
Multiple Adaptive Bayesian Linear Regression for Scalable Bayesian Optimization with Warm Start

Valerio Perrone, Rodolphe Jenatton, Matthias Seeger, Cédric Archambeau
Amazon
Berlin, Germany
{vperrone, jenatton, matthis, cedrica}@amazon.com

Abstract

Bayesian optimization (BO) is a model-based approach for gradient-free black-box function optimization. Typically, BO is powered by a Gaussian process (GP), whose algorithmic complexity is cubic in the number of evaluations. Hence, GP-based BO cannot leverage large amounts of past or related function evaluations, for example, to warm start the BO procedure. We develop a multiple adaptive Bayesian linear regression model as a scalable alternative whose complexity is linear in the number of observations. The multiple Bayesian linear regression models are coupled through a shared feedforward neural network, which learns a joint representation and transfers knowledge across machine learning problems.

1 Introduction

Bayesian optimization (BO) is a well-established methodology to optimize expensive black-box functions [20]. It relies on a probabilistic model of an unknown target function $f(\mathbf{x})$, which is repeatedly queried until one runs out of budget (e.g., time). The queries consist in evaluations of f at hyperparameter configurations $\mathbf{x}^1, \dots, \mathbf{x}^n$ selected according to an explore-exploit trade-off criterion (e.g., expected improvement). The hyperparameter configuration corresponding to the best query is then returned. One popular approach is to impose a Gaussian process (GP) prior over f and, in light of the observed queries $f(\mathbf{x}^1), \dots, f(\mathbf{x}^n)$, to compute the posterior GP. The GP model maintains posterior mean and posterior variance functions as required by the explore-exploit criterion.

Despite their flexibility, GPs scale cubically with the number of observations [18]. Hence, they cannot be applied in situations where f has been or can be queried a very large number of times. In this work, we are interested in such a setting as we would like to warm start BO by, e.g., transferring information obtained from previous runs of the BO routine, or learn across similar problems (e.g., a given classifier applied across different datasets [2, 25, 8, 9, 10]), which we will call *tasks*. To tackle the scalability limitation of GPs and ease transfer learning in BO, we propose to fall back to adaptive Bayesian linear regression (BLR) [3], ABLR for short, which scales linearly with the number of observations and cubically in the dimension of the basis function expansion. Sparse GPs [16] or multi-task GPs [23] have been respectively developed to scale up GPs and make them suitable for multi-task learning. ABLR offers a simple alternative combining the strengths of these two approaches.

Our main contribution is to learn a suitable representation of a variety of tasks with a feedforward neural network (NN), provided it is fed with enough data. We consider conditionally independent task-specific BLR models, which share a NN that learns the basis expansion. We compare to random Fourier basis expansions [17] as they have already been successfully applied to BO [11, 12]. While more scalable, they are less flexible in learning a useful representation.

Closest to our work is [21], where BO is scaled up by replacing the GP with an ABLR model. The authors consider a single task setting, with a two-step inference procedure. First, they train the NN with a squared loss at the output layer to learn a maximum a posteriori estimate of the NN parameters. This requires evaluating a number of candidate queries to feed the NN training algorithm. They then fix the network architecture and replace the output layer by a BLR layer to run the BO routine. Instead, we *jointly* learn the basis expansion, that is, the NN, and the task-specific BLR models in one go. Our objective function corresponds to a sum of log-marginal likelihood terms, each term corresponding to one of the underlying tasks. As a result, in contrast with [21] who use the squared loss, we can handle heterogeneous signals, each having its own marginal likelihood. In this sense, we borrow the strength of the likelihood of multi-output GPs while maintaining the scalability of [21].

Another related model is presented in [22]. The authors propose Bayesian NNs to sample from the posterior over f , and add task-specific embeddings to the NN inputs to handle multiple tasks. While allowing for a principled treatment of uncertainties, fully Bayesian NNs are computationally more expensive and their training can be sensitive to the stochastic gradient MCMC hyperparameters. Our model allows for simpler inference and is more suitable for large scale deployment.

