
GPU Accelerated Exhaustive Search for Optimal Ensemble of Black-Box Optimization Algorithms

Jiwei Liu
Nvidia
RAPIDS
Pittsburgh, PA, USA
jiweil@nvidia.com

Bojan Tunguz
Nvidia
RAPIDS
Greencastle, IN, USA
btunguz@nvidia.com

Gilberto Titericz
Nvidia
RAPIDS
Curitiba, Brazil
gtitericz@nvidia.com

Abstract

Black-box optimization is essential for tuning complex machine learning algorithms which are easier to experiment with than to understand. In this paper, we show that a simple ensemble of black-box optimization algorithms can outperform any single one of them. However, searching for such an optimal ensemble requires a large number of experiments. We propose a Multi-GPU-optimized framework to accelerate a brute force search for the optimal ensemble of black-box optimization algorithms by running many experiments in parallel. The lightweight optimizations are performed by CPU while expensive model training and evaluations are assigned to GPUs. The multi-GPU solution achieves 10x speedup of the CPU implementation. With the optimal ensemble found by GPU-accelerated exhaustive search, we won the 2nd place of NeurIPS 2020 black-box optimization challenge¹.

1 Introduction

Black-box optimization (BBO) has become the state-of-the-art methodology for parameter tuning of complex machine learning models [1, 2]. The optimization process is considered black-box because the details of the underlying machine learning models, datasets and the objective functions are hidden from the optimizer. The optimizer finds best hyper-parameters by experimenting with the machine learning models and observing the performance scores [3]. Popular BBO algorithms include random search [4] and Bayesian optimization [5]. Random search is proven to be more effective than brute-force grid search [4]. Bayesian optimization (BO), on the other hand, utilizes probabilistic models to find better hyper-parameters and it outperforms random search consistently [5]. Many advanced BO algorithms have been proposed to improve its scalability [6, 2]. Popular BBO libraries such as Optuna [7] and Hyperopt [8] adopt the Tree of parzen estimators (TPE) [9] algorithm and efficient pruning techniques.

As advances made in BBO improve accuracy, efficiency and usability, they also become increasingly complicated and opaque to users, just like another black box. The NeurIPS BBO challenge [10] provides a great opportunity to study and improve them. We made the following observations from what we learned:

- BBO algorithms excel in different machine learning models, datasets and objective functions.
- The overall execution time is dominant by model evaluation, which could cost 100x time of the actual optimization.

¹Source code: <https://github.com/daxiongshu/rapids-ai-BBO-2nd-place-solution>

Table 1: Bayesmark Overview

| | |
|------------------|---|
| Optimizers | hyperopt [8], nevergrad [14], opentuner [15], pysot [16], skopt [17], turbo [2] |
| Models [13] | DT, MLP-adam, MLP-sgd, RF, SVM, ada, kNN, lasso, linear |
| Dataset [13, 18] | breast, digits, iris, wine, boston, diabetes |
| Metrics [13, 18] | nll, acc, mse, mae |

These insights inspire us to treat BBO algorithms as black-boxes and search for an ensemble of BBO algorithms that outperforms the best single BBO algorithm. A simple exhaustive search is proven to be effective and it is enabled by accelerating massive parallel model evaluations on GPUs. Specifically we made the following contributions:

- An ensemble algorithm that allows multiple BBO algorithms to collectively make suggestions and share observations, within the same time budget as a single BBO algorithm.
- A multi-GPU optimized exhaustive search framework to find BBO candidates for the optimal ensemble.
- A suite of GPU-optimized cuML [11] models including scikit-learn counterparts, MLPs and Xgboost [12] are added to the Bayesmark toolkit to accelerate single model evaluation.
- A comprehensive evaluation and empirical analysis of both individual and ensemble optimization algorithms.

