MODEL-BASED ASYNCHRONOUS HYPERPARAMETER AND NEURAL ARCHITECTURE SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a model-based asynchronous multi-fidelity method for hyperparameter and neural architecture search that combines the strengths of asynchronous Successive Halving and Gaussian process-based Bayesian optimization. Our method is based on a probabilistic model that can simultaneously reason across hyperparameters and resource levels, and supports decision-making in the presence of pending evaluations. We demonstrate the effectiveness of our method on a wide range of challenging benchmarks, for tabular data, image classification and language modelling, and report substantial speed-ups over current state-of-the-art methods. Our new methods, along with asynchronous baselines, are implemented in a distributed framework which will be open sourced along with this publication.

1 INTRODUCTION

The goal of hyperparameter and neural architecture search (HNAS) is to automate the process of finding the right architecture or hyperparameters $\boldsymbol{x}_\star \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ defining a deep neural network by minimizing the validation loss $f(\boldsymbol{x})$, observed through noise: $y_i = f(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), i = 1, \ldots, n$. Bayesian optimization (BO) is an effective model-based approach for these expensive optimization problems (Jones et al., 1998; Shahriari et al., 2016). It constructs a probabilistic surrogate model of the loss function $p(f \mid \mathcal{D})$ based on previous evaluations $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Searching for the global minimum of f is then driven by trading off exploration in regions of uncertainty and exploitation in regions where the global optimum is likely to reside.

For HNAS problems, standard BO needs to be augmented in order to remain competitive. Training runs for unpromising configurations $\boldsymbol{x}$ can be stopped early, and thereby serve as low-fidelity approximations of $f(\boldsymbol{x})$ (Swersky et al., 2014; Domhan et al., 2015; Klein et al., 2017b). Further, evaluations of f can be executed in parallel to reduce the wall-clock time required for finding a good solution. Several methods for multi-fidelity and/or distributed BO have been proposed (Kandasamy et al., 2017; 2018a; Takeno et al., 2020), but they rely on rather complicated mechanisms to select either the fidelity level or the next candidate. In this work we aim to adopt the desiderata of Falkner et al. (2018), namely, that of simplicity, which often leads to more robust methods in practice.

A simple, easily parallelizable multi-fidelity scheduling algorithm is successive halving (SH) (Karnin et al., 2013; Jamieson & Talwalkar, 2016), which iteratively eliminates poorly performing neural networks over time. Hyperband (HB) (Li et al., 2017) iterates over multiple rounds of SH with varying ratios between the number of configurations and the minimum amount of resources spent per configuration. Falkner et al. (2018) introduced a hybrid model, called BOHB, that uses a probabilistic model to guide the search while retaining the efficient any-time performance of HB. However, both HB and BOHB can be bottlenecked due to their *synchronous* nature: stopping decisions are done only after synchronizing all training jobs at certain resource levels (called *rungs*). This approach is wasteful when the evaluation of some configurations take longer than others, as it is often the case when training neural networks (Ying et al., 2019), and can substantially delay progress towards high-performing configurations (see Figure 1 for an illustration).

Recently, Li et al. (2018) proposed ASHA, which adapts successive halving to the asynchronous parallel case. Even though ASHA only relies on the random sampling of new configurations, it has been shown to outperform synchronous HB and BOHB. In this work, we augment ASHA with a Gaussian process (GP) surrogate, which jointly models the performance across configurations and rungs to improve the already strong performance of ASHA.

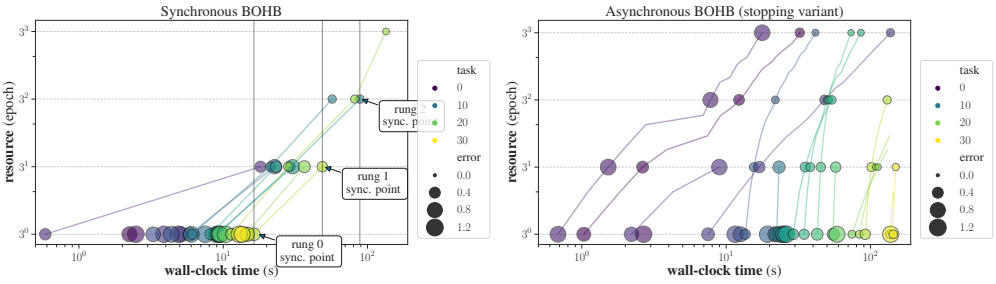

Figure 1: Evolution of synchronous (left) and asynchronous (right) SH. x: wall-clock time (150 secs), y: training epochs done (log scale). Coloured lines trace training jobs for HP configurations, with marker size proportional to validation error (smaller is better). Lines not reaching the maximum number of authorized epochs (27) are stopped early. For synchronous SH, rungs (epochs 1, 3, 9) need to be filled completely before any configuration can be promoted, and each synchronization results in some idle time. In asychronous SH, configurations proceed to higher epochs faster. Without synchronization overhead, more configurations are promoted to the highest rung in the same wall-clock time (figure best seen in colours).

1.1 Contributions

Dealing with hyperparameter optimization (HPO) in general, and HNAS for neural networks in particular, we would like to make the most efficient use of a given parallel computation budget (e.g., number of compute instances) in order to find high-accuracy solutions in the shorted wall-clock time possible. Besides exploiting low-fidelity approximations, we demonstrate that *asynchronous parallel* scheduling is decisive factor in cost-efficient search. Our novel combination of asynchronous successive halving with multi-fidelity BO, dubbed MOdel Based aSynchronous mulTi fidelity optimizER (MOBSTER) substantially outperforms either of them in isolation. More specifically:

- We clarify differences between existing asynchronous SH extensions: ASHA, as described by Li et al. (2018) and an arguably simpler stopping rule variant, related to the median rule (Golovin et al., 2017) and first implemented in Ray Tune (Liaw et al., 2018). Although their difference is subtle, they lead to substantially behaviours in practice.
- We present an extensive ablation study, comparing and analysing different components for asynchronous HNAS. While these individual components are not novel, one of our main contribution is to show that, by systematically combining them to form MOBSTER, we obtain a reliable and more efficient method than other recently proposed approaches.
- On a variety of neural network benchmarks, we show that MOBSTER is more efficient in terms of wall-clock time than other state-of-the-art algorithms. Unlike BOHB it does not suffer from substantial synchronization overheads when evaluations are expensive. As a result, we can achieve the same performance in the same amount of wall-clock time but with half the computational resources compared to random sampling.

Next, we relate our work to approaches recently published in the literature. In Section 2, we review synchronous SH and HB, as well as two asynchronous extensions. Our novel method is presented in Section 3. We present empirical evaluations for the HNAS of various neural architecture types in Section 4, and finish with conclusions and future work in Section 5.