2 Multiple Adaptive Bayesian Linear Regression Model

Consider T tasks defined by a set of black-box target functions $\{f_t(\cdot)\}_{t=1}^T$ we would like to optimize. Let $\mathcal{D}_t = \{(\mathbf{x}_t^n, y_t^n)\}_{n=1}^{N_t}$ be the set of N_t pairs of inputs and responses associated to task t . We further denote the stacked response vector associated to task t by $\mathbf{y}_t \in \mathbb{R}^{N_t}$ and the corresponding stacked matrix of inputs by $\mathbf{X}_t \in \mathbb{R}^{N_t \times P}$. We assume the task responses $\{\mathbf{y}_t\}_{t=1}^T$ are drawn from independent BLR models conditioned on the shared feature map $\phi_{\mathbf{z}}(\mathbf{x}) : \mathbb{R}^P \mapsto \mathbb{R}^D$, which is parametrized by \mathbf{z} , and the residual noise parameters $\{\alpha_t\}_{t=1}^T$:

$$\mathbf{y}_t \mid \mathbf{X}_t, \mathbf{w}_t, \alpha_t, \mathbf{z} \sim \mathcal{N}(\Phi_{\mathbf{z}}(\mathbf{X}_t)\mathbf{w}_t, \alpha_t^{-1}\mathbf{I}_{N_t}),$$

where $\Phi_{\mathbf{z}}(\mathbf{X}_t) = [\phi_{\mathbf{z}}(\mathbf{x}_t^n)]_n \in \mathbb{R}^{N_t \times D}$ is the feature matrix, $\mathbf{w}_t \in \mathbb{R}^D$ a weight vector, and $\alpha_t \in \mathbb{R}^+$ a precision (i.e., inverse variance). To complete the model, we impose a zero-mean isotropic Gaussian prior on \mathbf{w}_t and denote its precision by $\beta_t \in \mathbb{R}^+$. In the remainder, we will use Φ_t for $\Phi_{\mathbf{z}}(\mathbf{X}_t)$.

2.1 Posterior inference

The posterior distribution over the weight parameters is analytically tractable in this model, as well as the predictive distribution, both of which are multivariate Gaussian distributions [3]. The predictive mean and the predictive variance at a new input \mathbf{x}_t^* are respectively given by

$$\boldsymbol{\mu}_t(\mathbf{x}_t^*; \mathcal{D}_t, \alpha_t, \beta_t, \mathbf{z}) = \frac{\alpha_t}{\beta_t} \phi_{\mathbf{z}}(\mathbf{x}_t^*)^\top \mathbf{K}_t^{-1} \Phi_t^\top \mathbf{y}_t = \frac{\alpha_t}{\beta_t} \mathbf{c}_t^\top \mathbf{L}_t^{-1} \phi(\mathbf{x}_t^*), \quad (1)$$

$$\sigma_t^2(\mathbf{x}_t^*; \mathcal{D}_t, \alpha_t, \beta_t, \mathbf{z}) = \frac{1}{\beta_t} \phi_{\mathbf{z}}(\mathbf{x}_t^*)^\top \mathbf{K}_t^{-1} \phi_{\mathbf{z}}(\mathbf{x}_t^*) + \frac{1}{\alpha_t} = \frac{1}{\beta_t} \|\mathbf{L}_t^{-1} \phi(\mathbf{x}_t^*)\|^2 + \frac{1}{\alpha_t}, \quad (2)$$

where $\mathbf{K}_t = \frac{\alpha_t}{\beta_t} \Phi_t^\top \Phi_t + \mathbf{I}_D$. The right hand side reformulations (1) and (2) ensure numerical stability. They are obtained by decomposing \mathbf{K}_t in terms of its Cholesky factor $\mathbf{K}_t = \mathbf{L}_t \mathbf{L}_t^\top$, so that $\phi(\mathbf{x}_t^*)^\top \mathbf{K}_t^{-1} \phi(\mathbf{x}_t^*) = \|\mathbf{L}_t^{-1} \phi(\mathbf{x}_t^*)\|^2$ and $\mathbf{K}_t^{-1} \Phi_t^\top \mathbf{y}_t = \mathbf{L}_t^{-\top} \mathbf{c}_t$ with $\mathbf{c}_t = \mathbf{L}_t^{-1} \Phi_t^\top \mathbf{y}_t$.

Each task-specific BLR depends on the hyperparameters α_t and β_t , as well as the set of hyperparameters \mathbf{z} defining the feature map. In particular, \mathbf{z} will represent the weights of a NN (see Section 2.2). We adopt an empirical Bayes approach and jointly learn all these hyperparameters by optimizing the marginal likelihood of the data [15]. More specifically, we integrate out the model parameters $\{\mathbf{w}_t\}_{t=1}^T$ and minimize the sum of the negative log-marginal likelihoods of each task:

$$\rho(\mathbf{z}, \{\alpha_t, \beta_t\}_{t=1}^T) = - \sum_{t=1}^T \left[\frac{N_t}{2} \log \alpha_t - \frac{\alpha_t}{2} \left(\|\mathbf{y}_t\|^2 - \frac{\alpha_t}{\beta_t} \|\mathbf{c}_t\|^2 \right) - \sum_{i=1}^D \log([\mathbf{L}_t]_{ii}) \right]. \quad (3)$$

