Modern Bayesian Optimization
Research Statement and Proposal

Han Liang Wee Eric
PhD Student, NUS

April 28, 2021

Bayesian Optimization (BO) has been successful (i.e. AlphaGo) in addressing expensive low-dimensional black-box optimization in small continuous search spaces. Modern application of machine learning often requires dealing with high-dimensional, often large datasets. In addition, search spaces for these problems are not always continuous and are sometimes discrete. The research goal is to understand and propose methods to enable the successful application of BO in the digital age. In this paper, recent research experiences and future research direction is discussed. First, BO and its successes and motivation for this research are briefly discussed. Here, we also discuss one of the recent publications dealing with adapting BO to higher dimensions. Finally, we present one future research direction in adapting BO to discrete search spaces: combinatorial Bayesian Optimization.

1 Research Statement

Introduction to Bayesian Optimization. Bayesian Optimization (BO) is a popular method for sequential global optimization, and is suited to situations in which the objective black-box function f is expensive to evaluate [SLA12; Moc94]. For example, we consider finding $x_{\max} = \arg\max_{x \in \mathcal{X}} f(x)$ for some function f. If f is known and known/assumed to be convex, we can simply apply convex optimization. However, f is not known exactly and/or evaluations expensive and/or noisy in many real problems such as hyperparameter optimization. BO can overcome such conditions to find $x_{\max}$ with the *least* possible number of observations by making two key assumptions: surrogate model that will capture our prior beliefs on f and acquisition function that can be inexpensively optimized. BO translates the original problem to yet another optimization problem. Gaussian Processes (GP) are commonly used as the surrogate because it allows us to model the prior and marginal log-likelihood.

Motivation. BO has found success in many areas; most notably in automatically tuning the game playing hyperparameters of AlphaGo [Che+18]. BO enhanced the playing power of AlphaGo while decreasing resources needed for manual tuning. However, such success have been often found in low dimensional [Wan+13] continuous spaces [Sha+16] with small datasets [Amb+15].

Active Research. From the lines of inquiry, we discuss possible research directions below:

- Scaling to higher dimensions: two approaches via making additional assumptions,
  - Low effective dimensionality, only few dimensions significantly affect f [CCK12].
  - Additive structure, small subsets of variables interact with each other [KSP15].
- Adapting to discrete spaces: various approaches to tackle the problem,

  - Replace/adapt surrogate model with random forests or another suitable model [HHL13].
  - Introduce smoothness, attempting to coerce discrete spaces into continous spaces [Oh+19].
  - Feature engineering, through feature expansion of discrete with continous variables [Dax+19].
- Scaling to more observations: two common approaches,
  - Replace surrogate model, similar to above, with another suitable model [HHL11].
  - Introduce efficient matrix operations, such as sparsification of covariance matrix [SWL03].

We are also interested in other adjacent directions, such as: improving the robustness of BO methods.

We recently discussed a method via the additive structure assumption [HAS21], where we facilitate faster model learning by *reducing the model complexity* while retaining the *sample-efficiency* of existing additive methods. Specifically, we constrain the underlying additive structure to trees, so we facilitate scalable learning of the structure learning and optimization of the acquisition function. We show that our method is competitive on datasets, and the computation can be significantly reduced while maintaining optimization performance.

2 Research Proposal

Here, we focus our discussion on adapting BO to discrete spaces. Most BO methods are focused on continuous search spaces due to the underlying assumption of using GPs as the prior, which relies on the smoothness defined by a kernel to model uncertainty. Real-world problems can have a mix of both ordinal and categorical variables, presenting the following issues, that BO is suited to solve:

1. black-box objectives for which gradient-based optimizers are not amenable.
2. expensive evaluation procedures for which methods with low sample efficiency.
3. noisy and highly non-linear objectives for which simple and exact solutions are inaccurate.

There have been attempts to adapt BO to this problem by simply using continuous kernels and augment the result to become discrete. However, some combinations of discrete variables are impossible configurations, and it is difficult for the model to capture.

