
A Quantile-based Approach for Hyperparameter Transfer Learning

David Salinas¹ Huibin Shen² Valerio Perrone²

Abstract

Bayesian optimization (BO) is a popular methodology to tune the hyperparameters of expensive black-box functions. Traditionally, BO focuses on a single task at a time and is not designed to leverage information from related functions, such as tuning performance objectives of the same algorithm across multiple datasets. In this work, we introduce a novel approach to achieve transfer learning across different *datasets* as well as different *objectives*. The main idea is to regress the mapping from hyperparameter to objective quantiles with a semi-parametric Gaussian Copula distribution, which provides robustness against different scales or outliers that can occur in different tasks. We introduce two methods to leverage this mapping: a Thompson sampling strategy as well as a Gaussian Copula process using such quantile estimate as a prior. We show that these strategies can combine the estimation of multiple objectives such as latency and accuracy, steering the hyperparameters optimization toward faster predictions for the same level of accuracy. Extensive experiments demonstrate significant improvements over state-of-the-art methods for both hyperparameter optimization and neural architecture search.

1. Introduction

Tuning complex machine learning models such as deep neural networks can be daunting. Object detection or language understanding models often rely on deep neural networks with many tunable hyperparameters, and automatic hyperparameter optimization (HPO) techniques such as Bayesian optimization (BO) are critical to find good hyperparameters in short time. BO addresses the black-box optimization

¹NAVER LABS Europe (work started while being at Amazon) ²Amazon Web Services. Correspondence to: David Salinas <david.salinas@naverlabs.com>, Huibin Shen <huibishe@amazon.com>, Valerio Perrone <vperrone@amazon.com>.

problem by placing a probabilistic model, typically a Gaussian process (GP), on the function to minimize; then, the hyperparameters to evaluate next are determined through an acquisition function that trades off exploration and exploitation. Interest in BO has been originally motivated by speeding up HPO pipelines where the function to optimize usually takes hours to evaluate or even thousand of GPU days in total in the case of neural architecture search (NAS) (Zoph & Le, 2017; Real et al., 2018b). While traditional BO focuses on each problem in isolation, recent years have seen a surge of interest in *transfer learning* for HPO. The key idea is to exploit evaluations from previous, related *tasks* (e.g., the same neural network tuned on multiple datasets) to further speed up the hyperparameter search.

A key challenge for joint models is that different black-boxes exhibit heterogeneous scale and noise levels (Bardenet et al., 2013; Yogatama & Mann, 2014; Wistuba et al., 2018; Feurer et al., 2018). The straightforward approach of standardizing outputs (Yogatama & Mann, 2014) only works for tasks with normally distributed observations and no outliers. While rank estimation can be used to alleviate scale issues, the difficulty of feeding back rank information to GPs leads to restricting assumptions. For instance, Bardenet et al. (2013) does not model rank estimation uncertainty, while Feurer et al. (2018) uses independent GPs removing the adaptivity of the GP to the current task.

This paper shows how semi-parametric Gaussian Copulas effectively handle heterogeneous scales across tasks, giving rise to several algorithmic instantiations for hyperparameter transfer learning. Our key contributions are as follows:

- We propose using Gaussian Copulas instead of standardization to map observations from different tasks to comparable distributions;
- Two novel methods leveraging this finding, namely a Thompson sampling and Gaussian Copula process combined with a joint parametric prior;
- An extensive empirical study demonstrating substantial improvements over state-of-the-art transfer learning methods on real-world datasets, including on neural architecture search (NAS);
- A simple extension that scalarizes Gaussian Copula

objectives to achieve multi-objective Bayesian optimization.

2. Related work

A variety of methods have been developed to induce transfer learning in HPO. The most common approach is to model tasks jointly or via a conditional independence structure, which has been explored through multi-output GPs (Swersky et al., 2013), weighted combination of GPs (Schilling et al., 2016; Wistuba et al., 2018; Feurer et al., 2018), and neural networks, either fully Bayesian (Sprinzenberg et al., 2016) or hybrid (Snoek et al., 2015; Perrone et al., 2018; Law et al., 2018). A different line of research has focused on the setting where tasks come over time as a sequence and the models need to be updated online as new problems accrue. A way to achieve this is to fit a sequence of surrogate models to the residuals relative to predictions of the previously fitted model (Golovin et al., 2017; Poloczek et al., 2016). Specifically, the GP over the new task is centered on the predictive mean of the previously learned GP. Finally, rather than fitting a surrogate model to all past data, some transfer can be achieved by warm-starting BO with the solutions to the previous BO problems (Feurer et al., 2015; Wistuba et al., 2015b).

Some methods have instead focused on the search-space level, aiming to prune it to focus on regions of the hyperparameter space where good configurations are likely to lie. An example is Wistuba et al. (2015a), where related tasks are used to learn a promising search space during HPO, defining task similarity in terms of the distance of the respective dataset meta-features. An alternative was proposed in Perrone et al. (2019), where a restricted search space in the form of a low-volume hyper-rectangle or hyper-ellipsoid is learned from the optimal hyperparameters of related tasks. Gaussian Copula Process (GCP) (Wilson & Ghahramani, 2010) can also be used to alleviate scale issues on a single task at the extra cost of estimating the CDF of the data. Using GCP for HPO was proposed in Anderson et al. (2017) to handle potentially non-Gaussian data, albeit only considering non-parametric homoskedastic priors for the single task and single objective case.

3. Gaussian Copula regression

For each task j denote with $f^j : \mathbb{R}^p \rightarrow \mathbb{R}$ the error function one wishes to minimize, and with $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N_j}$ the evaluations available for an arbitrary task with $y_i = f(x_i)$. Given the evaluations on M tasks

$$\mathcal{D}^M = \bigcup_{j=1}^M \{x_i^j, y_i^j\}_{i=1}^{N_j},$$

we are interested in speeding up the optimization of an arbitrary new task f , namely in finding $\arg \min_{x \in \mathbb{R}^p} f(x)$ in the least number of evaluations. We assume the functions $\{f^j\}_{j=1}^M$ and f to be *related*, such as the error function of the same algorithm over several datasets. In the following, we refer to *task* as the problem of tuning a given algorithm on a dataset, with different datasets corresponding to related tasks.