1.2 Related Work

A range of prior Gaussian process (GP) based BO work exploits multiple fidelities of the objective (Kennedy & O'Hagan, 2000; Klein et al., 2017a; Poloczek et al., 2018; Cutajar et al., 2018). A joint GP model across configurations and tasks was used by Swersky et al. (2013), allowing for trade-offs between cheap auxiliaries and the expensive target. Klein et al. (2017a) presented a continuous multi-fidelity BO method, where training set size is an input. It relies on a complicated and expensive acquisition function, and experiments with asynchronous scheduling are not provided.

Kandasamy et al. (2017) presented BOCA, a Bayesian optimization method that exploits general continuous fidelities of the objective function. While BOCA uses a model similar to ours, they employ a different strategy to select the fidelity, which appears to work less efficiently in practice, as we show in our experiments. The "freeze-thaw" approach (Swersky et al., 2014) allows for asynchronous parallel evaluations, yet its scheduling is quite different from asynchronous HB. They propose an exponential decay kernel, which we use here as well. No results for deep neural network tuning were presented, and no implementation is publicly available. Previous work on asynchronous BO has demonstrated performance gains over synchronous BO methods (Alvin et al., 2019; Kandasamy et al., 2018a), yet this work did not exploit multiple fidelities of the objective. Concurrently to our work, Takeno et al. (2020) proposed an asynchronous multi-fidelity BO method. However, it is based on an information-theoretic acquisition function that is substantially more complex than our method. Since no implementation is publicly available they are omitted from our comparisons.

Falkner et al. (2018) combine synchronous HB with TPE (Bergstra et al., 2011a) as a model. Their method, called BOHB, combines early speed-ups of HB with fast convergence later on, and shows competitive performance for NAS (Ying et al., 2019; Dong & Yang, 2020). However, compared to our approach, it fits independent kernel density estimator at each rung level, which prevents the interpolation to higher rung levels. Apart from this, kernel density estimators do not permit fantasizing the potential outcome of a pending configuration, which makes it unclear how to extend BOHB to the asynchronous setting, and are notoriously sensitive to their bandwidth parameters.

2 Synchronous and Asynchronous Multi-Fidelity Scheduling

Multi-fidelity HNAS considers an objective $f(\boldsymbol{x}, r)$, indexed by the *resource level* $r \in \{r_{\text{min}}, \ldots, r_{\text{max}}\}$. The target function of interest is $f(\boldsymbol{x}) = f(\boldsymbol{x}, r_{\text{max}})$, while evaluations of $f(\boldsymbol{x}, r)$ for $r < r_{\text{max}}$ are cheaper, lower-fidelity evaluations that may be correlated with $f(\boldsymbol{x})$. Here, we consider the number of training epochs (*i.e.*, full sweeps over the data) as the resource r, though other choices are possible, such as training subset ratio (Klein et al., 2017a).

Two kinds of decisions are made in multi-fidelity HNAS: *choosing* the configurations to evaluate and *scheduling* their evaluations. Scheduling decisions are stop/go decisions that happen at certain resource levels called *rungs*. In this work, we chose them according to SH for its simplicity and strong empirical performance (Jamieson & Talwalkar, 2016; Li et al., 2018). Let $\eta \in \{2, 3, 4\}$ be the halving constant, r_{max} and r_{min} the maximum and minimum resource level for an evaluation. We assume, for simplicity, that $r_{\text{max}}/r_{\text{min}} = \eta^K$. The full set of rungs is $\mathcal{R} = \{r_{\text{min}}\eta^k \mid k = 0, \ldots, K\}$. In the sequel, the term "rung" will be overloaded to both mean the resource level at which stop/go decisions are made, and the list of configurations that reached this level. SH is obtained as a special case of HB, using only a single *bracket* B_0. Brackets are (ordered) subsets $B_s = \{r_{\text{min}}\eta^{k+s} \mid k = 0, \ldots, K-s\} \subseteq \mathcal{R}$, where $s = 0, \ldots, K$. In an outer loop, HB cycles over these brackets until the total budget is spent.

We distinguish between *synchronous* and *asynchronous* scheduling. Each rung has an *a priori* fixed size (*i.e.*, number of slots for configurations evaluated until the rung level), the size ratio between successive rungs being η^{-1}. Configurations promoted to a rung are evaluated until the corresponding resource level, making use of parallel computation if available. Once all of them finish, the η^{-1} fraction of top performing configurations are promoted to the next rung while the others are terminated. Each rung is a synchronization point for many training tasks: it has to be populated entirely before any configuration can be promoted (see Figure 1 left for a visualization).

Synchronous scheduling makes less efficient use of parallel computation then asynchronous scheduling. Configurations in the same rung may require quite different compute times, for example, if we search over hyperparameters that control the network size. Hence, some workers may just be idle at a synchronization point. Evaluations at larger resource levels are observed earlier with asynchronous scheduling. However, the risk is to continue mediocre configurations because they were selected earlier. Figure 1 illustrates the differences between asynchronous and synchronous scheduling.

We also distinguish two variants of asynchronous SH proposed in prior work. During the optimization process, for any hyperparameter and architecture configuration $\boldsymbol{x} \in \mathcal{X}$ that is currently being evaluated, a decision needs to be made once it reaches the next rung $r \in B_s$. Given the new observed

data point y at level r, the binary predicate $\texttt{continue}_s(\boldsymbol{x}, r, y)$ evaluates to true iff y is in the top η^{-1} fraction of records at the rung. The two variants differ in the case if `continue` is true:

Stopping variant. This is a simple extension of the median stopping rule (Golovin et al., 2017), first implemented in Ray Tune (Liaw et al., 2018). As soon as a job for $\boldsymbol{x}$ reaches a rung r, if $\texttt{continue}_s(\boldsymbol{x}, r, y)$ is true, it continues towards the next rung level. Otherwise, it is *stopped*, and the worker becomes free to pick up a novel evaluation. As long as fewer than η configurations have reached a rung, the job is always continued.

Promotion variant. This variant of asynchronous SH was presented by Li et al. (2018) and called *ASHA*. Once a job for some $\boldsymbol{x}$ reaches a rung r, it is *paused* there, and the worker is released. The evaluation of $\boldsymbol{x}$ can be *promoted* (*i.e.*, continued to the next rung) later on. When a worker becomes available, we scan the rungs in descending order, running a job to promote the first paused $\boldsymbol{x}$ for which $\texttt{continue}_s(\boldsymbol{x}, r, y)$ is true. When no such paused configuration exists, we start a new evaluation from scratch. With fewer than η metrics recorded at a rung, no promotions happen there. The promotion variant can be made efficient with pause-and-resume training of the neural network.[1]

The stopping and promotion variants can exhibit rather different behaviours (see Figure 2) which we will also show in the experiments. Initially, the stopping variant gives most configurations the benefit of doubt, while the promotion variant pauses evaluations until they can be compared against a sufficient number of competitors.