2 Overview of Bayesmark and Provided Optimizers

Bayesmark [3] is the framework for the BBO challenge, which has scikit-learn [13] (sklearn) models and datasets built-in to evaluate BBO algorithms. The Bayesmark provides 6 optimization algorithms, 9 machine learning models and 6 datasets, which are summarized in table 1.

Machine learning models are from scikit-learn toolkit [13] and each model has two variants for classification and regression, respectively. In the competition, each function is optimized in $N_STEP = 16$ iterations with batch size of $N_BATCH = 8$ per iteration. The optimizer for Bayesmark implements an suggest-observe interface as shown in Algorithm 1 (Left). Each iteration the optimizer suggests N_BATCH new guesses for evaluating the function. Once the evaluation is done, the scores are passed back as observations to update the optimizers. The function to be optimized is simply the cross validation score of a machine learning model on a dataset with a loss function.

3 Motivation Study

We study the performance of given optimizers with $N_STEP = 16$, $N_BATCH = 8$ and repeat 3 times ($N_REPEAT = 3$). Experiments are run using the default script [19] for all sklearn models, datasets and metrics.

Table 2 shows the optimization timing breakdown averaged per iteration and function for each optimizer. The time budget for optimization, which is the limit for the total time of “suggest” and “observe”, cannot exceed 40 seconds at each iteration. It is apparent that the time budget is more than enough to run multiple optimizations, suggesting an opportunity for an ensemble of optimizers.

Table 3 summarizes the normalized mean score [20] of each optimizer over all sklearn models (lower the better). The score is normalized to $(-1, 1)$ w.r.t the random search where 1 means random search level performance and 0 means the optimum found by random search [20]. We make two observations:

- Overall *turbo* is the best optimizer and *nevergrad* is the worst optimizer, in terms of number of models with which the optimizer achieves the lowest loss.
- Each optimizer is good at different models, which also presents a chance of an ensemble of optimizers.

Table 2: Average optimization time breakdown per iteration: seconds. The budge is 40 seconds per iteration.

| | hyperopt | nevergrad | opentuner | pysot | skopt | turbo |
|---------|----------|-----------|-----------|-------|-------|-------|
| suggest | 0.008 | 0.009 | 0.038 | 0.026 | 0.345 | 0.385 |
| observe | 0.00006 | 0.009 | 0.005 | 0.001 | 0.185 | 0.004 |

Table 3: Average minimum loss of each optimizer over all scikit-learn models. The lowest loss from each model (per row) is highlighted.

| | hyperopt | nevergrad | opentuner | pysot | skopt | turbo |
|----------|--------------|-----------|--------------|---------------|---------------|---------------|
| DT | 0.049 | 0.161 | 0.11 | -0.01 | -0.015 | -0.007 |
| MLP-adam | 0.011 | 0.023 | 0.039 | 0.016 | 0.037 | -0.007 |
| MLP-sgd | 0.035 | 0.059 | 0.084 | 0.016 | 0.026 | -0.008 |
| RF | 0.006 | 0.033 | 0.081 | -0.02 | -0.024 | 0.007 |
| SVM | 0.035 | 0.074 | 0.039 | 0.025 | 0.035 | 0.012 |
| ada | 0.132 | 0.125 | 0.069 | 0.101 | 0.104 | 0.099 |
| kNN | 0.022 | 0.065 | 0.0 | 0.047 | 0.056 | 0.045 |
| lasso | 0.027 | 0.068 | 0.03 | 0.041 | 0.062 | 0.034 |
| linear | 0.012 | 0.06 | 0.065 | -0.001 | 0.022 | 0.013 |