One possible approach to speed up the optimization of f is to build a surrogate model $\hat{f}(x)$. While using a Gaussian process is possible, scaling this approach to the large number of evaluations available in a transfer learning setting is challenging. Instead, we propose fitting a parametric estimate of the distribution of $\hat{f}_\theta(x)$ which can be later used in HPO strategies as a prior of a Gaussian Copula Process. A key requirement here is to learn a joint model, namely we would like to find θ which fits well all observed tasks f^j . We show how this can be achieved with a semi-parametric Gaussian Copula in two steps. First, all evaluations are mapped to quantiles with the empirical CDF. Then, we fit a parametric Gaussian distribution on quantiles mapped through the Gaussian inverse CDF.

We make the modeling assumption that every y_i comes from the same distribution for an arbitrary task. The probability integral transform results in $u_i = F(y_i)$, where F is the cumulative distribution function of y . The CDF of (u_1, \dots, u_N) is modeled with a Gaussian Copula:

$$C(u_1, \dots, u_N) = \phi_{\mu, \Sigma}(\Phi^{-1}(F(y_1)), \dots, \Phi^{-1}(F(y_N))),$$

where Φ is the standard normal CDF and $\phi_{\mu, \Sigma}$ is the CDF of a normal distribution parametrized by μ and Σ . Assuming F to be invertible, we define the change of variable $z = \Phi^{-1} \circ F(y) = \psi(y)$ and $g = \psi \circ f$, see Figure 1. Note that if z is regressed perfectly against x , then finding the minimum of f is trivial as a parameter x minimizing $\psi(f(x))$ also minimizes $f(x)$ since ψ is monotonically increasing.

The conditional distribution of $P(z|x)$ is estimated with a Gaussian distribution whose mean and variance are two deterministic parametric functions given by

$$\begin{aligned} P(z|x) &\sim \mathcal{N}(\mu_\theta(x), \sigma_\theta^2(x)) \\ &= \mathcal{N}(w_\mu^T h_{w_h}(x) + b_\mu, (\Psi(w_\sigma^T h_{w_h}(x) + b_\sigma))^2), \end{aligned}$$

where $h_{w_h}(x) \in \mathbb{R}^d$ is the output of a multi-layer perceptron (MLP), $w_h, w_\mu \in \mathbb{R}^d, b_\mu \in \mathbb{R}, w_\sigma \in \mathbb{R}^d, b_\sigma \in \mathbb{R}$ are free parameters and $\Psi(t) = \log(1 + \exp t)$ is an activation mapping to positive numbers. The parameters $\theta = \{w_h, w_\mu, b_\mu, w_\sigma, b_\sigma\}$ are learned by minimizing the Gaussian negative log-likelihood on the available evaluations $\mathcal{D}^M = \bigcup_{1 \leq j \leq M} \{x_i^j, z_i^j\}_{i=1}^{N_j}$, e.g., by minimizing with SGD

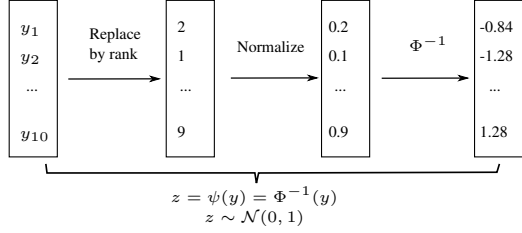


Figure 1. Illustration of the steps to compute of $z = \psi(y)$. The empirical CDF is computed on each task individually and truncated with Winsorized cut-off.

$$\sum_{(x,z) \in \mathcal{D}^{\mathcal{M}}} \frac{1}{2} \log 2\pi\sigma_{\theta}(x)^2 + \frac{1}{2} \left(\frac{z - \mu_{\theta}(x)}{\sigma_{\theta}(x)} \right)^2 \quad (1)$$

Let us make two remarks. First, the parameters θ are *independent* of the task chosen so that minimizing Eq. (1) gives a joint model over all available tasks, the hope being it can generalize to new tasks, see Figure 2. Second, each task is assumed to have the same number of observations. If this is not the case, each term in Eq. (1) can be weighted inversely to the number of task observations.

The transformation ψ requires F , which needs to be estimated. Rather than using a parametric or density estimation approach, we use the empirical CDF $\tilde{F}(t) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{y_i \leq t}$. While it has the advantage of being non-parametric, this estimator leads to infinite value when evaluating ψ at the minimum or maximum of y . To avoid this issue, we use the Winsorized cut-off estimator

$$F(t) \approx \begin{cases} \delta_N & \text{if } \tilde{F}(t) < \delta_N \\ \tilde{F}(t) & \text{if } \delta_N \leq \tilde{F}(t) \leq 1 - \delta_N \\ 1 - \delta_N & \text{if } \tilde{F}(t) > 1 - \delta_N \end{cases}$$

where N is the number of observations of y and choosing $\delta_N = \frac{1}{4N^{1/4}\sqrt{\pi \log N}}$ strikes a bias-variance trade-off (Liu et al., 2009). This approach is semi-parametric in that the CDF is estimated with a non-parametric estimator and the Gaussian Copula is estimated with a parametric approach.

The benefit of using a non-parametric estimator for the CDF is that it allows us to map the observations of each task to comparable distributions. Indeed, each $u^j = F^j(y^j)$ is uniformly distributed by the probability integral transform property. Since $z^j = \Phi^{-1}(u_j)$, all tasks j are normally distributed, namely $z^j \sim \mathcal{N}(0, 1)$ for all j . This is critical to allow the joint learning of the parametric estimates μ_{θ} and σ_{θ} , which share their parameters θ across all tasks.

Another advantage of this view is that one can easily assess whether transfer learning helps. Indeed, a constant predictor $\hat{z} = 0$ yields a RMSE of 1 as

$$\begin{aligned} \text{RMSE}(\hat{z})^2 &= \mathbf{E}[(z - \hat{z})^2] \\ &= \mathbf{E}[z^2] - 2 \mathbf{E}[z] \mathbf{E}[\hat{z}] + \mathbf{E}[\hat{z}^2] \\ &= 1 \end{aligned}$$

using the independence of z and \hat{z} , and the fact $z \sim \mathcal{N}(0, 1)$. Our experiments will show that one can regress a parametric estimate that has a RMSE lower than 1. This means that information can be leveraged from the evaluations obtained on related tasks, compared to the result of the constant predictor which would be the best predictor if no information was given (assuming of course absence of overfitting).