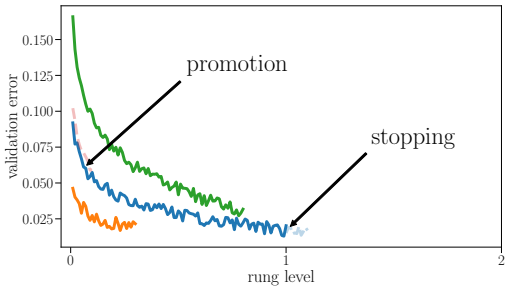

Figure 2: Difference between the promotion and stopping variants of SH. The three curves correspond to the evaluations of three configs. The blue curve reaches the smallest rung level first, which triggers the scheduling process. While the stopping variant would continue the evaluation (shaded blue curve), the promotion variant would stop, and proceed with a new configuration (shaded red curve). The evaluation of blue would only be continued once enough datapoints at this rung level have been observed. As shown in the experiments, the stopping variant achieves often better performance in early iterations, whereas the promotion variant obtains usually a better final performance.

3 Model-based Asynchronous Multi-Fidelity Optimizer

We now describe our new model-based method that is able to make asynchronous decisions across different fidelities of the objective function. Motivated by the fact that in standard sequential settings, Gaussian process (GP) based BO tends to outperform random search (Shahriari et al., 2016), we aim to equip asynchronous SH/HB with decisions based on a *GP surrogate model*. Following a range of prior work on non-parallel multi-fidelity BO (Swersky et al., 2013; Klein et al., 2017a; Poloczek et al., 2018; Swersky et al., 2014), we place a GP prior over $f(\boldsymbol{x}, r)$ to jointly model not only the autocorrelation within each fidelity, but also the cross-correlations between fidelities,

$$f(\boldsymbol{x}, r) \sim \mathcal{GP}(\mu(\boldsymbol{x}, r), k((\boldsymbol{x}, r), (\boldsymbol{x}', r'))).$$

This is in contrast to synchronous BOHB (Falkner et al., 2018), which just employs an independent TPE (Bergstra et al., 2011b) model over $\boldsymbol{x}$ for each rung level $r \in B_s$, which prevents it to transfer information from lower to higher rungs.

Since multiple configurations are evaluated in parallel, we have to deal with *pending* evaluations when proposing new configurations. We do so by *fantasizing* their outcomes, as detailed by Snoek et al. (2012). In a nutshell, we marginalize the acquisition function over the GP predictive posterior for pending configurations. In practice, we approximate this by averaging the acquisition function over sampled function values. It is important to emphasize that, as long as kernel hyperparameters remain constant, fantasizing comes almost for free. For GP regression, the posterior covariance matrix does not depend on the function values (only on the inputs), which means that fantasizing does not require additional cubic-complexity computations.

[1] Supporting pause-and-resume by checkpointing is difficult in the multi-node distributed context. It is simpler to start training from scratch whenever a configuration is promoted. This is done in our implementation.

How do pending evaluations and fantasizing generalize to the asynchronous successive halving case? When we start to evaluate $\boldsymbol{x}$ in bracket s, we register a pending evaluation for $(\boldsymbol{x}, r_{\min}\eta^s)$. Once it reaches the next rung $r \in B_s$, we convert the *pending* to a *labeled* evaluation $(\boldsymbol{x}, r)$. Then, as detailed above, we make a stopping decision. If the configuration survives (i.e `continue` evaluates to true), we register $(\boldsymbol{x}, r\eta)$ as *pending*, since $r\eta$ is the next rung it will reach. Details about how posterior distributions are represented, are given in Appendix A.1.

Kernel choice. We implemented and compared a range of different surrogate models for $f(\boldsymbol{x}, r)$: a joint Matérn$5/2$ kernel with automatic relevance determination (ARD) (Snoek et al., 2012) over $(\boldsymbol{x}, r)$, with and without warping of r; the additive exponential decay model of Swersky et al. (2014); a novel non-additive extension of the latter, relaxing underlying assumptions which may not hold in practice. In all cases, the base kernel on $\boldsymbol{x}$ alone is Matérn$5/2$ ARD. Details for our new surrogate model are provided in Appendix A.2.

Acquisition strategy. We follow synchronous BOHB in choosing $\boldsymbol{x}$ by optimizing the expected improvement (EI) acquisition function (Shahriari et al., 2016) over the posterior on $f(\cdot, r_{\text{acq}})$, where r_{acq} is fixed for each decision. We choose $r_{\text{acq}} \in \mathcal{R}$ as the largest rung for which at least L metric values have been observed, where L is typically set to the input dimensionality (*i.e.* number of hyperparameters). After some initial period, r_{acq} eventually converges to $r_{\max}$. This strategy does not take into account that the choice of $\boldsymbol{x}$ may have different downstream stopping or promotion consequences. We leave the exploration of alternative strategies as a direction for future work.

4 EXPERIMENTS

We compare MOBSTER to a range of state-of-the-art algorithms. For synchronous HB and BOHB, we used the `HpBandSter` package.[2], which uses a simple mechanism to reduce the amount of idling workers: if there are more free workers than slots in the rung, left-over workers are used to immediately start the next bracket. There is no open-source implementation of Swersky et al. (2014) and Takeno et al. (2020) (for which the complexity considerable), which is why omitted them here.

We implemented MOBSTER, as well as all other asynchronous techniques, in `AutoGluon`.[3], which will all be open-sourced along with this paper. A number of worker nodes run training jobs, publishing metrics after each epoch. One of the workers doubles as master, concurrently running the HP optimization and scheduling.

We also compare against Dragonfly (Kandasamy et al., 2019). While `AutoGluon` allows to parallelize both across instances and processes on a single instance, Dragonfly only supports the latter. Hence, for a fair comparison, we run our method only on a single instance if we compare against Dragonfly. For all other experiments we parallelize across multiple instances.

For all methods using SH the halving rate was $\eta = 3$. We report the *immediate regret* $r_t = y_t - y_\star$ after wall-clock time t, where $y_\star$ is the best observed validation error across all methods, runs and time steps on this dataset. Using the regret instead of just the validation error gives a more detailed picture of progress towards the global optimum. For all classification datasets we cap the regret at a lower bound of 10^{-3}, which corresponds to a 0.1% difference in classification accuracy to the global optimum. We consider lower values to be not significant in practice and attributed them to the intrinsic noise of the objective function. For each method we report mean and the standard error of the mean across multiple runs with different seeds of the random number generator.

We run experiments on the following set of benchmarks (for further details regarding the search spaces see Appendix B.All experiments were conducted on AWS EC2, using `m4.xlarge` instances for CPU- and `g4dn.xlarge` instances for GPU-based computations, respectively.

Tabular Datasets: First, we optimized the hyperparameters of a multi-layer perceptron (MLP) on the tabular classification dataset *electricity* gathered from OpenML (Vanschoren et al., 2014). The MLP has two hidden layers with ReLU activations. We tuned 9 hyperparameters: the learning rate of ADAM (Kingma & Ba, 2015), batch size, weight decay, and, for each layer, dropout rate, number of units and the scale of a uniform distribution to initialize weights. The number of epochs r varies between $r_{\min} = 1$ and $r_{\max} = 81$.