2.2 Learning a joint representation with feedforward neural networks

We learn the nonlinear map $\phi_{\mathbf{z}}(\mathbf{x})$ with a feedforward NN. For some input vector \mathbf{x} , we consider the following L -layer feedforward transformation parametrized by the weight matrices $\{\mathbf{Z}_l\}_{l=1}^L$:

$$\phi_{\mathbf{z}}(\mathbf{x}) = a_L (\mathbf{Z}_L a_{L-1} (\dots \mathbf{Z}_2 a_1 (\mathbf{Z}_1 \mathbf{x}) \dots)).$$

The parameter vector \mathbf{z} is a flattened version of the stacked weight matrices. In practice, a_l are set as \tanh functions and $L = 3$ (as [21]), but any more complex NN architecture can be used. Interestingly, we depart from [21] regarding the optimization of \mathbf{z} . While their squared-loss formulation naturally lends itself to stochastic gradient descent (SGD), in a regime with moderate values of T (typically several tens in our settings) the evidence (3) is better suited to batch optimization. In our experiments, L-BFGS [4] worked well. Unlike [21], an important by-product of this choice is that we need not find hyperparameters for SGD that should work robustly across a broad set of BO problems.

2.3 Random Fourier representation

An alternative approach is to use random kitchen sinks (RKS) for a random Fourier basis expansion [17]. Let $\mathbf{U} \in \mathbb{R}^{D \times P}$ and $\mathbf{b} \in \mathbb{R}^D$ be such that $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\{b_j\}_{j=1}^D \sim \mathcal{U}([0, 2\pi])$. For a vector \mathbf{x} , RKS defines the mapping $\phi_{\mathbf{z}}(\mathbf{x}) = \sqrt{2/D} \cos(\frac{1}{\sigma} \mathbf{U}\mathbf{x} + \mathbf{b})$, where $\sigma \in \mathbb{R}^+$ is the bandwidth of the approximated RBF kernel. The parameter vector \mathbf{z} is a flattened version of $\{\mathbf{U}, \mathbf{b}, \sigma\}$. Unlike the NN, the RKS representation contains only one hyperparameter to optimize (\mathbf{U} and \mathbf{b} are randomly generated). This reduces the complexity of learning the map, but is less expressive as we show in the following section. To optimize σ , we proceed as for the weights of the NN (see Section 2.2).

3 Results

The following subsections illustrate the benefits of multiple ABLR in a variety of settings. Sections 3.1 and 3.2 evaluate its ability to gather knowledge from multiple tasks, respectively on synthetic and OpenML data [24]. Section 3.3 shows how it can also be applied to exploit information from multiple heterogeneous signals. By doing so, we intend to learn more meaningful representations, which can be leveraged to accelerate the hyperparameter optimization. We could further generalize the model to handle multiple tasks and multiple signals at the same time, but leave this for future work.

We implemented multiple ABLR in GPyOpt [1], with a backend in MxNet [6], fully benefiting from the symbolic computation to obtain the derivatives of the mappings $\mathbf{z}, \{\alpha_t, \beta_t\}_{t=1}^T \rightarrow \rho(\mathbf{z}, \{\alpha_t, \beta_t\}_{t=1}^T)$, together with $\mathbf{x}_t^* \rightarrow \mu_t(\mathbf{x}_t^*; \mathcal{D}_t, \alpha_t, \beta_t, \mathbf{z})$ and $\mathbf{x}_t^* \rightarrow \sigma_t^2(\mathbf{x}_t^*; \mathcal{D}_t, \alpha_t, \beta_t, \mathbf{z})$. In particular, we leverage the backward operator for the Cholesky decomposition [19]. Interestingly, this allows us to jointly optimize all the model hyperparameters and perform exact BLR on top of an arbitrarily complex NN.

3.1 Transfer learning across parametrized quadratic functions

We first consider a set of T tasks. A task takes the form of a parametrized 3-dimensional quadratic function $f_t(\mathbf{x}) = \frac{1}{2}a_t\|\mathbf{x}\|_2^2 + b_t\mathbf{1}^\top\mathbf{x} + c_t$, where $(a_t, b_t, c_t) \in [0.1, 10]^3$. We call the triplet (a_t, b_t, c_t) the context associated to each task t . In a real-world setting, the contextual information would correspond to meta-data, e.g., the data set size or its dimensionality, as we shall see in the next section. We generated $T = 30$ different tasks by drawing (a_t, b_t, c_t) uniformly at random, and evaluated ABLR in a leave-one-task-out fashion. Specifically, we optimized each one of the 30 tasks after warm starting the optimization with 10 observations for the remaining 29 tasks. We compared single task ABLR-based and standard GP-based hyperparameter optimization (HPO), both denoted by `plain`, with their transfer learning counterparts, both denoted by `transfer`. We perform transfer learning with standard GPs by stacking all observations together and augmenting the input space with the corresponding contextual information [14]. For ABLR with transfer, we took our approach, i.e., one marginal likelihood per task, with and without the contextual information.