What is the problem you want to address in your work? Combinatorial Bayesian Optimization, to extend BO to discrete spaces. Our method should also work on mixed spaces, ie. both continuous and discrete spaces.

Why is it a problem? Optimization in structured domains using BO was raised in [Neu17]; BO has sample efficiency and global optimum convergence advantages when compared to other approaches such as evolution or genetic algorithms. An application would be deciding on a deep neural network architecture where the parameters may depend on the number of layers [Ben09]. There have been attempts to tackle this problem simply using continuous kernels in BO and augment the result to become discrete. However, some combinations of discrete variables are impossible configurations that are complex for the model to capture. It faces the problem of combinatorial explosion, as instead of modelling the structural dependencies, BO embed the discrete variables into a $\mathbb{R}^d$ box. These methods work but heavily rely on discretization.

What is the solution you developed in your work? We focus on the important considerations of: allowing information sharing via encoding relationships between the variables (both ordinal and categorical), efficiently optimize the discrete variables (selection of the next discrete values) with high sample and computational efficiency, encode the constraints between the variables, allowing the model to exploit them. We extend COMBO [Oh+19], where GP surrogate model is defined

over the graph using Graph Fourier Transform (GFT). Instead of using an ordered sub-graph for every continuous variable, we do not discretize, considering the continuous variables outside of the graph structure. It might be possible to define a kernel grammar over continuous and discrete variables by treating them separately.

Why is it a solution? COMBO uses hamming distance on both continuous and discrete variables, which works well on discrete variables. Hamming distance is an unnatural metric on continuous spaces, which should be instead modelled using a continuous distribution. There might be some penalty associated with GFT, such as performance issues, which might need to be addressed.

2.1 Relevant Related Work

In this section, we identify closely related works and their key ideas/contributions below:

[Swe+20] Use generative model to generate candidates for acquisition optimization. Implementation at `https://github.com/google-research/google-research/tree/master/amortized_bo`.

[Dax+19] Decouples the continuous and discrete components of the function using feature expansion. No implementation published.

[Oh+19] Use a combinatorial graph to quantify smoothness on combinatorial search spaces. Implementation at `https://github.com/QUVA-Lab/COMBO`.

[BP18] Uses second order regression modelling of the interactions. Implementation at `https://github.com/baptistar/BOCS`.

[Jen+17] Use a surrogate tree model where GPs are at the leafs. No implementation published.

Other related methods:

1. Traditional methods; generally not sample efficient and may not converge to global optimum (although sometimes, convergence is guaranteed).
   a) Local search methods such as random optimization, hill climbing, simulated annealing, etc.
   b) Evolutionary algorithms.
2. Related BO methods; generally unable to handle general constraints over the discrete variables.
   a) SMAC [HHL11] is a BO that uses random forests as a surrogate model, which accommodates mixed variable inputs. However, there are limitations as the frequentist uncertainty estimate provided by random forests may suffer from variance degradation.
   b) TPE [BYC13] uses a non-parametric kernel density estimator to identify inputs, likely to improve upon and unlikely to perform worse than the best input found so far.
3. Hyperband [Li+17] is a variant of random search that exploits cheap but less accurate approximations of the objective to allocate resources for function evaluations dynamically.
4. BOHB [FKH18] is the model-based counterpart of Hyperband, based on TPE.
5. Arc-Kernel [Swe+14] is a kernel designed to encode information about which parameters are relevant in a given structure.

Acknowledgements

We adapted this research template from [Rue18]. This work is done in collaboration with my PhD supervisor, Prof. Scarlett Jonathan. His patience and guidance are greatly appreciated, without which the research would not be possible; to family and friends who are of great support.

References

[Amb+15] Sivaram Ambikasaran et al. "Fast direct methods for Gaussian processes." In: *IEEE transactions on pattern analysis and machine intelligence* 38.2 (2015), pp. 252–265.

[Ben09] Yoshua Bengio. *Learning deep architectures for AI*. Now Publishers Inc, 2009.

