Supplementary Material: A simple transfer-learning extension of Hyperband

Lazar Valkov*, **Rodolphe Jenatton***, **Fela Winkelmolen***, **Cédric Archambeau***
Amazon AWS AI* & Amazon research*
l.valkov@sms.ed.ac.uk {jenatton, winkelmo, cedrica}@amazon.com

Abstract

We provide in this supplementary material details about about the experimental settings.

1 XGBoost binary classification tasks

In the second experiment of the paper, we focus on the tuning of XGBoost binary classifiers to optimize the validation AUC. We tune 8 HPs of XGBoost (reusing the terminology from its API):

- `eta` in $[0, 1]$
- `subsample` in $[0.5, 1]$
- `colsample_bytree` in $[0.3, 1]$
- `gamma` in $[2^{-20}, 64]$
- `min_child_weight` in $[2^{-8}, 64]$
- `alpha` in $[2^{-20}, 256]$
- `lambda` in $[2^{-10}, 256]$
- `max_depth` in $[2, 128]$

noting that all other HPs are left to their default values, in particular the booster is equal to `gbtree`. As discussed in the main paper, `num_round` acts as the resource parameter of Hyperband, in the range $[1, 81]$.

We consider the following subset of $T = 25$ datasets from the `libsvm` repository [35]: {`australian, fourclass, german.numer, gina_agnostic, madelon, splice, breast-cancer, higgs_small, a6a, a7a, a8a, ijcnn1, mushrooms, phishing, rcv1.binary, skin_nonskin, spambase, susy, svmguide1, w6a, w7a, w8a, cod-rna, a1a, w1a`}.

2 Using prior information in the definition of the search space

In the two experimental settings we have considered in the paper—tuning of SGD learning rates and XGBoost binary classifiers, we have not made assumptions on the search spaces. In particular, we have not used any prior information in the form of warping transformations, e.g., logarithmic transformations of some HP ranges spanning several order of magnitudes.

To assess the effect of injecting prior knowledge in the form of warping transformations, we run `random` for XGBoost in an appropriately transformed search space (with `log2_gamma`, `log2_min_child_weight`, `log2_lambda`, `log2_alpha` and a discretized list of `max_depth`={2, 3, 4, 6, 8, 11, 16, 23, 32, 45, 64, 91, 128} encoded as an integer HP corresponding to the indices in this list). We refer to this method as `random-log`.

2nd Workshop on Meta-Learning at NeurIPS 2018, Montréal, Canada.

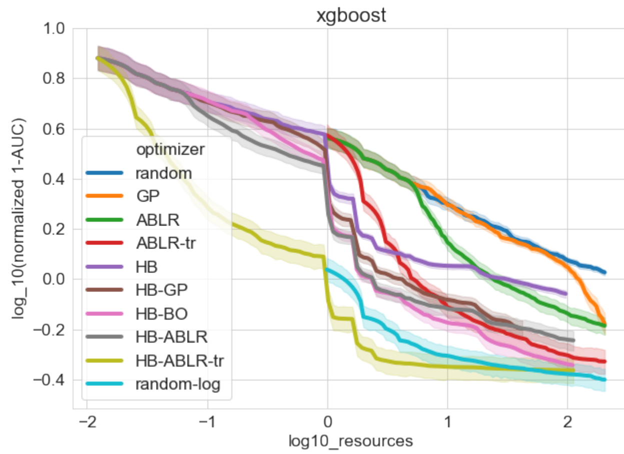

Figure 1: Comparison of different competing methods for the tuning of XGBoost binary classifiers.

In Figure 1, we can observe that appropriate transformations in the search space, if available, lead to a clear improvement. `HB_ABLR_transfer` is the only method operating in the "raw" search space that can first outperform, and then match, `random-log`.

As future work, we plan to further evaluate all the other methods in the transformed search space to better understand the impact of this prior information on the performance.

References

[1] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. Technical report, preprint arXiv:1603.06560, 2016.

[2] L. Bottou and Y. LeCun. Large scale online learning. In *Advances in Neural Information Processing Systems*, volume 16, pages 217–224, 2004.