4. Copula based HPO

We now show how the estimator introduced in the previous section can be leveraged to design two novel HPO strategies. We first introduce Copula Thompson sampling (CTS), a simple method to exploit information from related tasks. We then build on it to develop Gaussian Copula Process, which can additionally adapt to the new task.

4.1. Copula Thompson sampling

Given the predictive distribution $P(z|x) \sim \mathcal{N}(\mu_{\theta}(x), \sigma_{\theta}^2(x))$, we can derive a Thompson sampling strategy (Thompson, 1933) exploiting knowledge from previous tasks. Given N candidate hyperparameter configurations x_1, \dots, x_N , we sample from each predictive distribution $\tilde{z}_i \sim \mathcal{N}(\mu_{\theta}(x_i), \sigma_{\theta}^2(x_i))$ and then evaluate $f(x_i)$ where $i = \arg \min_i \tilde{z}_i$. Pseudo-code is given in Algorithm 1.

While this approach can re-use information from previous tasks, it does not exploit the evaluations from the current task as each draw is independent of the observed evaluations. This can become an issue if the new black-box significantly differs from previous tasks. We now show that Gaussian Copula regression can be combined with a GP to learn from previous tasks while also adapting to the current task.

Algorithm 1 Copula Thompson sampling

Learn the parameters θ of $\mu_{\theta}(x)$ and $\sigma_{\theta}(x)$ on hold-out evaluations \mathcal{D}^M by minimizing (1).

while Has budget **do**

Sample N candidate hyperparameters x_1, \dots, x_N from the search space

Draw $\tilde{z}_i \sim \mathcal{N}(\mu_{\theta}(x_i), \sigma_{\theta}^2(x_i))$ for $i = 1, \dots, N$

Evaluate $f(x_i)$ where $i = \arg \min_i \tilde{z}_i$

end while

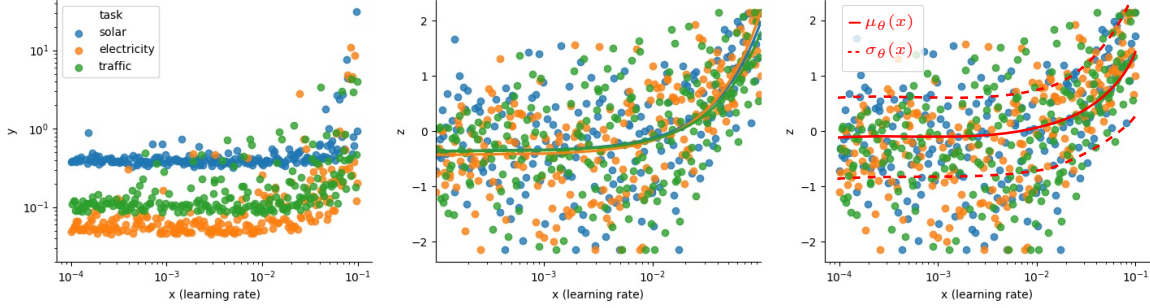


Figure 2. Plot of x (single learning rate in this example) against the blackbox error y in log-space (left) and $z = \Phi^{-1} \circ F(y)$ (middle and right) where F is approximated with the empirical CDF of every task. The middle plot shows a running mean fitted for every task individually. The last plot pictures the predictive distribution of a parametric prior given by an MLP $\mu_\theta(x), \sigma_\theta(x)$. Note that since θ is tied across all tasks, the predictions are the same for all tasks which can only work when distributions are similar across tasks.

4.2. Gaussian Copula Process

Instead of modeling observations with a GP, we model them as a Gaussian Copula Process (GCP) (Wilson & Ghahramani, 2010). Observations are mapped through the bijection $\psi = \Phi^{-1} \circ F$, where we recall that Φ is the standard normal CDF and that F is the CDF of y . As ψ is monotonically increasing and mapping into the line, we can alternatively view this modeling as a warped GP (Snelson et al., 2004) with a non-parametric warping. One advantage of this transformation is that $z = \psi(y)$ follows a normal distribution, which may not be the case for y . In the specific case of HPO, for instance y may represent accuracy scores in $[0, 1]$ of a classifier where a Gaussian cannot be used. Furthermore, we can use the information gained on other tasks with μ_θ and σ_θ by using them as prior mean and standard deviation. To do so, the following residual is modeled with a GP:

$$r(x) = \frac{g(x) - \mu_\theta(x)}{\sigma_\theta(x)} \\ \sim \text{GP}(m(x), k(x, x')),$$

where $g = \psi \circ f$. We use a Matérn-5/2 covariance kernel with automatic relevance determination hyperparameters, and optimize the GP hyperparameters by type-II maximum likelihood (Rasmussen & Williams, 2006). Categorical hyperparameters are handled by one-hot encoding. Fitting the GP gives the predictive distribution of the residual surrogate

$$\hat{r}(x) \sim \mathcal{N}(\mu_r(x), \sigma_r^2(x)).$$

Because μ_θ and σ_θ are deterministic functions, the predictive distribution of the surrogate \hat{g} is given by

$$\hat{g}(x) = \hat{r}(x)\sigma_\theta(x) + \mu_\theta(x) \\ \sim \mathcal{N}(\mu_{\hat{g}}(x), \sigma_{\hat{g}}^2(x)) \\ \sim \mathcal{N}(\mu_r(x)\sigma_\theta(x) + \mu_\theta(x), (\sigma_r(x)\sigma_\theta(x))^2).$$

Using this predictive distribution, we can select the hyperparameter configuration maximizing the Expected Improvement (EI) (Mockus et al., 1978) of $g(x)$. The EI can then be defined in closed form as

$$\text{EI}(x) = \mathbf{E}[\max(0, g(x_{\min}) - \hat{g}(x))] \\ = \sigma_{\hat{g}}(x)(v(x)\Phi(v(x)) + \phi(v(x))),$$

Algorithm 2 Gaussian Copula process with parametric prior

Learn the parameters θ of $\mu_\theta(x)$ and $\sigma_\theta(x)$ on hold-out evaluations \mathcal{D}^M by minimizing (1).