4 A Heuristic Ensemble Algorithm for Optimizers.

We propose a heuristic algorithm for an ensemble of two optimizers, which can be easily generalized to multiple optimizers. In Algorithm 1 (Right), two optimizers opt_1 and opt_2 are initialized. At each iteration, the optimizer is supposed to suggest N_BATCH guesses. Instead, opt_1 and opt_2 each contribute half of the N_BATCH guesses. A key design choice is that when the evaluation scores return, they are passed to both optimizers so that the two optimizers can learn from each other’s suggestions.

```

opt = Some_Optimizer()
for iter_id = 1 to N_STEP do
    params_list = opt.suggest(N_BATCH)
    scores = evaluate(params_list)
    opt.observe(scores)
end

opt_1 = Some_Optimizer()
opt_2 = Another_Optimizer()
for iter_id = 1 to N_STEP do
    params_list_1 = opt1.suggest(N_BATCH/2)
    params_list_2 = opt2.suggest(N_BATCH/2)
    # concatenate two lists
    params_list = params_list_1 + params_list_2
    scores = evaluate(params_list)
    opt_1.observe(scores)
    opt_2.observe(scores)
end

```

Algorithm 1: Left: workflow of a single optimizer. Right: A heuristic ensemble algorithm for two optimizers.

5 GPU Accelerated Exhaustive Search

A key question is which two optimizers to choose for the ensemble. Based on the motivation study in Table 3, we believe that an exhaustive search of all possible combinations is the most reliable method. However, using the default models and datasets of Bayesmark has the following downsides:

- The data size is small. The number of samples of built-in toy dataset ranges from 150 to 1797. Small dataset introduces randomness and it is prone to overfitting.
- Evaluating scikit-learn models on the CPU is slow. For example, the total running time to obtain the results of Table 3 is 42 hours using the default scripts. It should be noted that this

experiment is performed on the small sklearn toy data. Running additional data will be even more time consuming.

To make the experiments more representative and robust, We add three new real-world datasets: California housing, hotel booking and Higgs Boson. Each dataset is down-sampled to 10000 samples to make experiments faster.

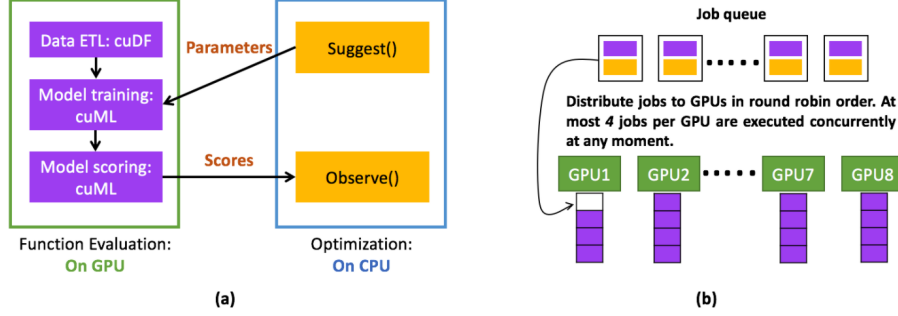


Figure 1: GPU acceleration of BBO. (a) GPUs are used to execute computing intensive function evaluations with cuDF and cuML libraries. (b) Parallel execution of function evaluation and optimization on multiple GPUs.

To accelerate the experiments, we implement the entire evaluation pipeline on GPU utilizing the RAPIDS GPU data science framework. Data loading and preprocessing are boosted by cuDF while scikit-learn models and scorers are replaced with their GPU counterparts in cuML library [11]. Xgboost [12] is also added to the benchmark suite, which supports both GPU and CPU modes. We also implemented a GPU-optimized MLP using pytorch [21] where data loader is implemented with cupy [22] and DLPACK [23] to keep data on GPU.

Another benefit of moving function evaluation onto GPUs is workload partitioning as shown in Figure 1(a). The more computing intensive workloads are on GPU while the relatively lightweight optimization is on the CPU. This also facilitates parallel experiments when multiple GPUs are available. The workloads are distributed to each GPU in round robin order until each GPU has N jobs to execute concurrently. For Bayesmark with cuML workloads, we experiment with several values for N and set $N = 4$ for optimal GPU utilization and memory consumption.

