Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

MULTI-FIDELITY BAYESIAN OPTIMIZATION WITH TRACE OBSERVATIONS

Saul Toscano-Palmerin

The School of Operations Research and Information Engineering
Cornell University
206 Frank HT Rhodes Hall
Ithaca, NY 14853, USA

ABSTRACT

We propose a new provably-convergent continuous-fidelity Bayesian optimization method where fidelity is controlled by one or more continuous settings, and we have trace observations. For example, we may wish to adjust training data size and number of training iterations, for an optimal accuracy run-time tradeoff. We make two innovations: (1) we fix the widely spread issue that common continuous-fidelity methods always prefer a very low fidelity point no matter how much actual value it can provide, and our fix is theoretically sound; (2) our method is designed in a decision-theoretic manner in light of the fact that we have trace observations. Numerical experiments show that our method outperforms state-of-art algorithms when optimizing synthetic functions, tuning feedforward neural networks on MNIST, tuning convolutional neural networks (CNNs) on CIFAR-10 and SVHN, and in large-scale kernel learning.

1 INTRODUCTION

We consider the problem of finding an optimal solution x in some set $\mathbb{A}$ to minimize an objective function $f(x)$, i.e., to solve

$$\min_{x \in \mathbb{A}} f(x) \tag{1}$$

Evaluating $f(x)$ can take substantial time and computational power (Bergstra and Bengio 2012), and may not provide gradient evaluations. This problem is commonly found in optimization via simulation with steady state simulations (Goldsman et al. 2002), and machine learning (Snoek et al. 2012). In hyperparameter tuning of machine learning models, we seek to find a set of hyperparameters x in some set $\mathbb{A}$ to minimize the validation error $f(x)$. Thus, machine learning practitioners have turned to Bayesian optimization for solving (1) (Snoek et al. 2012) because it tends to find good solutions with few function evaluations (Jones et al. 1998).

As the computational expense of training and testing a modern deep neural network for a single set of hyperparameters has grown, researchers have sought ways to solve (1) more quickly by supplanting some evaluations of $f(x)$ with computationally inexpensive low-fidelity approximations. These approximations perform the same training and testing steps, but use fewer training iterations than required for convergence or ignore some training data. Recently developed solution approaches of this type include the Bayesian optimization methods FaBOLAS (Klein et al. 2016; Klein et al. 2015), Freeze-Thaw Bayesian Optimization (Swersky et al. 2014), BOCA (Kandasamy et al. 2017), a predictive entropy search method for a single continuous fidelity in McLeod et al. (2017), and early-stopping SMAC (Domhan et al. 2015). They also include the bandit method Hyperband (Li et al. 2016). Previous literature demonstrates these approaches can find good hyperparameter settings significantly faster than methods that always train until convergence using the full training data.

While many of these approaches build on previous work on multi-fidelity Bayesian optimization, training iterations have unique characteristics that differ from fidelity controls typically considered in that

literature. First, training with n iterations naturally produces evaluations of the low-fidelity approximation for *all* training iterations less than or equal to n. While computing the test error after each iteration does require an extra step of evaluating in the test data, this expense is typically very small compared to training. Second, if we cache the state of our training algorithm once an evaluation with n iterations is complete, we can restart evaluation from this state if we wish to evaluate for $n' > n$ evaluations, significantly reducing the expense of evaluation. Succinctly, we observe the full *trace* of performance with respect to training iterations, not just a single point.

The fact that we observe traces over training iterations raises two key challenges: (1) Balancing between increasing training iterations at previously evaluated hyperparameters vs. starting from scratch at a new set of hyperparameters that seem promising based on other results; (2) Extracting the most useful information from a set of training iterations without burdening a Gaussian process inference procedure (which scales as the cube of the number of observations) with a large number of observations.

Contributions: We propose a novel acquisition function, the trace-aware knowledge gradient, and a provably-convergent method for maximizing it, that can choose the training data size, training iterations, and hyperparameters at which to evaluate. Our approach is applicable to problems with trace observations along one or more fidelity controls, including hyperparameter optimization while varying both training iterations and training data. It addresses the challenges presented by trace observations by considering the reduced cost of adding iterations at a previously evaluated point, and using an intelligent selection scheme to choose a subset of the observed training iterations to include in inference. It can be used in either a batch or sequential setting, and can leverage gradient information if it is available.

Our numerical experiments demonstrate a significant improvement over FaBOLAS, Hyperband, and BOCA.

REFERENCES

Bergstra, J., and Y. Bengio. 2012. "Random Search for Hyper-Parameter Optimization". *Journal of Machine Learning Research* 13(2):281–305.

Domhan, T., J. Springenberg, and F. Hutter. 2015. "Speeding Up Automatic Hyperparameter Optimization of Deep Neural Networks by Extrapolation of Learning Curves". In *IJCAI*, edited by Q. Yang and M. Wooldridge, Volume 15, 3460–3468. Palo Alto, California: AAAI Press.

Goldsman, D., S. H. Kim, W. S. Marshall, and B. L. Nelson. 2002. "Ranking and Selection for Steady-State Simulation: Procedures and Perspectives.". *INFORMS Journal on Computing* 1(14):2–19.

Jones, D. R., M. Schonlau, and W. J. Welch. 1998. "Efficient Global Optimization of Expensive Black-Box Functions". *Journal of Global optimization* 13(4):455–492.

Kandasamy, K., G. Dasarathy, J. Schneider, and B. Poczos. 2017. "Multi-Fidelity Bayesian Optimisation with Continuous Approximations.". *arXiv preprint arXiv:1703.06240.*.

Klein, A., S. Bartels, S. Falkner, P. Hennig, and F. Hutter. 2015. "Towards Efficient Bayesian Optimization for Big Data". In *Advances in Neural Information Processing Systems*, Volume 134, 98. Lake Tahoe, Nevada: Curran Associates, Inc.

Klein, A., S. Falkner, S. Bartels, P. Hennig, and F. Hutter. 2016. "Fast Bayesian Optimization of Machine Learning Hyperparameters on Large Datasets". *arXiv preprint arXiv:1605.07079*.

Li, L., K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. 2016. "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization.". *arXiv preprint arXiv:1603.06560.*.

McLeod, M., M. A. Osborne, and S. J. Roberts. 2017. "Practical Bayesian Optimization for Variable Cost Objectives". *arXiv preprint arXiv:1703.04335*.

Snoek, J., H. Larochelle, and R. P. Adams. 2012. "Practical Bayesian Optimization of Machine Learning Algorithms". In *Advances in neural information processing systems*, 2951–2959. Lake Tahoe, Nevada: Curran Associates, Inc.

Swersky, K., J. Snoek, and R. P. Adams. 2014. "Freeze-Thaw Bayesian optimization". *arXiv preprint arXiv:1406.3896*.