Sample an initial set of evaluations $\mathcal{D} = \{(x_i, f(x_i))\}_{i=1}^{N_0}$ via CTS.

while Has budget **do**

Estimate CDF F on the current task observations $\{f(x_i)\}$ to obtain ψ

Fit the GP surrogate \hat{r} to the observations $\{(x, \frac{\psi(y) - \mu_\theta(x)}{\sigma_\theta(x)}) \mid (x, y) \in \mathcal{D}\}$

Sample N candidate hyperparameters x_1, \dots, x_N from the search space

Compute the hyperparameter maximizing EI $x^* = \arg \max \text{EI}$

Evaluate $y^* = f(x^*)$ and update $\mathcal{D} = \mathcal{D} \cup \{(x^*, y^*)\}$.

end while

where $v(x) := \frac{\mu_{\hat{g}}(x) - g(x_{\min})}{\sigma_{\hat{g}}(x)}$, and Φ, ϕ denote the CDF and PDF of the standard normal, respectively. When no observations are available, the empirical CDF \tilde{F} is not defined. Therefore, we warm-start the optimization on the new task by sampling a set of $N_0 = 5$ hyperparameter configurations via Thompson sampling, as described above. Pseudo-code is given in Algorithm 2.

4.3. Computational complexity

We assume that the tasks in \mathcal{D}^M contain n observations each, so that we have Mn offline evaluations in total, and

that N evaluations have been queried for the new task. Fitting a GP on all tasks costs $O((Mn + N)^3)$, which prevents from using exact approaches when many offline evaluations are available. As finding parameters of the parametric prior takes $O(Mn)$, running CTS costs $O(Mn + N)$ and running GCP with a parametric prior costs $O(Mn + N^3)$. One benefit of using a parametric prior is to avoid the cubical complexity in the number of offline evaluations. The next section demonstrates that it also improves accuracy compared to state-of-the-art HPO and transfer learning methods.

5. Experiments

We consider three algorithms in the HPO context: XGBoost (Chen & Guestrin, 2016), a 2-layer feed-forward neural network (FCNET) (Klein & Hutter, 2019), and the RNN-based time series prediction model proposed in Salinas et al. (2017) (DeepAR). As advocated in Eggenberger et al. (2012) and Klein & Hutter (2019), we compute tabular evaluations (log) uniformly beforehand on multiple datasets to compare methods with sufficiently many random repetitions. Each optimization problem is then discrete as we select from a list of precomputed solutions. While we consider hyperparameter spaces with small to moderate dimensions, optimizing the acquisition over a continuous domain is better suited in higher dimensional spaces. In this setting, GCP is readily applicable, for instance, by optimizing EI with LBFSGS. We run each experiment with 30 random seeds on AWS batch with m4.xlarge instances.

We also run experiments on NAS-Bench-201 (Dong & Yang, 2020). In this benchmark, all possible 15625 configurations of a specific cell search space were evaluated on 3 datasets. Each architecture can be represented as 6 categorical variables, each containing 5 different types of connections. As in Dong & Yang (2020), every model gets as input x the concatenation of the 6 one-hot vectors, resulting in $x \in \mathbb{R}^{30}$. Optimizers are allowed to query the black-box for 70 observations, which corresponds to roughly 12000 seconds in total. More details on each black-box can be found in Table 1, including the number of hyperparameters (HPs) for each problem. The list of the datasets is in the appendix.

The MLP $h_{w_h}(x)$ used to regress μ_θ and σ_θ has 3 layers with 50 nodes, a dropout rate of 0.1 after each hidden layer and relu activation functions. The learning rate is set to 0.01, and ADAM is run over 1000 gradient updates three times, lowering the learning rate by 5 each time with a batch size of 64. In practice, optimizing the MLP take less than two minutes on a laptop.

Baselines. We compare against random search (RS) and GP-based BO (GP), the two most popular HPO methods, as well as four transfer learning baselines. The first one is warm-start GP (WS-GP) (Feurer et al., 2015), which uses

tasks	# datasets	# HPs	# evaluations	objective
DeepAR	10	6	2420	quantile loss
FCNET	4	9	248832	MSE
XGBoost	9	9	45000	1-AUC
NAS-Bench-201	3	6	46875	accuracy, runtime

Table 1. Statistics of the black-boxes considered.

the best evaluation from all related tasks to warm-start the GP after standardizing objectives values for each dataset. The second one is AutoGP (Perrone et al., 2019), which transfers information by fitting a bounding box that contains the best hyperparameters from each previous task, and runs a GP-based BO in the learned search space. The third one is SGPT (Wistuba et al., 2018), which combines $M + 1$ GPs with a specific acquisition function using rank-matching-based weighting to transfer information across related tasks. The last baseline is ABLR (Perrone et al., 2018), a multi-task model consisting of a shared-across-task neural network with a per-task Bayesian linear regression layer on top.

For NAS, we also compare with the 4 best methods out of the 10 considered in Dong & Yang (2020): RS, REINFORCE (Williams, 1992), BOHB (Falkner et al., 2018) and REA (Real et al., 2018a).¹ The transfer learning capabilities of each method are evaluated in a leave-one-task-out setting: one dataset is sequentially left out to assess how much transfer can be achieved from the other datasets, and overall results are aggregated.

Ablation study. In addition to comparing to baselines, we perform an ablation to measure the benefits of 1) using a Gaussian Copula Process rather than sampling independently with Thompson sampling, 2) applying Gaussian Copula to normalize the data using ψ , and 3) using a parametric prior. When ψ is not used, data is instead standardized with

$$z = \Delta(y) = \frac{y - \mathbf{E}(y)}{\sqrt{\mathbf{V}(y)}}.$$

Specifically, we evaluate the following variants of our model:

- CTS: The CTS model with a parametric prior described in Section 4.2.
- TS: The model with a parametric prior described in Section 4.2 where data is standardized with Δ instead of ψ .
- GCP+prior: The GCP model with a parametric prior described in Section 4.2.

¹These baselines are not evaluated on other black-boxes as they are either designed for NAS or require multi-fidelity information that is not available for all black-boxes.