[BP18] Ricardo Baptista and Matthias Poloczek. "Bayesian optimization of combinatorial structures." In: *International Conference on Machine Learning*. PMLR. 2018, pp. 462–471.

[BYC13] 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*. PMLR. 2013, pp. 115–123.

[CCK12] Bo Chen, Rui M Castro, and Andreas Krause. "Joint optimization and variable selection of high-dimensional Gaussian processes." In: 2012, pp. 1379–1386.

[Che+18] Yutian Chen et al. "Bayesian optimization in alphago." In: *arXiv preprint arXiv:1812.06855* (2018).

[Dax+19] Erik Daxberger et al. "Mixed-variable bayesian optimization." In: *arXiv preprint arXiv:1907.01329* (2019).

[FKH18] Stefan Falkner, Aaron Klein, and Frank Hutter. "BOHB: Robust and efficient hyperparameter optimization at scale." In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1437–1446.

[HAS21] Eric Han, Ishank Arora, and Jonathan Scarlett. "High-Dimensional Bayesian Optimization via Tree-Structured Additive Models." In: *Proc. of AAAI Conference on Artificial Intelligence* (2021).

[HHL11] Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. "Sequential model-based optimization for general algorithm configuration." In: *International conference on learning and intelligent optimization*. Springer. 2011, pp. 507–523.

[HHL13] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. "An evaluation of sequential model-based optimization for expensive blackbox functions." In: *Proceedings of the 15th annual conference companion on Genetic and evolutionary computation*. 2013, pp. 1209–1216.

[Jen+17] Rodolphe Jenatton et al. "Bayesian Optimization with Tree-structured Dependencies." In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 1655–1664. URL: `http://proceedings.mlr.press/v70/jenatton17a.html`.

[KSP15] Kirthevasan Kandasamy, Jeff Schneider, and Barnabás Póczos. "High dimensional Bayesian optimisation and bandits via additive models." In: 2015, pp. 295–304.

[Li+17] Lisha Li et al. "Hyperband: A novel bandit-based approach to hyperparameter optimization." In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 6765–6816.

[Moc94] Jonas Mockus. "Application of Bayesian approach to numerical methods of global and stochastic optimization." In: *Journal of Global Optimization* 4.4 (1994), pp. 347–365.

[Neu17] NeurIPS. *NeurIPS 2017 Schedule*. `https://nips.cc/Conferences/2017/Schedule?showEvent=8773`. Accessed: 2020-Apr-28. 2017.

[Oh+19] Changyong Oh et al. "Combinatorial bayesian optimization using the graph cartesian product." In: *arXiv preprint arXiv:1902.00448* (2019).

[Rue18] Stefan T. Ruehl. *Latex Template for Research Proposals*. `https://github.com/stefantruehl/research-proposal-template`. 2018.

[Sha+16] Bobak Shahriari et al. "Taking the Human Out of the Loop: A Review of Bayesian Optimization." In: *Proceedings of the IEEE* 104.1 (2016), pp. 148–175. DOI: 10.1109/JPROC.2015.2494218.

[SLA12] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. "Practical Bayesian optimization of machine learning algorithms." In: 2012, pp. 2951–2959.

[Swe+14] Kevin Swersky et al. "Raiders of the lost architecture: Kernels for Bayesian optimization in conditional parameter spaces." In: *arXiv preprint arXiv:1409.4011* (2014).

[Swe+20] Kevin Swersky et al. "Amortized Bayesian Optimization over Discrete Spaces." In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. Ed. by Jonas Peters and David Sontag. Vol. 124. Proceedings of Machine Learning Research. PMLR, Mar. 2020, pp. 769–778. URL: http://proceedings.mlr.press/v124/swersky20a.html.

[SWL03] Matthias Seeger, Christopher Williams, and Neil Lawrence. *Fast forward selection to speed up sparse Gaussian process regression*. Tech. rep. 2003.

[Wan+13] Ziyu Wang et al. "Bayesian optimization in high dimensions via random embeddings." In: 2013.