6 Experiments and Results

6.1 Hardware Configuration

The experiments are performed on a DGX-1 workstation with 8 NVIDIA V100 GPUs and two 20-core Intel Xeon CPUs. An exhaustive search is implemented for all $\binom{M}{2}$ combinations of M provided optimizers. In this case, $M = 5$ so there are 10 ensembles and 5 single optimizers to experiment. We exclude nevergrad because it has the worst performance and crashes randomly in ensemble for GPU estimators.

6.2 Evaluation of the Ensemble of Optimizers

Figure 2 summarizes the performance of each optimization algorithm. In Bayesmark, the data is split into training data and hold-out test data. The cross validation score of the training data is visible to optimizers and it is the score optimized. The validation score of the hold-out test data represents the generalization capability of the optimizers to new data. We argue that the generalization score is more important because the hidden models and dataset in the competition must be different. For example, *turbo* has some degree of overfitting since it is the best optimizer for cross validation but it is not in top 5 in terms of generalization. Based on these results, we believe the best three optimizers overall are *turbo - hyperopt*, *pysot* and *turbo - skopt*.

Figure 3 shows the iterative performance of each optimization algorithm. Since the cross validation score is visible to the optimizers and the cumulative minimum is used, the curve always goes down.

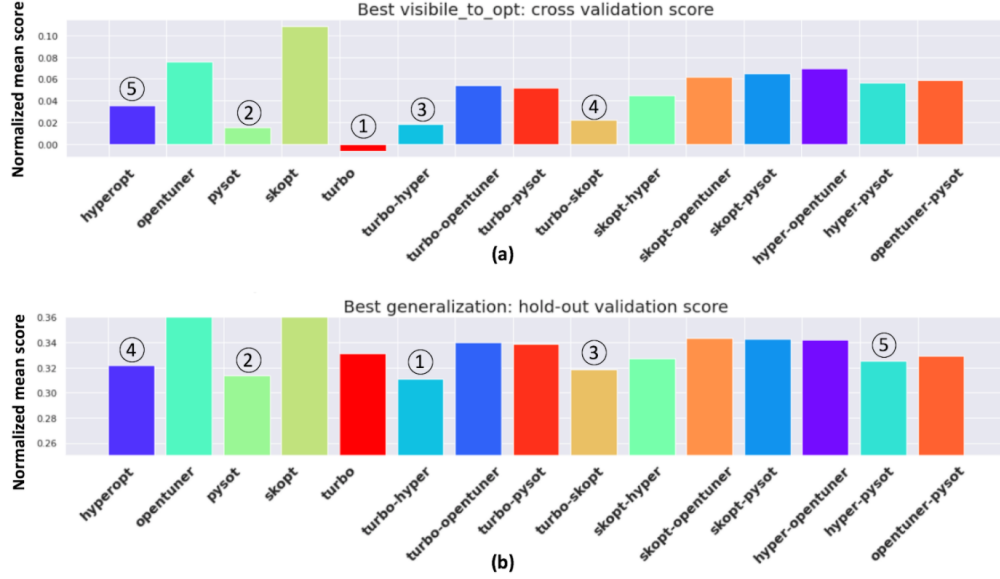


Figure 2: Performance of optimization algorithms in terms of (a) cross validation score that is visible to and minimized by optimizers and (b) holdout validation score which represents the generalization ability of the optimizer. The y-axis is normalized mean score [20] and lower is better. The top 5 optimizers are highlighted in each sub-figure