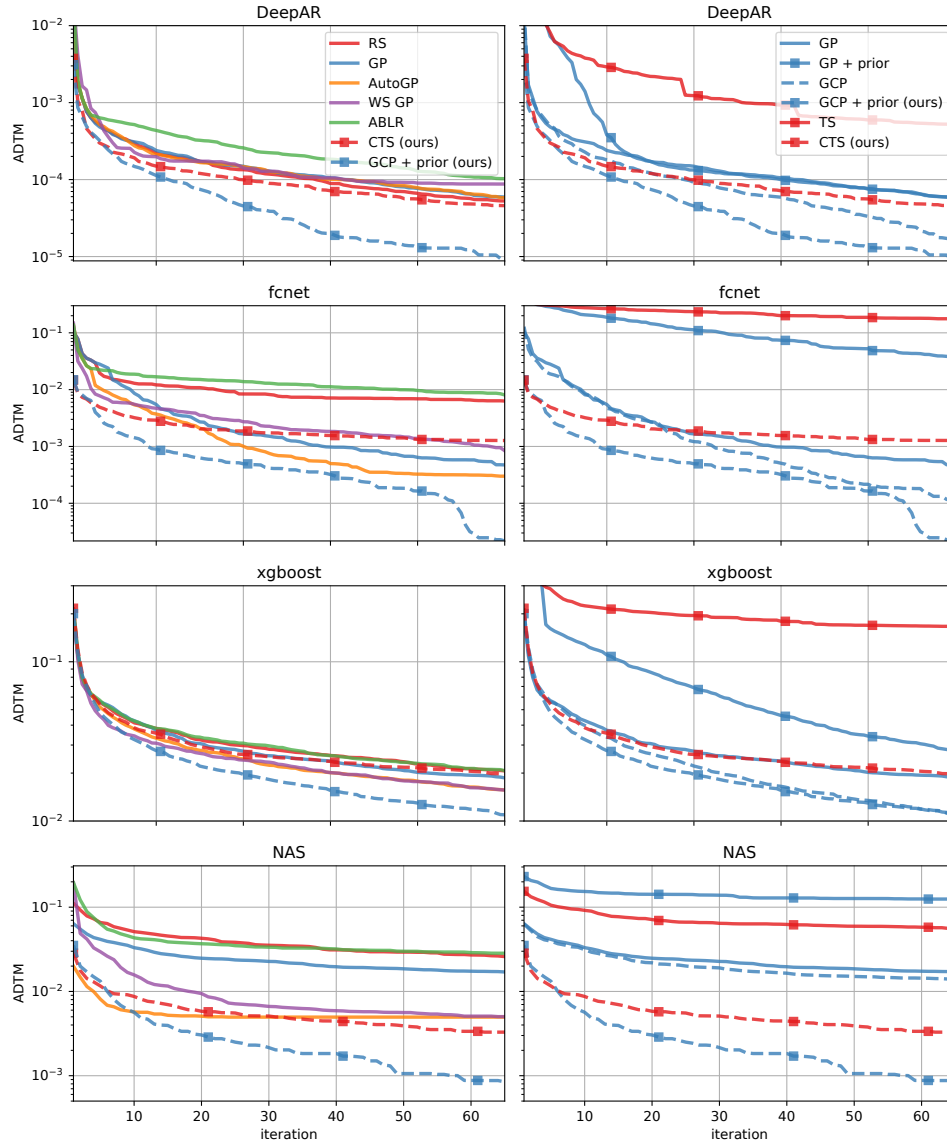


Figure 3. ADTM over iterations for baselines (left) and ablation variants (right). Methods using a parametric prior are depicted with a square and methods using Gaussian Copula are represented by a dashed line.

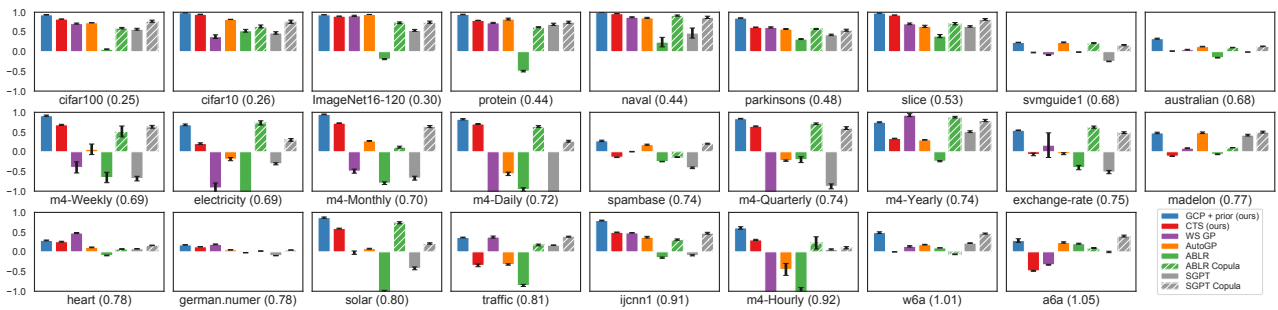


Figure 4. DTM improvement over RS (higher is better) for datasets of all black-boxes. Datasets are sorted by increasing RMSE error, which is indicated in parenthesis. Transfer learning difficulty is the lowest for the top-left task and increases when going to the right. Most transfer learning methods struggle to keep good performance as the difficulty of transfer learning increases.

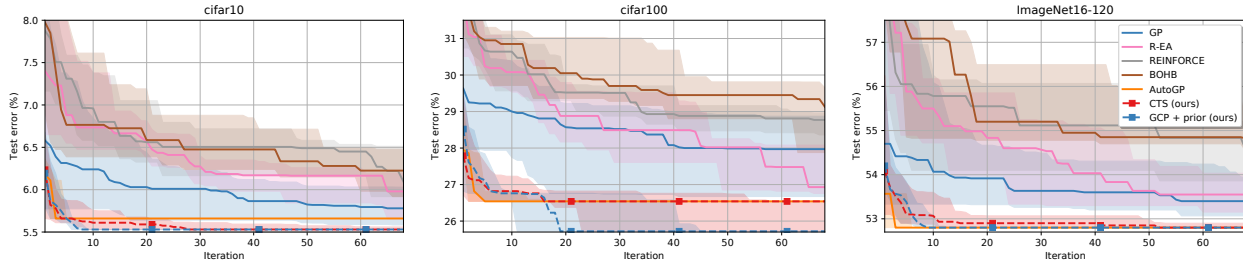


Figure 5. Performance comparison on NAS on the 3 NAS-Bench-201 datasets, with shaded areas representing 80% confidence intervals.

- GCP: A Gaussian Copula Process with a standard prior.
- GP+prior: A Gaussian Process where a parametric prior is estimated but data is standardized with Δ instead of using ψ .