[2] `https://github.com/automl/HpBandSter`

[3] `https://github.com/awslabs/autogluon`

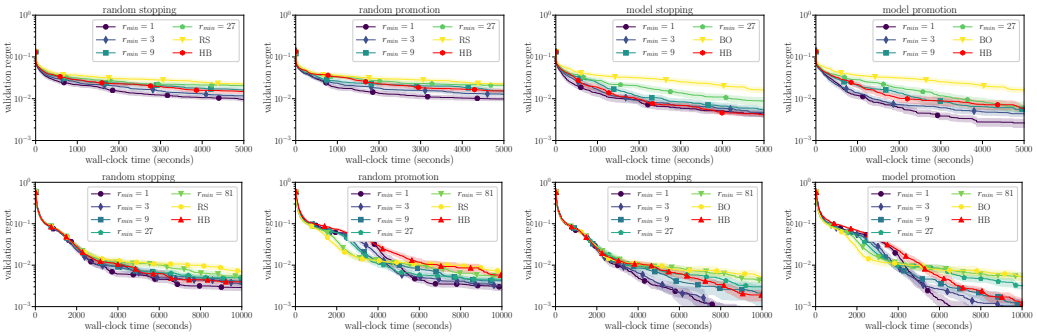

Figure 3: Comparison of asynchronous HB and SH (stopping / promotion scheduling variants and sampling from model / random) on the MLP benchmark (top) and NASBench201 (bottom). SH with the most aggressive bracket works as well and often better than HB across all variants.

Neural Architecture Search: To evaluate our approach in a NAS setting, we used the NASBench201 datasets from Dong & Yang (2020) which consists of a offline evaluated grid of neural network architectures for three different datasets: CIFAR-10, CIFAR-100 (Krizhevsky, 2009) and Imagenet16 (Chrabaszcz et al., 2017). Categorical hyperparameters were one-hot encoded. The minimum and maximum number of epochs were $r_{\text{min}} = 1$ and $r_{\text{max}} = 200$. To simulate realistic dynamics of asynchronous NAS, we put each worker to sleep for the time per epoch each real evaluation took. This does not speed up experiments, but decreases cost, since no GPU instances are required. Compared to the other benchmarks which contains integer and continuous hyperparameters, this benchmark is purely discrete.

Image Classification: Next, we tuned ResNet (He et al., 2016) on the image classification dataset CIFAR-10 (Krizhevsky, 2009), the batch size, weight decay, initial learning rate and momentum. We used 40000 images for training, 10000 for validation, and applied standard data augmentation (cropping; horizontal flips). SGD is used for training, the learning rate is multiplied by 0.1 after 25 epochs. We used $r_{\text{min}} = 1$ and $r_{\text{max}} = 27$. As our goal is to evaluate our approach, we omitted additional tricks-of-the-trade required to obtain state-of-the-art accuracy (Yang et al., 2020).

Language Modelling: Lastly, we considered an LSTM (Hochreiter & Schmidhuber, 1997), applied to language modelling on the WikiText-2 dataset (Merity et al., 2017) with the default training/validation/test split. We optimized the initial learning rate of SGD, batch size, dropout rate, gradient clipping range, and decay factor multiplied to the learning rate if the validation error has not improved upon the previous epoch. The LSTM consisted of two layers with 1500 hidden units each. For the word embeddings, we also used 1500 hidden units and tied them with the output. Epochs vary between $r_{\text{min}} = 1$ and $r_{\text{max}} = 81$.

4.1 Ablation Study

We start with an in-depth analysis of our proposed method. If not explicitly stated otherwise, we report the immediate regret over wall-clock time of the MLP benchmark on the *electricity* dataset for 4 workers and the NASBench201 CIFAR-10 dataset for 8 workers. These two benchmarks cover continuous as well as discrete hyperparameter optimization problems.

Successive Halving vs Hyperband: First, we compare asynchronous scheduling based on SH versus HB. We run different instances of SH, where we vary the minimum rung level r_{min} in the same way as HB. Figure 3 shows the results for both promotion and stopping based scheduling, combined with sampling from our model or random. Similar to results reported by Li et al. (2018), the performance of SH is decaying with an increasing r_{min} both for our model-based and random sampling. Also, the most aggressive bracket, where $r_{min} = 1$ epoch, works as good or even better than HB, most distinctly for the promotion based scheduling. For the remaining experiments we therefore use SH with a minimum rung level of $r_{min} = 1$ epoch.

Kernels: We analyse the influence of the choice of kernel and mean function for $f(\boldsymbol{x}, r)$ on the performance of the stopping and promotion variant of asynchronous BOHB. We compare different

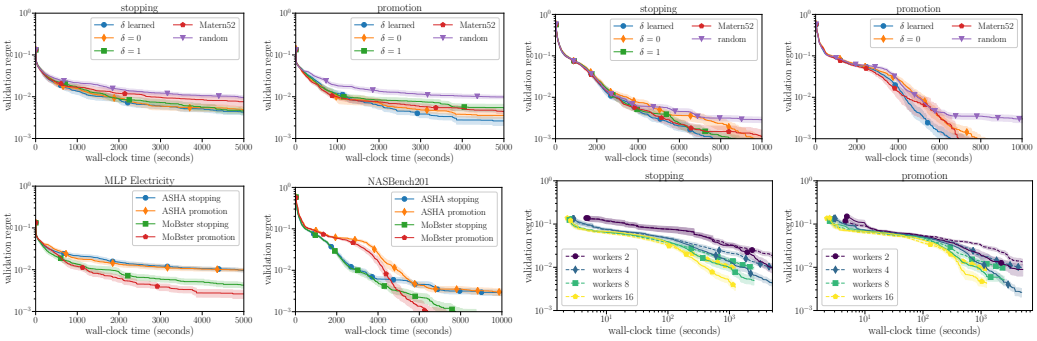

Figure 4: **Top:** Effect of the kernel choice on the MLP benchmark (left, left middle) and NASBench201 (right middle, right). We also include ASHA (random choices) for comparison. There is not a consistent trend across the difference kernels (note the standard error of the mean). **Bottom:** Comparison of ASHA and MoBster (stopping and promotion variant). Left: MLP on *electricity* (4 workers). Left Middle: NASBench201 on CIFAR-10 (8 workers). Right Middle: scaling with the number of workers for stopping and Right: promotion

variants of the exponential decay model (see Appendix A.2. for more details): fixing $\delta = 0$ as in Swersky et al. (2014), $\delta = 1$, or learning δ with the other GP hyperparameters. We use a Matérn$^5/_2$ ARD kernel for the $k_{\mathcal{X}}$ model between configurations $\boldsymbol{x}$ Snoek et al. (2012). As an additional baseline, we use a Matérn$^5/_2$ ARD kernel and constant mean function on inputs $(\boldsymbol{x}, r)$.