Figure 1(left) shows the current best minimum at each of 50 iterations of HPO. The results are averaged over 10 random initializations and 30 leave-one-task-out runs. HPO converges to the minimum much faster than plain ABLR or plain GP when we exploit the information from the related tasks. In addition, the RKS representation with $D = 100$ performs slightly worse than the NN with 3 hidden layers of 50 units each per layer (as advocated in [21]). Including the contextual information did not yield clear improvements, hence, for simplicity, we do not use it in the following experiments. The GP-based HPO with transfer performs slightly better on this toy example, but is not applicable in large-scale settings, such as the one in the next section (with $\sum_t N_t \approx 7.5 \times 10^5$). Figure 1(right) compares the compute time of HPO with GP and NN-based ABLR, suggesting that the linear scaling with the number of evaluations of the latter allows us to apply ABLR in the large-scale setting. The RKS basis expansion further decreases the computational time (at the expense of performance).

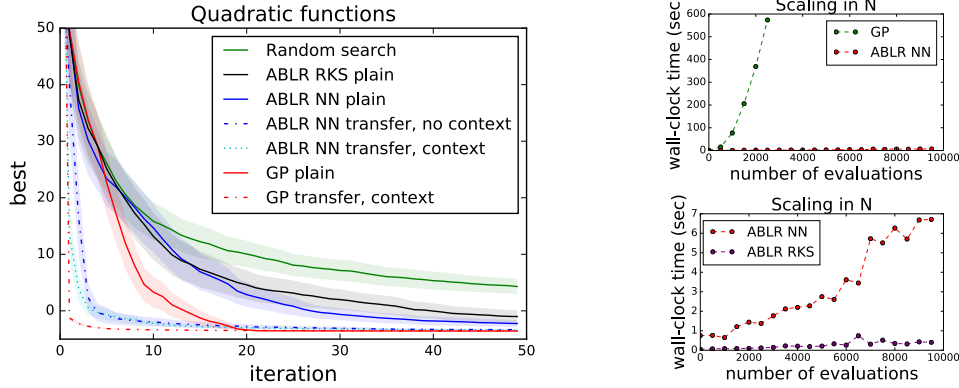


Figure 1: *Left*: Transfer learning across parametrized quadratic functions. *Right-top*: GP (cubical scaling) vs ABLR (linear scaling). *Right-bottom*: NN vs RKS basis expansion.

3.2 Transfer learning across OpenML black-box functions

We consider the OpenML platform [24], which contains a large number of evaluations for a wide range of machine learning algorithms (referred to as flows in OpenML) over different datasets. In particular, we focus on a random forest model (`flow_id 6794`) and apply ABLR to optimize its hyperparameters. We filtered the $T = 30$ most evaluated datasets for this `flow_id`, which amounts to $\sum_t N_t \approx 7.5 \times 10^5$ evaluations (with N_t ranging from 9.940 to 64.284). In this setting, the linear scaling of ABLR is particularly appealing. As previously, we apply a leave-one-task-out protocol, where each task stands for a dataset. For the left-out task being optimized, say t_0 , we use the surrogate modeling approach from [7]. We compare GP `plain` and ABLR `plain`, which use evaluations of task t_0 only, with ABLR `transfer`, which is warm-started with the evaluations of all the other tasks. The results are reported in Figure 2(left), showing that ABLR is able to gather knowledge from different datasets to speed up the convergence.

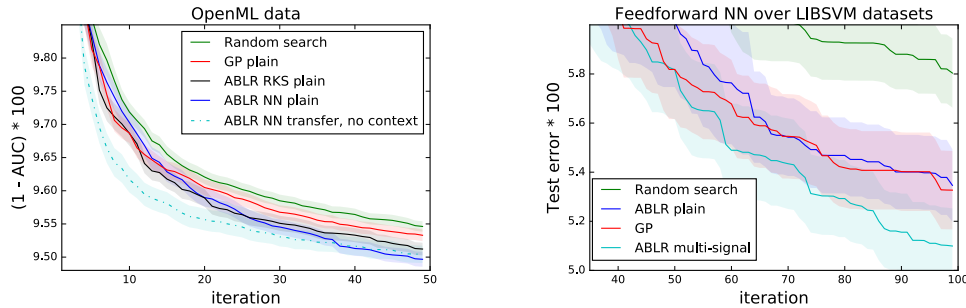


Figure 2: *Left*: OpenML data, multiple tasks. *Right*: LIBSVM data, multiple signals.