Average distance to the minimum. To be able to aggregate scores over tasks, we follow the approach from Wistuba et al. (2018) and measure performance on each task in terms of the normalized distance to the global minimum. This is defined as

$$DTM_{opt}^j(t) = \frac{y_{opt}^j(t) - y_{min}^j}{y_{max}^j - y_{min}^j},$$

where $y_{opt}^j(t)$ denotes the best performance, averaged over replicates of an optimizer after t iterations on task j , while y_{min}^j and y_{max}^j respectively denote the minimum and maximum objective computed across all offline evaluations available for task j . This score is in $[0, 1]$, making performance more comparable across datasets. The average DTM across tasks is defined as $ADTM_{opt}(t) = \frac{1}{M} \sum_{j=1}^M DTM_{opt}^j(t)$.

Figure 3 illustrates the performance of competing methods over time for each black-box in terms of ADTM. We report the mean ADTM across all seeds, noting that in this transfer learning setting standard deviation would emphasize the variance coming from the different meta-datasets. Additionally, Table 2 reports the average improvement over RS, defined as the average across datasets of $\frac{1}{T} \sum_{t=1}^T \frac{DTM_{RS}^j(t) - DTM_{opt}^j(t)}{DTM_{RS}^j(t)} \in]-\infty, 1]$. This shows how much each algorithm improves over RS, whose performance indicates the complexity of the tuning problem. Figure 4 shows the improvement over random search on each dataset.

5.1. Results Discussion

Figure 3 and Table 2 show that the Copula approach gives consistent improvement over both GP and TS. In particular, GCP is a strong baseline, which is expected as the modeled data after ψ is Gaussian as opposed to a standard GP. Critically, using a parametric prior is only beneficial in combination with Gaussian Copula as evidenced by the very poor

performance of TS and GP+prior. This issue also affects ABLR and SGPT, which are unable to consistently outperform GP even though they leverage additional information from other tasks while AutoGP and WS-GP are less affected as they only use the best hyperparameters evaluation of each task. In addition, Figure 4 and Table 2 report results when using Gaussian Copula in combination with these baselines (e.g., using ψ instead of Δ to normalize outputs). The quality of these methods is dramatically improved, showing how they are negatively affected by heterogeneous scales and non-normality.

While being able to transfer information from other datasets, CTS is unable to benefit from observations of the current task and is outperformed by other baselines given sufficiently many observations, especially on DeepAR and XGBoost. On these black-boxes, we observe modest performance for the other transfer learning baselines, which we believe is due to the lower correlation of hyperparameter performance between tasks. To investigate this further, Figure 4 shows the average improvement over random search computed separately on each dataset and sorted by the prior RMSE computed on the current task² with $\sqrt{\frac{1}{N} \sum_{i=1}^N (z_i - \mu_{\theta}(x_i))^2}$. As mentioned in Section 3, low RMSE values indicate that the current task is similar to other available tasks and consequently easier for transfer learning. Both CTS and transfer learning baselines show improvement over RS when the RMSE is low, while the performance of baselines deteriorates for tasks with higher RMSE and is even adversely affected when the test task excessively differs from the held-out datasets. On the other hand, being able to benefit both from other tasks and observations of the current task, GCP+prior is the best method overall. This can be observed on all black-boxes both at the beginning and at the end of the optimization. The results are summarized in Table 2, which gives the average rank of the 16 methods. Over the 26 datasets, GCP+prior is the best method 15 times and the second best 7 times, with an average rank of 1.5.

²Observations from the current task are only used to report RMSE but are not used when fitting our model.

Table 2. DTM normalized over random search (higher is better). The best two methods are in bold and the average rank of each method is indicated in parenthesis.

black-box	DeepAR	FCNET	XGBoost	NAS
RS (baseline)	0.00 (7.1)	0.00 (10.8)	0.00 (8.2)	0.00 (12.0)
TS	-21.02 (13.0)	-563.27 (13.0)	-6.28 (12.7)	-1.30 (14.3)
CTS (ours)	0.38 (4.5)	0.83 (2.5)	0.02 (7.4)	0.88 (2.7)
GP + prior	-5.92 (11.8)	-166.64 (12.0)	-1.70 (11.1)	-2.24 (15.3)
GCP	0.42 (4.3)	0.79 (4.0)	0.31 (3.1)	0.45 (7.3)
GCP + prior (ours)	0.73 (1.7)	0.94 (1.0)	0.37 (1.9)	0.94 (1.3)
GP	-0.25 (7.9)	0.53 (8.0)	0.00 (8.6)	0.38 (8.7)
AutoGP	-0.11 (7.3)	0.72 (5.2)	0.22 (4.2)	0.84 (2.3)
WS GP	-0.50 (7.6)	0.73 (5.2)	0.11 (5.9)	0.62 (5.7)
ABLR	-0.75 (10.2)	0.11 (10.2)	-0.05 (9.1)	0.13 (10.3)
ABLR Copula	0.53 (3.1)	0.71 (5.5)	0.08 (7.0)	0.63 (5.3)
SGPT	-0.38 (8.8)	0.56 (8.2)	-0.01 (8.4)	0.46 (8.0)
SGPT Copula	0.44 (3.7)	0.74 (5.2)	0.28 (3.3)	0.67 (5.0)
BOHB	-	-	-	-0.19 (14.3)
R-EA	-	-	-	0.19 (10.3)
REINFORCE	-	-	-	-0.09 (13.0)

Comparison with NAS. Figure 5 shows test accuracy over time compared to the 4 best baselines evaluated in Dong & Yang (2020). Interestingly, GP appears to be a satisfactory baseline even though it is rarely evaluated in this context (Ying et al., 2019; Dong & Yang, 2020). When combined with a prior, both GCP and CTS converge in a fraction of the number of iterations required by the other baselines. The only exception is AutoGP, which GCP+prior still outperforms given sufficiently many observations due to its greater ability to adapt to the target task.

5.2. Multi-Objective Optimization

The goal of multi-objective optimization (MO) is to optimize multiple objectives f_1, \dots, f_L *simultaneously*. This is relevant to many applications including NAS, where the device on which the model is deployed comes with additional memory or latency constraints (Tan et al., 2018; Elsken et al., 2019). Typically, no single solution minimizes all objectives at once and one seeks instead Pareto-optimal solutions. A solution x dominates x' if $f_l(x) \leq f_l(x')$ for all $l \leq L$ and there exists $l' \leq L$ such that $f_{l'}(x) < f_{l'}(x')$. The Pareto front is the set of all Pareto-optimal solutions, defined as points that are not dominated by any other points.