In Figure 4 we compare the various types of kernel. Maybe somewhat surprisingly, there is no significant difference between the surrogate models and the precise model for inter- or extrapolation along r seems less important. For the remaining experiments, we use the exponential decay kernel with learned δ as the most flexible choice which also obtains slightly better performance.

MoBster **vs ASHA:**. Next, we show the benefit of sampling from our model instead of uniformly at random for both scheduling variants (see Figure 4). Initially, MoBster (stopping / promotion) performs similar to its random-based counterpart ASHA (stopping / promotion). However, with enough datapoints, it rapidly approaches the global optimum, which is in line with typical improvements of BO over random search (Falkner et al., 2018).

Scaling with the number of workers As seen in Figure 4 right, the substantial speed-ups of MoBster over ASHA are observed consistently with 2, 4, 8, and 16 workers on the MLP benchmark. In fact, in most cases MoBster converges to a high-accuracy solution in about the same time as ASHA needs with twice the number of workers: *using a joint surrogate model can save half the resources compared to random search*.

Stopping versus promotion scheduling:. Next, we focus on the two variants of asynchronous scheduling (stopping and promotion) described in Section 2. The stopping rule achieves a good performance faster than the promotion variant (Figure 3, right) up to 4000 seconds on the NASBench201 benchmark and up to 500 seconds on the MLP benchmark. Recall from Section 2 that for the stopping rule, the initial η configurations are not stopped, while the promotion variant pauses all initial configurations at the lowest rung level of their bracket. However, at least for the MLP benchmark, after this initial phase the promotion-based scheduling achieves slightly better performance.

4.2 Comparison to State-of-the-Art Methods

We now compare MoBster against a range of state-of-the-art algorithms: ASHA (Li et al., 2018), asynchronous BO (Snoek et al., 2012) (A-BO), asynchronous random search (Bergstra et al., 2011a) (A-RS), synchronous BOHB (Falkner et al., 2018) (S-BOHB) , synchronous HB (Li et al., 2017) (S-HB), asynchronous BO with Thompson Sampling (Kandasamy et al., 2018b) (Dragonfly) and BOCA (Kandasamy et al., 2017) (Dragonfly-MF).

MLP and NASBench201 Figure 5 provides results for the MLP and the NASBench201 benchmark, both for the single-instance (left) and multi-instance setting (right) (see Appendix for results on the

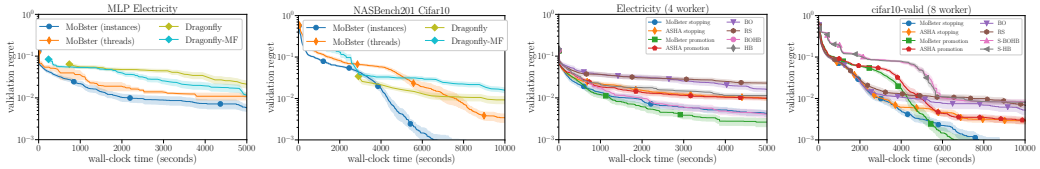

Figure 5: Comparison to various baselines across multiple processes on a single instance (left) and across instances (right) for the MLP and NASBench201 benchmarks. See main text for an analysis.

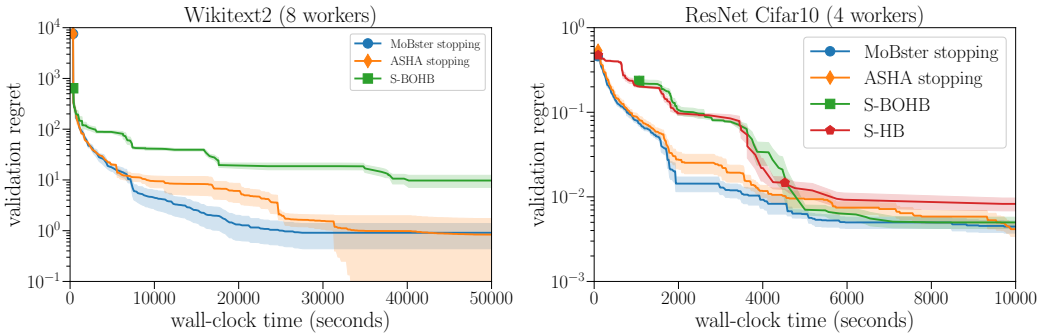

Figure 6: Left: Results of the LSTM benchmark. Right: Results for the ResNet benchmark.

other two NASBench201 benchmarks). MOBSTER outperforms both Dragonfly and Dragonfly-MF. Thereby, parallelizing across instances provides an additional speed-up compared to the single-instance setting. Somewhat surprisingly Dragonfly-MF is slower in terms of convergence than Dragonfly on the NASBench201 benchmark, which indicates that it cannot model the performance across fidelities well. For the multi-instance scenario, ASHA (stopping and promotion) performs better than synchronous HB and BOHB on the discrete NASBench201 benchmarks. But, due to its random sampling, it often does not converge to a sensible performance on the continuous MLP problem. We attributed S-BOHB poor performance on the NASBench201 datasets to its kernel density estimator which seems to struggle with discrete spaces. As described above, the stopping variant of MOBSTER works better in the beginning, but the promotion variant achieves a better final performance. However, both variants of MOBSTER outperform all other baselines.

ResNet on CIFAR-10. Based on the previous results we limit our comparison to ASHA as well as S-HB and S-BOHB. MOBSTER achieves consistently a lower regret than all three baselines.

LSTM on WikiText-2. Figure 6 left shows results for LSTM tuning on the WikiText-2 dataset. To save computations, we limited the comparison to MOBSTER, ASHA, and S-BOHB. Note that a single run (out of 5 per method) takes roughly 5 GPU days. Training a single neural network for the full 81 epochs takes roughly 12 hours, and S-BOHB is only able to execute a single bracket in the allocated time budget. Due to limited data, it cannot fit a surrogate model at the highest resource level r_{max}, which may explain its overall poor performance. Because of its asynchronous scheduling, both MOBSTER and ASHA perform substantially better. After around 7000 seconds, MOBSTER's model kicks in, allowing it to approach the optimum much faster than random-based ASHA.

5 CONCLUSIONS

We present a novel method, that combines asynchronous scheduling with multi-fidelity BO. Instead of sampling new configurations at random, they are chosen based on a joint Gaussian process surrogate model over resource levels, and asynchronous scheduling is accommodated by fantasizing the outcome of pending candidates. On a range of challenging benchmarks, we compare our proposal to prior synchronous and asynchronous HNAS state-of-the-art. MOBSTER significantly outperforms synchronous BOHB and AHSA. The former is slowed down by synchronization overhead, using its parallel resources inefficiently at higher resource levels. ASHA on the other hand does not suffer from synchronization points, but compared to MOBSTER, we often find it to plateau before solutions with state-of-the-art performance are found.

