Flexible Online Aggregations Using Basis Function Expansions

Abstract

Bayesian learning often necessitates online inference, adaptive models, and the combination of multiple distinct models. Recent advancements have demonstrated the use of random feature approximations for scalable, online aggregation of Gaussian processes, which possess favorable theoretical characteristics and practical uses. A crucial aspect of these methods is the incorporation of a random walk on model parameters, which introduces adaptability. We demonstrate that these methods can be readily extended to any model using basis function expansion and that employing alternative basis expansions, such as Hilbert space Gaussian processes, frequently leads to enhanced performance. To streamline the selection of a specific basis expansion, the versatility of our approach also enables the aggregation of several entirely different models, such as a Gaussian process and polynomial regression. Lastly, we introduce an innovative technique for combining both static and dynamic models.

1 Introduction

Numerous machine learning applications demand real-time, online data processing, a scenario that frequently requires substantial alterations to conventional techniques. Online adaptations of various methods have been developed, including kernel machines, (kernel) least-squares, and Gaussian processes. The field of online learning has also been thoroughly investigated from an optimization standpoint.

Online learning can be further complicated when model selection is needed, as the best-performing model is rarely evident at the outset of the learning process. One solution involves training multiple models concurrently and then combining them. In a Bayesian framework, Bayesian model averaging (BMA) has long been employed to combine online models, functioning by assigning weights to each "expert" model based on its supporting evidence.

More recently, it was shown how to adapt BMA to online Gaussian processes (GPs) in a technique called incremental ensembles of GPs. GPs are a versatile, non-parametric instrument in Bayesian machine learning that possesses universal approximation capabilities and provides well-founded uncertainty estimations. By employing a random Fourier feature (RFF) approximation for Gaussian processes, online learning can be executed, featuring closed-form Bayesian model averaging updates and a manageable regret analysis.

Besides an online ensemble of GPs, the advantages of incorporating random walks on model parameters were illustrated, which they term dynamic IE-GPs (DIE-GPs). This can significantly enhance performance when the learning task undergoes slight changes over time.

2 Related Work

The concept of combining random feature GPs, as introduced by IE-GPs, has demonstrated adaptability and effectiveness. Extensions to this framework encompass Gaussian process state-space models, deep Gaussian processes, and graph learning. Along with its extensions, DIE-GPs have been effectively applied in Bayesian optimization and causal inference.

However, the dependence on the RFF approximation implies that IE-GPs also inherit the limitations of random feature GPs. Specifically, the RFF approximation is a direct Monte Carlo approximation of the Wiener-Khinchin integral and thus is significantly impacted by the curse of dimensionality. Our findings reveal that on several real-world datasets, (D)IE-GPs exhibit performance that is comparable to or worse than that of simpler models, such as online Bayesian linear regression and one-layer RBF networks.

3 Methodology

In this paper, we present online ensembles of basis expansions (OEBEs), a generalization of IE-GPs that overcomes their dependence on RFF GPs and enhances performance across multiple real datasets. Our specific contributions are as follows:

1. We observe that the derivation of DIE-GPs does not rely on the RFF approximation, except for its role as a linear basis expansion. The same derivations and code can be reused to combine arbitrary Bayesian linear models with any design matrix. This allows for the combination of not only models of the same type but also various distinct basis expansions (e.g., B-splines, one-layer RBF networks, etc.). 2. We contend that a GP with a generalized additive model (GAM) structure is often more suitable when GP regression is the focus. To this end, we employ GAM Hilbert space Gaussian processes (HSGPs), which can be interpreted as a quadrature rule for the same integral that the RFF approximation addresses through direct Monte Carlo. Apart from theoretical considerations, empirical evidence indicates that HSGPs converge to the true approximated GP more rapidly (in terms of the number of basis functions) than RFF GPs. We offer a similar empirical evaluation. 3. We introduce a new method for integrating static and dynamic models, enabling the use of principled posteriors of static methods when appropriate and extending the expressiveness of dynamic methods otherwise. We demonstrate the necessity of this method by providing a constructive example on real data where the naive approach to combining static and dynamic methods is unsuccessful. 4. We provide Jax/Objax code at `https://www.github.com/danwaxman/DynamicOnlineBasisExpansions` that only requires the user to specify the design matrix, with several choices already implemented.