3.3 Tuning of feedforward neural networks from heterogeneous signals

Finally, we consider the tuning of feedforward NNs for binary classification. We show that our formulation can be seamlessly applied to the orthogonal problem of modeling S multiple output signals, possibly of heterogeneous nature, *at once*. Here, we optimize for the validation accuracy, using the training accuracy and CPU time as side information. Such side signals “come for free” while training machine learning algorithms, but are in general not exploited for efficient HPO. In comparison to multi-output GPs that scale as $\mathcal{O}(N^3 + S^3)$, ABLR scales as $\mathcal{O}(S(D^2N + D^3))$. The NN hyperparameters to tune are the number of hidden layers in $\{1, \dots, 4\}$, the number of units in $\{1, \dots, 50\}$, the amount of ℓ_2 regularization in $\{2^{-6}, 2^{-5}, \dots, 2^3\}$, the learning rate of Adam [13] in $\{2^{-6}, 2^{-5}, \dots, 2^{-1}\}$, and the number of epochs in $\{3, \dots, 10\}$. Figure 2(right) shows the results, which are averaged over 10 random initializations and 5 datasets (`w8a`, `sonar`, `w1a`, `phishing`, `australian`) from LIBSVM [5]. It can be observed that incorporating side signals in addition to the target signal, namely the validation accuracy of the NN classifier, speeds up the ABLR-based HPO.

References

- [1] GPyOpt: A Bayesian optimization framework in python. <http://github.com/SheffieldML/GPyOpt>, 2016.
- [2] R. Bardenet, M. Brendel, B. Kégl, and M. Sebag. Collaborative hyperparameter tuning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 199–207, 2013.
- [3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer New York, 2006.
- [4] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.
- [5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.
- [6] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. In *Neural Information Processing Systems, Workshop on Machine Learning Systems*, 2015.
- [7] K. Eggenberger, F. Hutter, H. Hoos, and K. Leyton-brown. Efficient benchmarking of hyperparameter optimizers via surrogates background: hyperparameter optimization. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 1114–1120, 2012.
- [8] M. Feurer, T. Springenberg, and F. Hutter. Initializing Bayesian hyperparameter optimization via meta-learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- [9] N. Fusi and H. M. Elibol. Probabilistic matrix factorization for automated machine learning. Technical report, preprint arXiv:1705.05355, 2017.
- [10] D. Golovin, B. Solnik, S. Moitra, G. Kochanski, J. 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.
- [11] J. M. Hernández-Lobato, J. Requeima, E. O. Pyzer-Knapp, and A. Aspuru-Guzik. Parallel and distributed Thompson sampling for large-scale accelerated exploration of chemical space. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.
- [12] R. Jenatton, C. Archambeau, J. Gonzales, and M. Seeger. Bayesian optimization with tree-structured dependencies. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.
- [13] D. Kingma and J. Ba. Adam: A method for stochastic optimization. Technical report, preprint arXiv:1412.6980, 2014.
- [14] A. Krause and C. S. Ong. Contextual gaussian process bandit optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2447–2455, 2011.
- [15] D. J. C. Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.
- [16] M. McIntire, D. Ratner, and S. Ermon. Sparse gaussian processes for Bayesian optimization. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2016.
- [17] A. Rahimi, B. Recht, et al. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, volume 3, page 5, 2007.
- [18] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- [19] M. Seeger, A. Hetzel, Z. Dai, and N. D. Lawrence. Auto-differentiating linear algebra. Technical report, preprint arXiv:1710.08717, 2017.
- [20] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

- [21] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, M. Prabhat, and R. Adams. Scalable Bayesian optimization using deep neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2171–2180, 2015.
- [22] J. T. Springenberg, A. Klein, S. Falkner, and F. Hutter. Bayesian optimization with robust Bayesian neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4134–4142, 2016.
- [23] K. Swersky, J. Snoek, and R. P. Adams. Multi-task Bayesian optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2004–2012, 2013.
- [24] J. Vanschoren, J. N. Van Rijn, B. Bischl, and L. Torgo. OpenML: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.
- [25] D. Yogatama and G. 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.