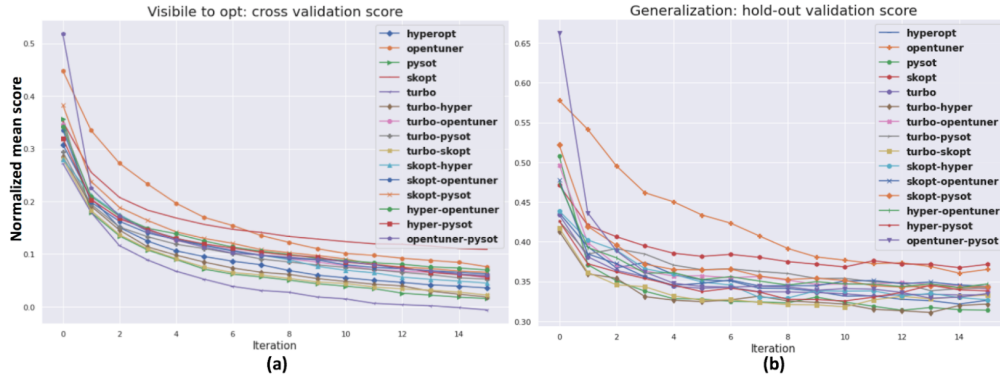


Figure 3: Performance of optimization algorithms per iteration: (a) cross validation score that is visible to optimizers. (b) holdout validation score which represents the generalization ability.

It is clear that optimizer *turbo* (purple) outperforms every other optimizer by a large margin as shown in Figure 3(a). However, Figure 3(b) shows a different pattern. The *turbo – hyper* (brown diamond) and *turbo – skopt* (yellow square) converge faster than other optimizers including the best single optimizer *pysot* (green). We believe that it is due to the diversity of ensemble compared to a single optimizer.

Figure 4 shows the performance breakdown of each optimizer on each of the cuML models, in terms of the generalization score. The optimizer *pysot* has the best performance for two tree based models: random forest and xgboost. The ensemble optimizer *turbo – skopt* shines at *MLP – adam*, the most widely used deep learning model in the benchmark suite. It is also interesting to note that the best ensemble optimizer *turbo – hyper* does not achieve the best performance for any model particularly.

Based on these results, we believe the best three optimizers overall are *turbo – hyperopt*, *pysot* and *turbo – skopt*, which can be good submissions to the competition. **We chose *turbo – skopt* as our final submission because 1) it has a Top-3 generalization score; 2) it converges faster than single optimizers and 3) it achieves best performance for a representative deep learning model.**

| | hyperopt | opentuner | pysot | skopt | turbo | turbohyper | turboopen | turbopysot | turbosk | skhyper | skopen | skpysot | hyperopen | hyperpysot | pysotopen |
|---------------|----------|-----------|-------|-------|-------|------------|-----------|------------|---------|---------|--------|---------|-----------|------------|-----------|
| model | | | | | | | | | | | | | | | |
| MLP-adam-cuml | 0.06 | 0.18 | 0.04 | 0.05 | 0.05 | 0.04 | 0.07 | 0.09 | 0.03 | 0.05 | 0.08 | 0.07 | 0.09 | 0.06 | 0.06 |
| MLP-sgd-cuml | 0.10 | 0.18 | 0.07 | 0.09 | 0.04 | 0.07 | 0.11 | 0.10 | 0.07 | 0.12 | 0.13 | 0.11 | 0.10 | 0.11 | 0.13 |
| RF-cuml | 0.25 | 0.31 | 0.18 | 1.00 | 0.27 | 0.29 | 0.31 | 0.36 | 0.37 | 0.27 | 0.34 | 0.32 | 0.37 | 0.26 | 0.33 |
| SVM-cuml | 0.41 | 0.41 | 0.36 | 0.45 | 0.41 | 0.42 | 0.38 | 0.28 | 0.45 | 0.38 | 0.36 | 0.44 | 0.34 | 0.40 | 0.50 |
| kNN-cuml | 0.52 | 0.53 | 0.52 | 0.62 | 0.55 | 0.52 | 0.47 | 0.55 | 0.56 | 0.53 | 0.49 | 0.62 | 0.52 | 0.47 | 0.50 |
| lasso-cuml | 0.37 | 0.33 | 0.37 | 0.31 | 0.41 | 0.32 | 0.34 | 0.34 | 0.31 | 0.35 | 0.33 | 0.40 | 0.35 | 0.35 | 0.41 |
| linear-cuml | 0.43 | 0.41 | 0.52 | 0.41 | 0.45 | 0.43 | 0.47 | 0.44 | 0.48 | 0.46 | 0.43 | 0.41 | 0.36 | 0.41 | 0.42 |
| xgb-cuml | 0.27 | 0.32 | 0.26 | 0.34 | 0.27 | 0.27 | 0.34 | 0.29 | 0.29 | 0.30 | 0.34 | 0.28 | 0.29 | 0.32 | 0.31 |