REFERENCES

A. Alvin, B. Ru, J. P. Calliess, S. Roberts, and M. A. Osborne. Asynchronous batch Bayesian optimisation with improved local penalisation. In *Proceedings of the 36th International Conference on Machine Learning (ICML'19)*, 2019.

J. Bergstra, R. Bardenet, Y. Bengio, and B. Kegl. Algorithms for hyperparameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems 24*, pp. 2546–2554. Curran Associates, 2011a.

J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In *Proceedings of the 24th International Conference on Advances in Neural Information Processing Systems (NIPS'11)*, 2011b.

P. Chrabaszcz, I. Loshchilov, and F. Hutter. A downsampled variant of ImageNet as an alternative to the CIFAR datasets. *arXiv:1707.08819 [cs.CV]*, 2017.

K. Cutajar, M. Pullin, A. Damianou, N. Lawrence, and J. Gonzáles. Deep Gaussian processes for multi-fidelity modeling. Technical report, NeurIPS Workshop on Bayesian Deep Learning, 2018.

T. Domhan, J. T. Springenberg, and F. Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI'15)*, 2015.

X. Dong and Y. Yang. Nas-bench-201: Extending the scope of reproducible neural architecture search. *arXiv:2001.00326 [cs.CV]*, 2020.

S. Falkner, A. Klein, and F. Hutter. BOHB: Robust and efficient hyperparameter optimization at scale. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, 2018.

D. Golovin, B. Solnik, S. Moitra, G. Kochanski, J. Karro, and D. Sculley. Google vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*, 2016.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 1997.

K. Jamieson and A. Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS'16)*, 2016.

D. Jones, M. Schonlau, and W. Welch. Efficient global optimization of expensive black box functions. *Journal of Global Optimization*, 1998.

K. Kandasamy, G. Dasarathy, J. Schneider, and B. Póczos. Multi-fidelity bayesian optimisation with continuous approximations. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, 2017.

K. Kandasamy, A. Krishnamurthy, J. Schneider, and B. Poczos. Parallelised Bayesian optimisation via Thompson sampling. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS'18)*, 2018a.

K. Kandasamy, A. Krishnamurthy, J. Schneider, and B. Poczos. Parallelised bayesian optimisation via thompson sampling. In *In International Conference on Artificial Intelligence and Statistics (AISTATS'18)*, 2018b.

K. Kandasamy, K. R. Vysyaraju, W. Neiswanger, B. Paria, C. R. Collins, J. Schneider, B. Poczos, and E. P. Xing. Tuning hyperparameters without grad students: Scalable and robust bayesian optimisation with dragonfly. In *Journal of Machine Learning Research 2020, Special Issue on Bayesian Optimization*, 2019.

Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *In Proceedings of the 30th International Conference on Machine Learning (ICML'13)*, 2013.

M. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR'15)*, 2015.

A. Klein, S. Falkner, S. Bartels, P. Hennig, and F. Hutter. Fast Bayesian hyperparameter optimization on large datasets. *Electronic Journal of Statistics*, 11, 2017a.

A. Klein, S. Falkner, J. T. Springenberg, and F. Hutter. Learning curve prediction with Bayesian neural networks. In *International Conference on Learning Representations (ICLR'17)*, 2017b.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: Bandit-based configuration evaluation for hyperparameter optimization. In *International Conference on Learning Representations (ICLR'17)*, 2017.

L. Li, K. Jamieson, A. Rostamizadeh, K. Gonina, M. Hardt, B. Recht, and A. Talwalkar. Massively parallel hyperparameter tuning. *arXiv:1810.05934 [cs.LG]*, 2018.

R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

S. Merity, N. S. Keskar, and R. Socher. Regularizing and optimizing LSTM language models. *arXiv:1708.02182 [cs.CL]*, 2017.

M. Poloczek, J. Wang, and P. Frazier. Multi-information source optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*. Curran Associates, 2018.

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

B. Shahriari, K. Swersky, Z. Wang, R. Adams, and N. de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 2016.

J. Snoek, H. Larochelle, and R. Adams. Practical Bayesian optimization of machine learning algorithms. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25*, pp. 2951–2959. Curran Associates, 2012.

K. Swersky, J. Snoek, and R. Adams. Multi-task Bayesian optimization. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26*. Curran Associates, 2013.

K. Swersky, J. Snoek, and R. Adams. Freeze-thaw Bayesian optimization. Technical Report arXiv:1406.3896 [stat.ML], ArXiv, 2014. URL `https://arxiv.org/abs/1406.3896`.

S. Takeno, H. Fukuoka, Y. Tsukada, T. Koyama, M. Shiga, I. Takeuchi, and M. Karasuyama. Multi-fidelity bayesian optimization with max-value entropy search and its parallelization. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, 2020.

J. Vanschoren, J. van Rijn, B. Bischl, and L. Torgo. OpenML: Networked science in machine learning. *SIGKDD Explorations*, 2014.

A. Yang, P. M. Esperança, and F. M. Carlucci. NAS valuation is frustratingly hard. In *International Conference on Learning Representations (ICLR'20)*, 2020.

C. Ying, A. Klein, E. Real, E. Christiansen, K. Murphy, and F. Hutter. NAS-Bench-101: Towards reproducible neural architecture search. *arXiv:1902.09635 [cs.LG]*, 2019.

A FURTHER DETAILS ABOUT THE MODEL

Here, we provide some additional details in order to support the main text.

A.1 POSTERIOR REPRESENTATION. FANTASIZING

We employ a standard representation of the Gaussian process posterior. Say that $k(\boldsymbol{x}, \boldsymbol{x}')$ is the kernel function, $\mu(\boldsymbol{x})$ the mean function, and the dataset $\mathcal{D}$ is given by $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, $\boldsymbol{y} \in \mathbb{R}^n$. For notational simplicity, $\boldsymbol{x}$ contains the resource attribute r here. The representation maintains $\boldsymbol{L}$ and $\boldsymbol{p}$, where

$$\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{K} + \sigma^2 \boldsymbol{I}_n, \quad \boldsymbol{p} = \boldsymbol{L}^{-1}(\boldsymbol{y} - \mu(\boldsymbol{X})).$$

Here, $\boldsymbol{K} = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)] \in \mathbb{R}^{n \times n}$ is the kernel matrix over the training inputs $\boldsymbol{X}$, σ^2 is the noise variance, and $\boldsymbol{L}$ is the Cholesky factor (lower triangular, positive diagonal).