A simple approach to MO is to scalarize the objective values y_1, \dots, y_L as $\sum_{l=1}^L \alpha_l y_l$ and fall back to single-objective minimization. However, combining multiple objectives poses challenges similar to the ones from the transfer learning setting. Objectives typically have different scales and mixing them linearly only allows for convex level sets, which is a poor approximation of the Pareto frontier geometry. We illustrate this behavior in Figure 6: no mixture coefficient properly approximates the Pareto front of latency and prediction error on Cifar10. On the other hand, as Binois et al. (2015) observed in the context of Pareto front

estimation, averaging the two Gaussian Copula objectives provides a good approximation of the Pareto front.

Motivated by this property, we extend our methods to MO by simply averaging observations from Gaussian Copulas. We compare with EHI (Emmerich et al., 2011), SMS (Ponweiser et al., 2008), SUR (Picheny, 2015) and EMI (Svenson & Santner, 2016), implemented in GPareto (Binois & Picheny, 2019). The suffix +MO is used to indicate a scalarization of the objective obtained by averaging observations after applying $Z = \psi$ for methods using Copula and $Z = \Delta$ for others. Performance at each BO iteration is evaluated by computing the Pareto hypervolume error, namely the hypervolume difference with the Pareto front. Consistently with the results of the previous section, linear scalarization performs poorly and both GP+prior and TS are strongly outperformed while GCP+prior and CTS compete or outperform GPareto baselines.

6. Hardware Specification

We used AWS batch with m4.xlarge instances for most of our experiments. Beside RS whose cost is almost negligible, evaluating an optimizer takes around 5 minutes for a seed. Excluding GPareto and NAS baselines, we then estimate the cost of running our experiments to be $\text{num_methods} \times \text{num_seeds} \times \text{num_datasets} \times \text{optimizer_time} \approx 12 \times 30 \times 26 \times 300$, which is around 32 days of a single machine.

7. Conclusions

We introduced a new class of methods to accelerate hyperparameter optimization by exploiting evaluations from previous tasks. The key idea was to leverage a semi-parametric Gaussian Copula prior to account for the different scale and noise levels across tasks. Experiments showed that competing methods are outperformed on both HPO and NAS, and our approach deals with heterogeneous tasks more robustly than a number of transfer learning strategies recently proposed in the literature. Finally, we showed that our framework seamlessly extends to combine multiple metrics, such as test error and latency, in a multi-objective BO framework.

A number of directions for future work are open. First, one could combine our Copula-based HPO strategies with Hyperband-style optimizers (Li et al., 2016). In addition, one could generalize our approach to deal with settings in which related problems are not limited to the same algorithm run over different datasets. This would allow for different hyperparameter dimensions across tasks, or perform transfer learning across different black-boxes.

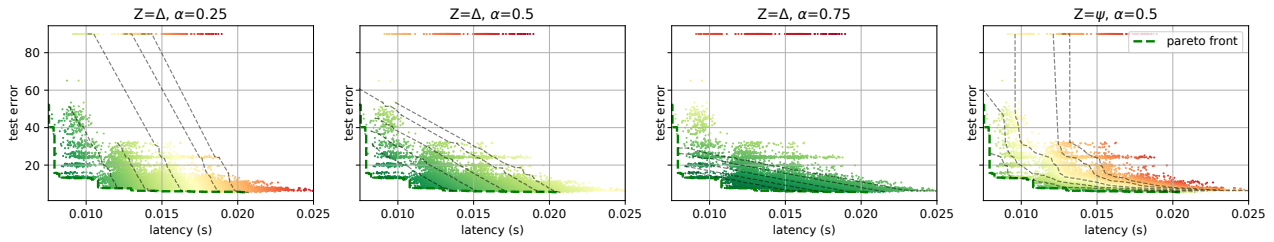


Figure 6. Latency and test error for each architecture evaluated on Cifar10. Colors represent values scalarized with $\alpha Z(y_{\text{error}}) + (1 - \alpha)Z(y_{\text{latency}})$ for $\alpha \in \{0.25, 0.5, 0.75\}$, $Z = \Delta$ for the first three plots, and $\alpha = 0.5$, $Z = \psi$ for the rightmost plot (lowest values are green, highest values are red). Level sets are linear for $Z = \Delta$ and no single value of α can approximate the geometry of the Pareto front. In contrast, ψ better approximates the shape of the front, which is followed closely by the top values of the scalarized objective.

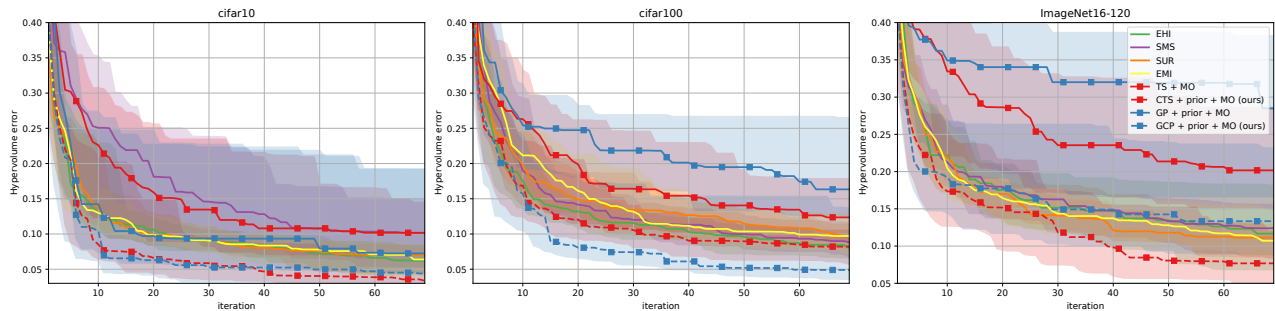


Figure 7. NAS multi-objective experiments. Hypervolume error at each iteration, with shaded areas representing 80% confidence intervals. Methods using a parametric prior are depicted with a square and methods using Gaussian Copula are represented by a dashed line.

Acknowledgements

We thank anonymous reviewers as well as Matthias Seeger, Aaron Klein and Michele Donini whose feedback greatly improved this paper.

