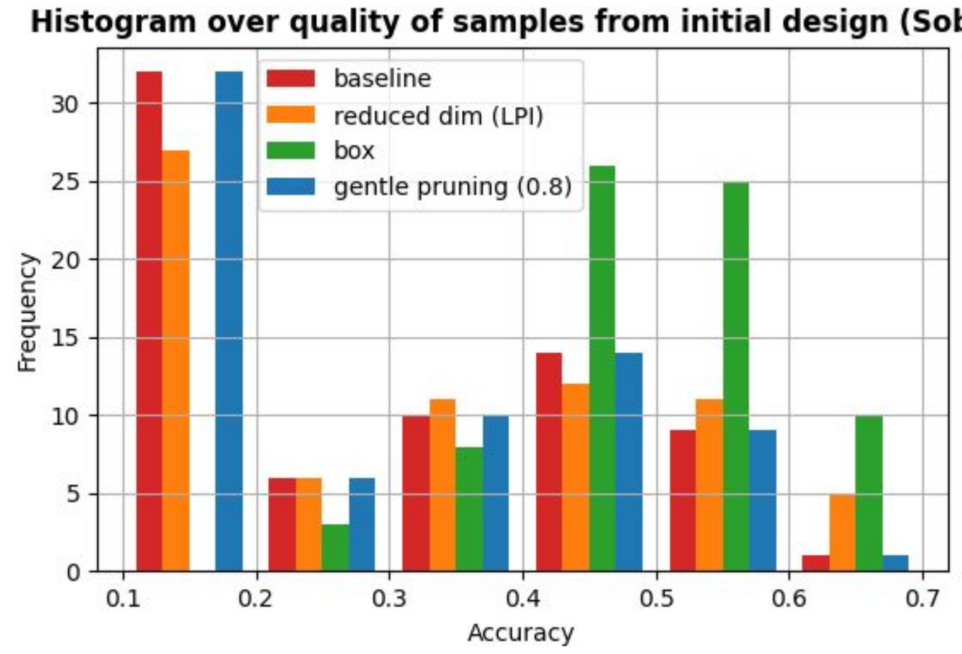
Approach 5: Defining a prior

This approach aims to **infer** a suitable prior on where the **promising regions** would be using *D*. The prior is then augmented into the acquisition function using the existing work on **πBO**[2]

- *Integer* hyperparameters: The prior is a normal distribution with a small sigma, centered around the hyperparameter value with the best cost on average.
- *Categorical* hyperparameters: We place probability weights on the different values that are proportional to their average cost, by fitting a sigmoid function.

Effect of pruning on initial sampling

Histogram over quality of samples from initial design (Sobol)



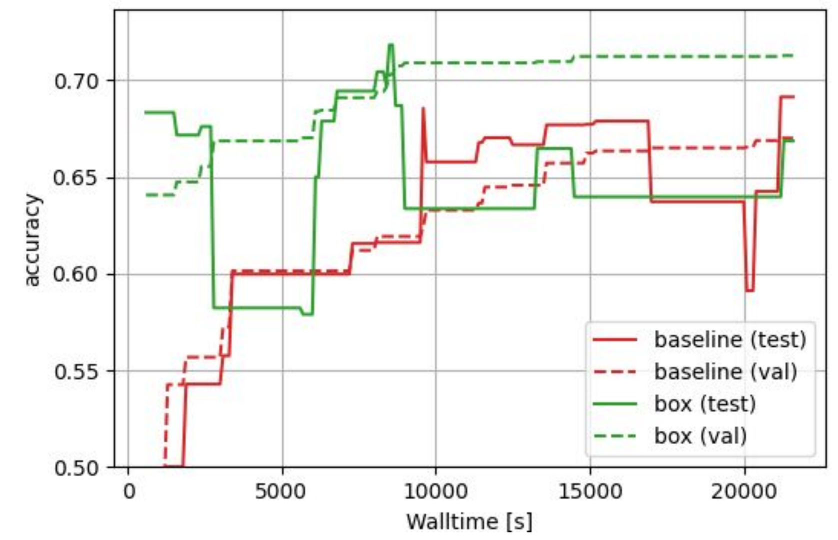
Key insights:

- Box pruning produces a higher proportion of configurations with a high validation accuracy, *however this does not indicate a good test performance*
- All variations produce configurations that are as good as the ones in warmstart dataset

Experimental Design

- 3 seeded runs (except *warmstart greedy*, *gentle pruning*: 2 seeds)
- 1 GPU (NVIDIA Tesla V100 w. 32GB memory), 4 CPUs at BWUniCluster2.0
- Limit at 6h (21600s) runtime OR 150 trials
- runs limited by n_trials used ≥ 20000s walltime (except *box (alternative)* but no detailed evaluation for this experiment)