Recall that we use *fantasizing* (Snoek et al., 2012) to deal with pending feedback. For M fantasy samples, we now have a target *matrix* $\boldsymbol{Y} \in \mathbb{R}^{n \times M}$. For any observed configuration $\boldsymbol{x}_i$, the corresponding row in $\boldsymbol{Y}$ is $y_i \boldsymbol{1}_M^T$. Now, $\boldsymbol{L}$ remains the same, and $\boldsymbol{P} = \boldsymbol{L}^{-1}(\boldsymbol{Y} - \mu(\boldsymbol{X})\boldsymbol{1}_M^T)$. Importantly, the dominating computation of $\boldsymbol{L}$ is independent of the number M of fantasy samples. Predictions on test points result in M different posterior means, but common posterior variances.

In practice, the surrogate model (kernel and mean function) comes with hyperparameters (*e.g.*, noise variance σ^2, length scales and variance of kernel $k_{\mathcal{X}}$, $\alpha, \beta, \gamma, \delta$), which have to be adjusted. All in all, recomputing the posterior representation involves (a) optimizing the hyperparameters on the observed dataset only, and (b) drawing fantasy targets, then computing the representation on the extended dataset, including pending configurations. Here, (a) is substantially more expensive than (b), as it involves computing the representation many times during the optimization.

In our implementation, we delay computations of the posterior representation until the next configuration has to be chosen. If running tasks report metric values, they are inserted into the dataset, replacing pending configurations entered there before. Our implementation offers a number of strategies to skip the expensive part (a) of the computation. Here, (a) is not skipped initially, until the dataset size passes a threshold. After that, we can skip (a) for all but every k-th update. Another strategy is to skip (a) as long as the number of *full resource* datapoints $(\boldsymbol{x}_i, r_i = r_{\text{max}})$ does not grow. This ensures that we do not spend more time on surrogate model updates in the multi-fidelity case than in standard BO with full resource evaluations only.

A.2 EXPONENTIAL DECAY SURROGATE MODEL

The model proposed by Swersky et al. (2014) is based on the assumption $f(\boldsymbol{x}, r) = f(\boldsymbol{x}) + e^{-\lambda r}$, where $\lambda \sim p(\lambda)$ is drawn from a Gamma prior, and the asymptote $f(\boldsymbol{x})$ (*i.e.*, $f(\boldsymbol{x}, r)$ as $r \to \infty$) from a GP prior. They end up using a zero-mean GP prior with kernel function $k((\boldsymbol{x}, r), (\boldsymbol{x}', r')) = k_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}') + k_{\mathcal{R}}(r, r')$ and

$$k_{\mathcal{R}}(r, r') := \int_0^\infty e^{-\lambda r} e^{-\lambda r'} p(\lambda)\, d\lambda = \kappa(r + r'),$$
$$\kappa(u) := \frac{\beta^\alpha}{(u + \beta)^\alpha}, \quad \alpha, \beta > 0.$$

This proposal has several shortcomings. First, $f(\boldsymbol{x}, r) \to f(\boldsymbol{x}) + 1$ as $r \to 0$, independent of what metric is used. Second, the random function $e^{-\lambda r}$ assumption implies a non-zero mean function, and a kernel taking this into account. A better assumption would be $f(\boldsymbol{x}, r) = f(\boldsymbol{x}) + \gamma e^{-\lambda r}$, $\gamma > 0$, which implies a mean function $\mu(\boldsymbol{x}, r) = \mu_{\mathcal{X}}(\boldsymbol{x}) + \gamma\kappa(r)$ and a covariance function

$$k((\boldsymbol{x}, r), (\boldsymbol{x}', r')) = k_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}') + \gamma^2 \tilde{k}_{\mathcal{R}}(r, r'), \tag{1}$$
$$\tilde{k}_{\mathcal{R}}(r, r') := \kappa(r + r') - \kappa(r)\kappa(r').$$

One advantage of this additive model structure is that an implied conditional independence relation can be exploited in order to speed up inference computations (Swersky et al., 2014). However, this model implies $f(\boldsymbol{x}, r) = f(\boldsymbol{x}) + \gamma$ as $r \to 0$: the metric for no training at all depends on $\boldsymbol{x}$ in the same way as the asymptote. A more sensible assumption would be $f(\boldsymbol{x}, r) = \gamma$ as $r \to 0$, since

Table 1: The hyperparameters and architecture choices for the fully connected networks.

Hyperparameter	Range
learning rate	$[10^{-6}, 1]$
batch size	$[2^3, 2^7]$
weight decay	$[10^{-8}, 1]$
dropout layer 1	$[0, .99]$
dropout layer 2	$[0, .99]$
units layer 1	$[2^4, 2^{10}]$
units layer 2	$[2^4, 2^{10}]$
scale layer 1	$[10^{-3}, 10]$
scale layer 2	$[10^{-3}, 10]$

without training, predictions should be random (*e.g.*, $\gamma = \frac{1}{2}$ for binary classification). We present a novel surrogate model which can incorporate this assumption. It is based on

$$f(\boldsymbol{x}, r) = \gamma e^{-\lambda r} + f(\boldsymbol{x}) \left(1 - \delta e^{-\lambda r}\right), \quad \delta \in [0, 1]. \tag{2}$$

Some algebra (provided below) gives

$$\begin{aligned}
\mu(\boldsymbol{x}, r) &= \gamma\kappa(r) + \mu_{\mathcal{X}}(\boldsymbol{x})(1 - \delta\kappa(r)), \\
k((\boldsymbol{x}, r), (\boldsymbol{x}', r')) &= (\gamma - \delta\mu_{\mathcal{X}}(\boldsymbol{x}))\tilde{k}_{\mathcal{R}}(r, r')(\gamma - \delta\mu_{\mathcal{X}}(\boldsymbol{x}')) \\
&\quad + k_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}')[1 - \delta(\kappa(r) + \kappa(r') \\
&\quad - \delta\kappa(r + r'))].
\end{aligned}$$

For $\delta = 0$, we recover the additive model of Eq. 1, while for $\delta = 1$, we have that $f(\boldsymbol{x}, r) = \gamma$ as $r \to 0$, independent of $\boldsymbol{x}$. This model comes with hyperparameters $\alpha, \beta, \gamma > 0$ and $\delta \in [0, 1]$. We provide an empirical analysis of this model in Section 4.1 of the main paper.