Figure 4: Generalization performance of optimizers on each cuML model. The best optimizer for each model (per row) is highlighted.

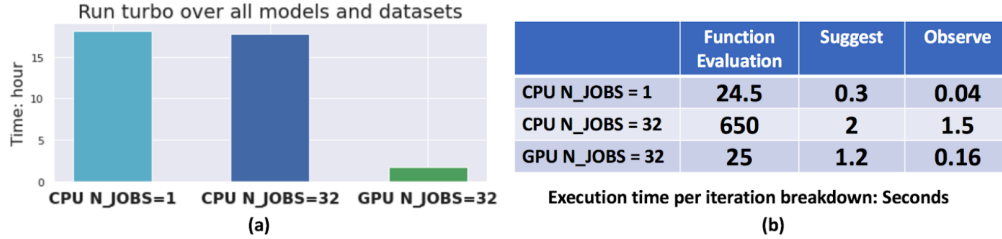


Figure 5: (a) Running time comparison between the proposed multi-GPU implementation and multi-core CPU implementation. (b) The breakdown of execution time per iteration.

Using this simple ensemble, we won the 2nd place of NeurIPS 2020 black-box optimization challenge.

6.3 GPU Speedup

The proposed multi-GPU implementation is the key to finish the exhaustive search in reasonable time. We define a job as one invocation of “bayesmark.experiment()” or one “bayesmark-launch” command with one model, one dataset, one metric and one optimizer. As shown in Figure 5(a), it takes 1.8 hours on GPUs to examine the optimizer turbo over all models, datasets and metrics, where 32 jobs run concurrently on GPUs. In contrast, the multi-core CPUs can’t take advantage of parallelism as shown in Figure 5(a) due to two reasons:

- Some models utilize multi-thread training by default such as xgboost and MLP. The multiple cores of CPUs are already busy even when running with one job at a time. Figure 5(b) shows that the function evaluation time of CPU with $N_JOBS = 32$ almost scales linearly with respect to CPU with $N_JOBS = 1$, indicating there are no speedup.
- The optimizers also run on CPUs so both models and optimizers are competing for CPU resources and slow down the overall performance.

In contrast, the multi-GPU implementation naturally isolates the workloads: models are on GPUs while optimizers are on CPUs. Each GPU is also isolated without any contention with other GPUs. Overall the proposed multi-GPU implementation achieves 10x speedup of the CPU counterpart and all the experiments (Figure 2, 3 and 4) **finish in 24 hours which consists of 4,230 jobs, 2.7 million model trainings and 541,440 optimizations**. The same workload would take at least 10 days on CPUs.

Figure 6 shows the detailed run time comparison of cuML models and sklearn models. Since the dataset in this experimentation is small, cuML models such as *knn* and *xgb* could be slower than their sklearn counterparts on CPU when only one job is training. When launching many jobs, cuML achieves significant speedup on all of the models.