The remainder of this paper is organized as follows: Section 2 reviews foundational concepts in linear basis expansions, GP regression, spectral approximations of GPs, and BMA. These concepts are put into practice in Section 3, where we present the OEBEs and several extensions, including applications to non-Gaussian likelihoods, and provide some concise theoretical observations. We offer further practical insights regarding the development of OEBEs, including a discussion on the composition of an ensemble and how to combine static and dynamic models in Section 4. The proposed models are empirically evaluated in Section 5. Finally, we present concluding remarks and suggest future directions in Section 6.

4 Experiments

We present three distinct experiments in the main text, with supplementary experiments in the appendices. In the first experiment (Section 5.1), we assess ensembles of several different basis expansions, demonstrating that the best-performing model varies considerably. In the second experiment (Section 5.2), we illustrate how model collapse can occur between static and dynamic models and how the model introduced in Section 4.2 mitigates this issue. Lastly, we demonstrate that E-DOEBE can effectively combine methods that are both static and dynamic, and of different basis expansions (Section 5.3).

The metrics we employ are the normalized mean square error (nMSE) and the predictive log-likelihood (PLL). The nMSE is defined as the MSE of yt with the predictive mean, divided by the variance of y1:T. Specifically, at time t, the nMSE is calculated as:

$nMSE_t = \frac{\sum_{\tau=1}^{t}(\mu_{y_\tau} - y_\tau)^2}{t \cdot Var(y_{1:T})}$

The predictive log-likelihood (PLL) is the average value of log p(yt+1|X1:t, y1:t), i.e.,

$PLL_t = \frac{\sum_{\tau=1}^{t} \log p(y_{\tau+1}|X_{1:\tau}, y_{1:\tau})}{t}$.

Across all experiments, we utilize several publicly available datasets, varying in both size and the number of features. A summary of dataset statistics is provided in Table 1. Friedman 1 and Friedman 2 are synthetic datasets designed to be highly nonlinear and, notably, are i.i.d. The Elevators dataset pertains to controlling the elevators on an aircraft. The SARCOS dataset uses simulations of a robotic arm, and Kuka 1 is a similar real dataset derived from physical experiments. CaData comprises California housing data, and the task of CPU Small is to predict a type of CPU usage based on system properties.

All hyperparameter optimization was performed on the first 1,000 samples of each dataset; since we already assume access, each dataset was additionally standardized in both x and y using the statistics of the first 1,000 samples. We follow prior work in setting a weight to 0 when it falls below the threshold of 10^{-16}.

4.1 Comparing Different Basis Expansions

To demonstrate that having a diverse set of basis expansion models available is beneficial, we evaluate several model types on each dataset listed in Table 1. Furthermore, we examine both static and dynamic versions of models to assess their performance.

Models used for comparison include an additive HSGP model [(D)OE-HSGP], an RFF GP [(D)OE-RFF], an ensemble of quadratic, cubic, and quartic polynomials with additive structure [(D)OE-Poly], linear regression [(D)OE-Linear], and a one-layer RBF network [(D)OE-RBF]. Apart from additional hyperparameter tuning in an ARD kernel, the (D)OE-RFF model is identical to the (D)IE-GP.

For RFF GPs, 50 Fourier features were employed (resulting in F = 2 × 50), and for HSGPs, Ž30a100/DŽ30b features were used for each dimension (resulting in F Ž272 100). An SE-ARD kernel was utilized in both cases. For RBF networks, 100 locations were initialized using K-means and subsequently optimized with empirical Bayes, along with ARD length scales. For all models except RBF networks, ensembles were generated using the process outlined in Section 4.1 Ž014 for RBF networks, the computation of the Hessian was too computationally demanding, so parameters were randomly perturbed by white Gaussian noise with variance 10^{-3} instead.

For dynamic models, Ď3c3^2 was set to 10^{-3}. The initial values of Ď3c3^{2}Ď3b8 and Ď3c3^{2}Ď3f5 were 1.0 and 0.25, respectively. Optimization was carried out using Adam.

Results of the average nMSE and PLL are presented in Table 2 and Table 3. We observe that the best-performing class of models varies significantly across datasets. Specifically, in terms of both nMSE and PLL, HSGPs, RFF GPs, and RBF networks each achieve the best performance on at least one dataset. This reinforces the notion that combining several different models is advantageous, as no single method consistently outperforms the others.