Derivation of Exponential Decay Surrogate Model

Recall that our novel surrogate model $k((\boldsymbol{x}, r), (\boldsymbol{x}', r'))$ and $\mu(\boldsymbol{x}, r)$ is based on the random function assumption (2). Since $\mathbb{E}[e^{-\lambda r}] = \kappa(r)$, it is clear that $\mathbb{E}[f(\boldsymbol{x}, r)] = \gamma\kappa(r) + \mu_{\mathcal{X}}(\boldsymbol{x})(1 - \delta\kappa(r)) = \mu_{\mathcal{X}}(\boldsymbol{x}) + (\gamma - \delta\mu_{\mathcal{X}}(\boldsymbol{x}))\kappa(r)$. Moreover,

$$\begin{aligned}
&k((\boldsymbol{x}, r), (\boldsymbol{x}', r')) = \mathbb{E}\left[f(\boldsymbol{x}, r) f(\boldsymbol{x}', r')\right] \\
&= (\gamma - \delta\mu_{\mathcal{X}}(\boldsymbol{x}))\tilde{k}_{\mathcal{R}}(r, r')(\gamma - \delta\mu_{\mathcal{X}}(\boldsymbol{x}')) + k_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}') \\
&\cdot\ \mathbb{E}\left[\left(1 - \delta e^{-\lambda r}\right)\left(1 - \delta e^{-\lambda r'}\right)\right] = (\gamma - \delta\mu_{\mathcal{X}}(\boldsymbol{x})) \\
&\cdot\ \tilde{k}_{\mathcal{R}}(r, r')(\gamma - \delta\mu_{\mathcal{X}}(\boldsymbol{x}')) + k_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}') \\
&\cdot\ (1 - \delta(\kappa(r) + \kappa(r') - \delta\kappa(r + r'))), \\
&\tilde{k}_{\mathcal{R}}(r, r') = \kappa(r + r') - \kappa(r)\kappa(r').
\end{aligned}$$

B Configuration Spaces

Table 1 shows the configuration space for the MLP benchmark, Table 2 for the residual network benchmark, Table 4 for NASBench201 and Table 3 for the LSTM benchmark. If we use an exponential notation than we optimized the corresponding hyperparameter on a logarithmic scale. We used a one-hot encoding for all categorical hyperparameters.

C Additional NASBench201 Experiments

Table 2: The configuration space for the residual neural network benchmark.

Hyperparameter	Range
learning rate	$[1^{-3}, 1^{-1}]$
batch size	$[8, 256]$
weight decay	$[10^{-5}, 10^{-3}]$
momentum	$[0, .99]$

Table 3: The configuration space for the LSTM benchmark.

Hyperparameter	Range
learning rate	$[1, 50]$
batch size	$[2^3, 2^7]$
dropout	$[0, 0.99]$
gradient clipping	$[0.1, 2]$
learning rate factor	$[1, 100]$

Table 4: The configuration space for NASBench201.

Hyperparameter	Range
edge 0	$\{none, skip\text{-}connect, nor\text{-}conv\text{-}1x1, nor\text{-}conv\text{-}3x3, avg\text{-}pool\text{-}3x3\}$
edge 1	$\{none, skip\text{-}connect, nor\text{-}conv\text{-}1x1, nor\text{-}conv\text{-}3x3, avg\text{-}pool\text{-}3x3\}$
edge 2	$\{none, skip\text{-}connect, nor\text{-}conv\text{-}1x1, nor\text{-}conv\text{-}3x3, avg\text{-}pool\text{-}3x3\}$
edge 3	$\{none, skip\text{-}connect, nor\text{-}conv\text{-}1x1, nor\text{-}conv\text{-}3x3, avg\text{-}pool\text{-}3x3\}$
edge 4	$\{none, skip\text{-}connect, nor\text{-}conv\text{-}1x1, nor\text{-}conv\text{-}3x3, avg\text{-}pool\text{-}3x3\}$
edge 5	$\{none, skip\text{-}connect, nor\text{-}conv\text{-}1x1, nor\text{-}conv\text{-}3x3, avg\text{-}pool\text{-}3x3\}$

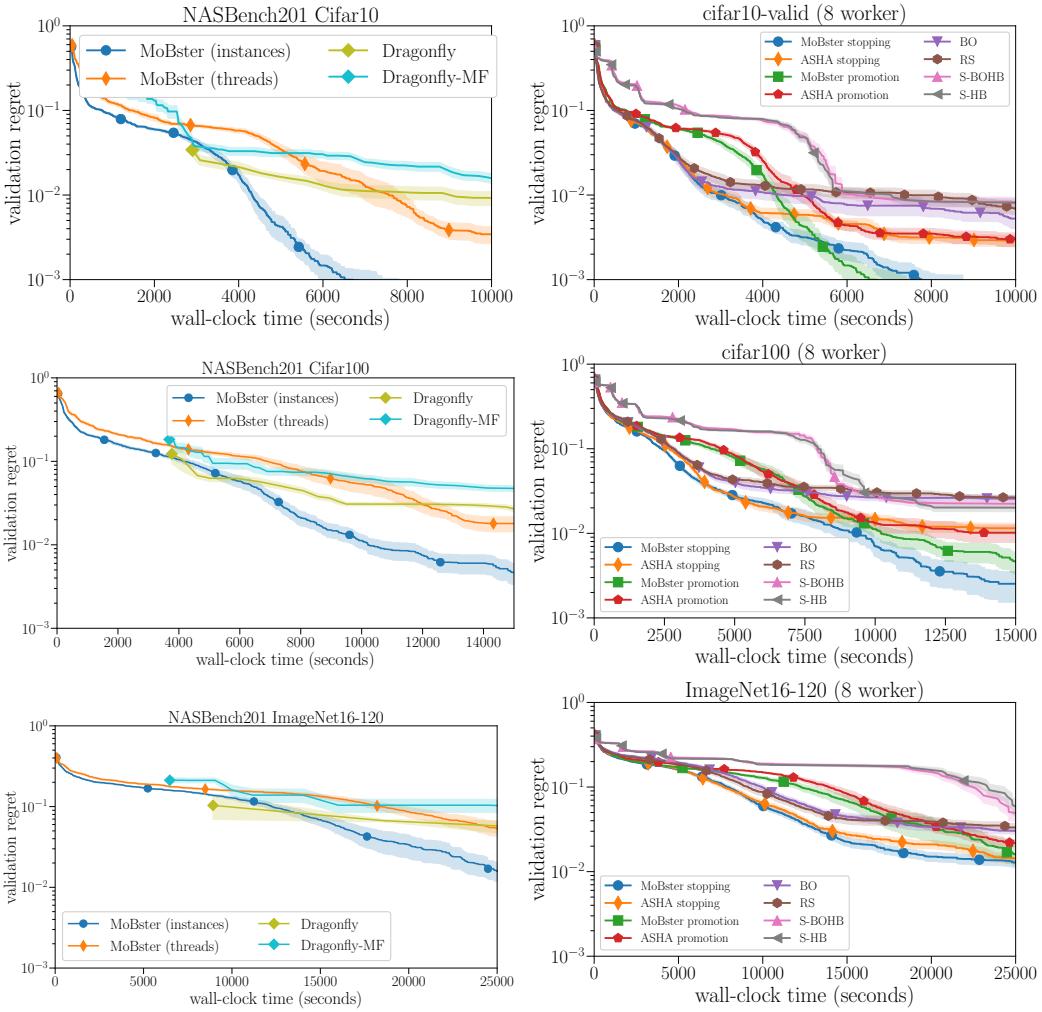

Figure 7: Comparison to various baselines across multiple processes on a single instance (left) and across instances (right) for the three different NASBench201 benchmarks. See main text for an analysis.