References

- Anderson, A., Dubois, S., Cuesta-Infante, A., and Veeramachaneni, K. Sample, estimate, tune: Scaling bayesian auto-tuning of data science pipelines. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 361–372. IEEE, 2017.
- Bardenet, R., Brendel, M., Kégl, B., and Sebag, M. Collaborative hyperparameter tuning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 199–207, 2013.
- Binois, M. and Picheny, V. Gpareto: An r package for gaussian-process-based multi-objective optimization and analysis. *Journal of Statistical Software*, 89(1), 2019.
- Binois, M., Rullière, D., and Roustant, O. On the estimation of Pareto fronts from the point of view of copula theory. *Information Sciences*, 324:270–285, 2015.
- Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
- Dong, X. and Yang, Y. Nas-bench-201: Extending the scope of reproducible neural architecture search. *International Conference on Learning Representations (ICLR)*, 2020.
- Eggenberger, K., Hutter, F., Hoos, H., and Leyton-brown, K. Efficient benchmarking of hyperparameter optimizers via surrogates background: hyperparameter optimization. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 1114–1120, 2012.
- Elsken, T., Hutter, F., and Metzen, J. H. Efficient multi-objective neural architecture search via Lamarckian evolution. *7th International Conference on Learning Representations, ICLR 2019*, pp. 1–23, 2019.
- Emmerich, M. T. M., Deutz, A. H., and Klinkenberg, J. W. Hypervolume-based expected improvement: Monotonicity properties and exact computation. In *2011 IEEE Congress of Evolutionary Computation (CEC)*, pp. 2147–2154, 2011.
- Falkner, S., Klein, A., and Hutter, F. Bohb: Robust and efficient hyperparameter optimization at scale. *Proceedings*

- of the *International Conference on Machine Learning (ICML)*, 2018.
- Feurer, M., Springenberg, T., and Hutter, F. Initializing Bayesian hyperparameter optimization via meta-learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- Feurer, M., Letham, B., and Bakshy, E. Scalable meta-learning for Bayesian optimization using ranking-weighted Gaussian process ensembles. In *ICML 2018 AutoML Workshop*, 2018.
- Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J., and Sculley, D. Google Vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1487–1495, 2017.
- Klein, A. and Hutter, F. Tabular benchmarks for joint architecture and hyperparameter optimization. *arXiv preprint arXiv:1905.04970*, 2019.
- Law, H. C. L., Zhao, P., Huang, J., and Sejdinovic, D. Hyperparameter learning via distributional transfer. Technical report, preprint arXiv:1810.06305, 2018.
- Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. Technical report, preprint arXiv:1603.06560, 2016.
- Liu, H., Lafferty, J., and Wasserman, L. The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. 10:2295–2328, 2009.
- Mockus, J., Tiesis, V., and Zilinskas, A. The application of bayesian methods for seeking the extremum. *Towards global optimization*, 2(117-129):2, 1978.
- Perrone, V., Jenatton, R., Seeger, M., and Archambeau, C. Scalable hyperparameter transfer learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- Perrone, V., Shen, H., Seeger, M., Archambeau, C., and Jenatton, R. Learning search spaces for bayesian optimization: Another view of hyperparameter transfer learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- Picheny, V. Multiobjective optimization using gaussian process emulators via stepwise uncertainty reduction. *Statistics and Computing*, 25(6):1265–1280, 2015.
- Poloczek, M., Wang, J., and Frazier, P. I. Warm starting Bayesian optimization. In *Winter Simulation Conference (WSC), 2016*, pp. 770–781. IEEE, 2016.
- Ponweiser, W., Wagner, T., Biermann, D., and Vincze, M. Multiobjective optimization on a limited budget of evaluations using model-assisted s-metric selection. In Rudolph, G., Jansen, T., Beume, N., Lucas, S., and Poloni, C. (eds.), *Parallel Problem Solving from Nature – PPSN X*, pp. 784–794, 2008.
- Rasmussen, C. and Williams, C. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- Real, E., Aggarwal, A., Huang, Y., and Le, Q. V. Regularized evolution for image classifier architecture search. *arXiv preprint arXiv:1802.01548*, 2018a.
- Real, E., Aggarwal, A., Huang, Y., and Le, Q. V. Regularized evolution for image classifier architecture search. *arXiv preprint arXiv:1802.01548*, 2018b.
- Salinas, D., Flunkert, V., and Gasthaus, J. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *CoRR*, abs/1704.04110, 2017.
- Schilling, N., Wistuba, M., and Schmidt-Thieme, L. Scalable hyperparameter optimization with products of Gaussian process experts. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 33–48. Springer, 2016.
- Snelson, E., Ghahramani, Z., and Rasmussen, C. E. Warped gaussian processes. In Thrun, S., Saul, L. K., and Schölkopf, B. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 337–344. 2004.
- Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, M., and Adams, R. Scalable Bayesian optimization using deep neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2171–2180, 2015.
- Springenberg, J. T., Klein, A., Falkner, S., and Hutter, F. Bayesian optimization with robust Bayesian neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4134–4142, 2016.
- Svenson, J. and Santner, T. Multiobjective optimization of expensive-to-evaluate deterministic computer simulator models. *Computational Statistics & Data Analysis*, 94: 250 – 264, 2016.
- Swersky, K., Snoek, J., and Adams, R. P. Multi-task Bayesian optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2004–2012, 2013.
- Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., and Le, Q. V. MnasNet: Platform-Aware Neural Architecture Search for Mobile. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2018.

- Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.
- Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.
- Wilson, A. G. and Ghahramani, Z. Copula processes. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, NeurIPS 2010, pp. 2460–2468, 2010.
- Wistuba, M., Schilling, N., and Schmidt-Thieme, L. Hyperparameter search space pruning—a new component for sequential model-based hyperparameter optimization. In *Machine Learning and Knowledge Discovery in Databases*, pp. 104–119. Springer, 2015a.
- Wistuba, M., Schilling, N., and Schmidt-Thieme, L. Learning hyperparameter optimization initializations. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pp. 1–10. IEEE, 2015b.
- Wistuba, M., Schilling, N., and Schmidt-Thieme, L. Scalable Gaussian process-based transfer surrogates for hyperparameter optimization. *Machine Learning*, 107(1): 43–78, 2018.
- Ying, C., Klein, A., Real, E., Christiansen, E., Murphy, K., and Hutter, F. Nas-bench-101: Towards reproducible neural architecture search. *Preprint arXiv:1902.09635*, 2019.
- Yogatama, D. and Mann, G. Efficient transfer learning method for automatic hyperparameter tuning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1077–1085, 2014.
- Zoph, B. and Le, Q. V. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1802.01548*, 2017.