[3] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1), 2009.

[4] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

[5] Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Technical report, preprint arXiv:1012.2599, 2010.

[6] 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, 2016.

[7] Kevin Swersky, Jasper Snoek, and Ryan P Adams. Multi-task Bayesian optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2004–2012, 2013.

[8] Kevin Swersky, Jasper Snoek, and Ryan Prescott Adams. Freeze-thaw Bayesian optimization. Technical report, preprint arXiv:1406.3896, 2014.

[9] Aaron Klein, Stefan Falkner, Simon Bartels, Philipp Hennig, and Frank Hutter. Fast Bayesian optimization of machine learning hyperparameters on large datasets. Technical report, preprint arXiv:1605.07079, 2016.

[10] Stefan Falkner, Aaron Klein, and Frank Hutter. Bohb: Robust and efficient hyperparameter optimization at scale. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1436–1445, 2018.

[11] Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. Technical report, preprint arXiv:1502.07943, 2015.

[12] Stefan Falkner, Aaron Klein, and Frank Hutter. Combining hyperband and bayesian optimization. In *Proceedings of BayesOpt NIPS workshop*, 2017.

[13] Hadrien Bertrand, Roberto Ardon, Matthieu Perrot, and Isabelle Bloch. Hyperparameter optimization of deep neural networks: Combining hyperband with bayesian model selection. In *Conférence sur l'Apprentissage Automatique (CAP 2017)*, 2017.

[14] Jiazhuo Wang, Jason Xu, and Xuejun Wang. Combination of hyperband and bayesian optimization for hyperparameter optimization in deep learning. Technical report, preprint arXiv:1801.01596, 2018.

[15] Rémi Bardenet, Mátyás Brendel, Balázs Kégl, and Michele Sebag. Collaborative hyperparameter tuning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 199–207, 2013.

[16] Dani Yogatama and Gideon Mann. Efficient transfer learning method for automatic hyperparameter tuning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1077–1085, 2014.

[17] Martin Wistuba, Nicolas Schilling, and Lars Schmidt-Thieme. Learning hyperparameter optimization initializations. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–10. IEEE, 2015.

[18] Matthias Feurer, T Springenberg, and Frank Hutter. Initializing Bayesian hyperparameter optimization via meta-learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[19] Nicolo Fusi and Huseyn Melih Elibol. Probabilistic matrix factorization for automated machine learning. Technical report, preprint arXiv:1705.05355, 2017.

[20] Valerio Perrone, Rodolphe Jenatton, Matthias Seeger, and Cédric Archambeau. Scalable hyperparameter transfer learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[21] Matthias Feurer, Benjamin Letham, and Eytan Bakshy. Scalable meta-learning for bayesian optimization using ranking-weighted gaussian process ensembles. In *ICML 2018 AutoML Workshop*, July 2018.

[22] Carl Rasmussen and Chris Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[23] Jost Tobias Springenberg, Aaron Klein, Stefan Falkner, and Frank Hutter. Bayesian optimization with robust Bayesian neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4134–4142, 2016.

[24] James Bergstra, Rémi Bardenet, Yoshua Bengio, Balázs Kégl, et al. Algorithms for hyperparameter optimization. In *Advances in Neural Information Processing Systems*, volume 24, pages 2546–2554, 2011.

[25] 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*, pages 785–794. ACM, 2016.

[26] Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. The application of Bayesian methods for seeking the extremum. *Towards Global Optimization*, 2(117-129):2, 1978.

[27] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

[28] F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *Proceedings of LION-5*, page 507?523, 2011.

[29] Mitchell McIntire, Daniel Ratner, and Stefano Ermon. Sparse gaussian processes for Bayesian optimization. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2016.

[30] Geoff Pleiss, Jacob R Gardner, Kilian Q Weinberger, and Andrew Gordon Wilson. Constant-time predictive distributions for gaussian processes. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[31] Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. Scalable Bayesian optimization using deep neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2171–2180, 2015.

[32] `GPyOpt`: A Bayesian optimization framework in python. `http://github.com/SheffieldML/GPyOpt`, 2016.

[33] 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.

[34] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

[35] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.