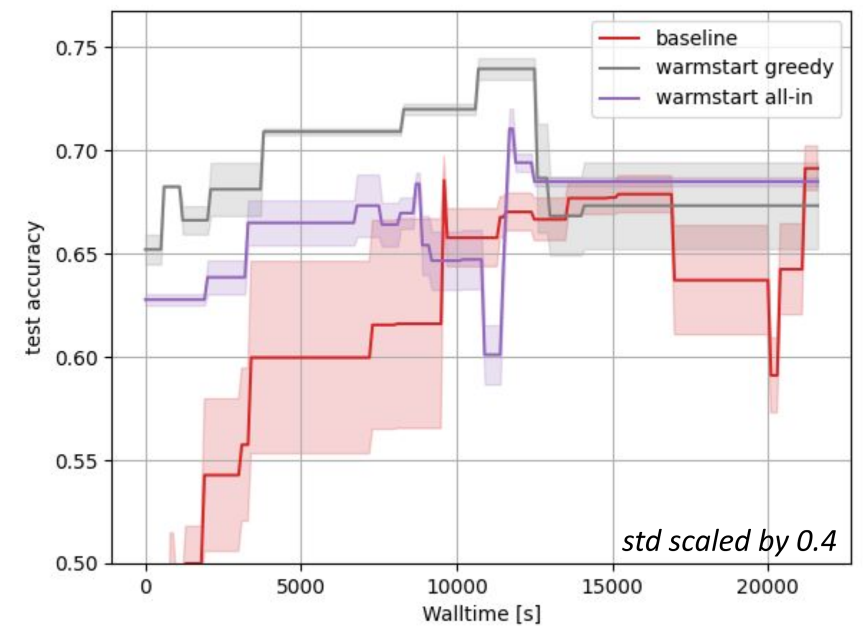
Results: Box Pruning



Poor generalization:
Test performance of box approach is much lower than its validation performance.

Limiting the configuration space around configurations from *D* **limits** optimizer to **suboptimal region**.

Results: Warmstarting

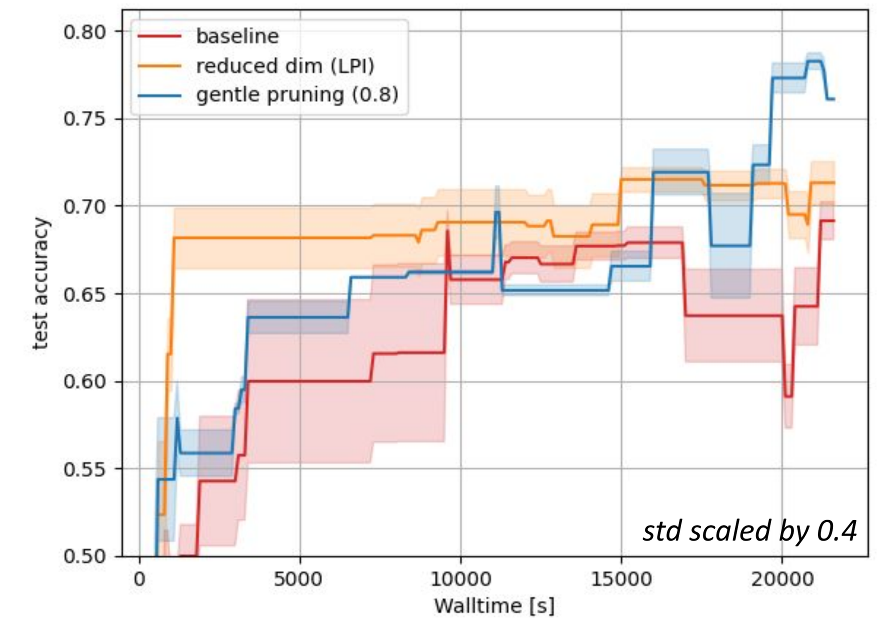


Information about suboptimal configurations does **not provide an advantage** to warmstarted SMBO.

Does warmstarting tie SMBO to similar regions as box?

Does warmstarting **hinder exploration?**

Results: Best Approaches



Gentle pruning yields **best accuracy**.

LPI dim reduction primarily yields **speed-up** plus minor accuracy improvement.

Both approaches do **not assume optimality** of data.

Further experimental results (w/o analysis): Prior: mean= 0.70, std= 0.02
Box (alternative): mean= 0.70 std= 0.03

Summary and Future Work

- **Tightening** search space bounds **risky**, if data is **not optimal.**
- Approaches **extracting information** from data and not assuming its optimality **performed best**.
  → recommended option, if data is known to miss optimal configurations OR there is **uncertainty** whether optimal values **translate** well to new problem instance
- Inferior performance of **warmstarting requires further analysis**:
  - Compare configuration footprint of baseline and warmstarting to confirm lack of exploration
  - Compare incumbents of warmstarting and box, to confirm tie to suboptimal region
- Racing on more seeds could improve generalization from validation to test performance
- **fANOVA** theoretically **more appropriate** to identify HPs to tune:
  - fANOVA handling **conditional HPs naturally** expected to yield better basis for HP selection
  - **Combine** fANOVA importances **with expert priors** (e.g. keep learning rate) for HP selection

[1]Wistuba, Martin, Nicolas Schilling, and Lars Schmidt-Thieme. "Hyperparameter search space pruning–a new component for sequential model-based hyperparameter optimization." *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part II 15*. Springer International Publishing, 2015.
[2]Hvarfner, Carl, et al. "$\pi $ BO: Augmenting acquisition functions with user beliefs for bayesian optimization." arXiv preprint arXiv:2204.11051 (2022).

AutoML.org
@AutoML_org