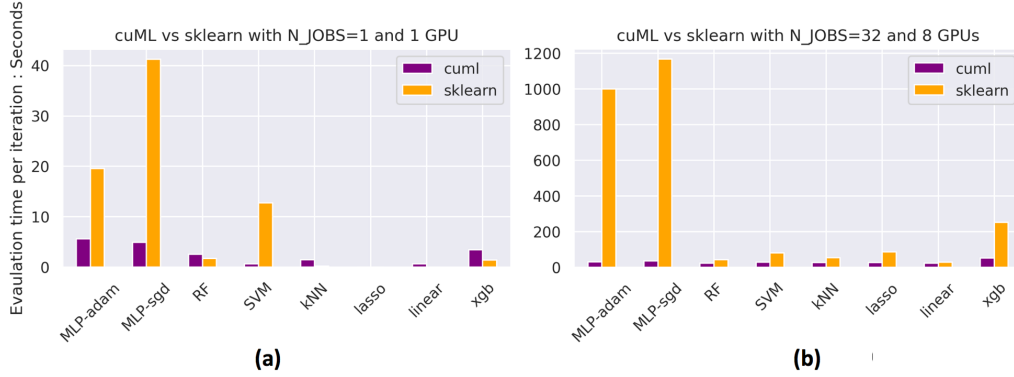


Figure 6: Run time comparison of cuML vs sklearn models.

7 Conclusion

In this paper, we propose a fast multi-GPU accelerated exhaustive search to find the best ensemble of optimization algorithms. The ensemble algorithm can be generalized to multiple optimizers and the proposed framework also scales with multiple GPUs.

Bibliography

- [1] Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John 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*, pages 1487–1495, 2017.
- [2] David Eriksson, Michael Pearce, Jacob Gardner, Ryan D Turner, and Matthias Poloczek. Scalable global optimization via local bayesian optimization. In *Advances in Neural Information Processing Systems*, pages 5496–5507, 2019.
- [3] Uber. uber/bayesmark. <https://github.com/uber/bayesmark>.
- [4] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.
- [5] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.
- [6] José Miguel Hernández-Lobato, James Requeima, Edward O Pyzer-Knapp, and Alán Aspuru-Guzik. Parallel and distributed thompson sampling for large-scale accelerated exploration of chemical space. *arXiv preprint arXiv:1706.01825*, 2017.
- [7] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.
- [8] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR, 2013.
- [9] James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyperparameter optimization. In *Advances in neural information processing systems*, pages 2546–2554, 2011.
- [10] Find the best black-box optimizer for machine learning. <https://bbochallenge.com/>.

- [11] Sebastian Raschka, Joshua Patterson, and Corey Nolet. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *arXiv preprint arXiv:2002.04803*, 2020.
- [12] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.
- [13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [14] J. Rapin and O. Teytaud. Nevergrad - A gradient-free optimization platform. <https://GitHub.com/FacebookResearch/Nevergrad>, 2018.
- [15] Jason Ansel, Shoaib Kamil, Kalyan Veeramachaneni, Jonathan Ragan-Kelley, Jeffrey Bosboom, Una-May O'Reilly, and Saman Amarasinghe. Opentuner: An extensible framework for program autotuning. In *International Conference on Parallel Architectures and Compilation Techniques*, Edmonton, Canada, Aug 2014.
- [16] David Eriksson, David Bindel, and Christine A Shoemaker. pysot and poap: An event-driven asynchronous framework for surrogate optimization. *arXiv preprint arXiv:1908.00420*, 2019.
- [17] Scikit-optimize: Sequential model-based optimization in python. <https://scikit-optimize.github.io/stable/>.
- [18] bayesmark data and metrics. <https://github.com/uber/bayesmark#launch-the-experiments>.
- [19] bbo challenge starter kit. https://github.com/rdturnermtl/bbo_challenge_starter_kit/blob/master/run_local.sh.
- [20] uber/bayesmark. How scoring works. <https://bayesmark.readthedocs.io/en/latest/scoring.html#mean-scores>.
- [21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d 'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [22] ROYUD Nishino and Shohei Hido Crissman Loomis. Cupy: A numpy-compatible library for nvidia gpu calculations.
- [23] Dlpack: Open in memory tensor structure. <https://github.com/dmlc/dlpack>.