Moreover, as anticipated, dynamic models can substantially outperform static models in specific scenarios (e.g., on SARCOS and Kuka 1) but yield a lower PLL on datasets where the data is reasonably i.i.d. (e.g., Friedman 1).

As expected, when an additive structure is a reasonable approximation, additive HSGP methods surpass RFF GPs, for instance, on Kuka 1 and CaData. The RFF GP approximation rarely exhibits particularly poor performance, making it a consistently "good" estimator, and it achieves the highest PLL on Friedman 2, SARCOS, and CPU Small. However, it is also occasionally outperformed by simpler methods, such as the RBF network, highlighting the potential advantages of employing diverse basis expansions.

Key Takeaways Key takeaways from this experiment include: (1) neither dynamic nor static methods are strictly superior across all settings, (2) no single basis expansion is superior across all datasets, and (3) RFF GPs consistently provide good performance, but this performance can often be improved upon by using other basis expansions.

4.2 The Necessity of Ensembles of Dynamic Ensembles

In this experiment, we demonstrate that the E-DOEBE model introduced in Section 4.2 can indeed prevent the premature collapse of BMA weights. While this premature collapse of BMA weights does not appear to be common in real datasets, it is not difficult to illustrate its possibility, even on real datasets with high-performing methods.

As a constructive example, we can create an ensemble of additive HSGPs on the Kuka 1 dataset, where dynamic models performed significantly better in Section 5.1. Specifically, we created an ensemble of two additive HSGPs, with the first model being dynamic ($\check{0}3c3(1)rw^2 = 10^{-3}$) and the second model being static ($\check{0}3c3(2) = 0$). The ensemble hyperparameters were determined using empirical Bayes, with initial length scale values set to the vector of ones. Subsequently, the resulting ensemble was trained online as a DOEBE and as an E-DOEBE, with $\check{0}3b4 = 10^{-2}$. Note that in this carefully controlled setting, each basis expansion is entirely deterministic given the hyperparameters, so the results are purely deterministic and cannot be attributed to poor random seeds.

The resulting weights demonstrate that premature collapse of BMA weights can be a problem. Numerically, the log-likelihood of the E-DOEBE model is dramatically better than that of the DOEBE model (Table 2), showing this collapse can be catastrophic.

This issue can be partially averted by eliminating the threshold of 10^{-16} when ensembling. Indeed, in this example, the weights reach a minimum of approximately 10^{-72}. However, with any finite precision arithmetic, there is always the potential for this type of collapse to occur due to numerical underflow. It is trivial to construct such examples by generating the first N1 samples with $\check{0}3c3(m)rw = 0$ until weight collapse occurs, and the rest of the dataset with $\check{0}3c3(m)rw > 0$.

Key Takeaway The key takeaway of this experiment is that an ensemble of dynamic and static models can catastrophically collapse Ž014 even when the discrepancy in performance along the entire dataset is large Ž014 and that the E-DOEBE approach proposed in Section 4.2 can avoid this collapse.

4.3 E-DOEBE Outperforms Other Methods

The ultimate goal of the E-DOEBE model is to combine static and dynamic models of several different types. To do so, we repeat the experiments of Section 5.1 while comparing to an E-DOEBE model. We restrict our attention to static and dynamic versions of the three best-performing families of models in Experiment 1 ((D)OE-HSGP, (D)OE-RFF, and (D)OE-RBF), and an E-DOEBE ensemble containing all of them. The E-DOEBE model is created with $\check{0}3b4 = 10^{-2}$, which was not tuned.

As desired, the E-DOEBE model can effectively ensemble dynamic and static models of different basis expansions. Across all experiments, the E-DOEBE model performs the best in terms of PLL, and is the best in terms of NMSE for all but one dataset (Friedman 2).

Key Takeaway The E-DOEBE can effectively ensemble several different ensembles of high-performing basis expansions, resulting in consistently better performance than any single method.

5 Conclusion

In this paper, we demonstrated that recent advancements in online prediction using RFF GPs can be extended to arbitrary linear basis expansions. This included several basis expansions that surpass RFF GPs on real and synthetic datasets. We show how different linear basis expansions can be combined within a simple framework, enhancing ensemble diversity. While several common choices of basis expansions were employed, it would be worthwhile to expand the tests even further, particularly with splines.

We also demonstrated that the premature collapse of BMA weights can be a concern in online combining. We introduced the E-DOEBE model, which mitigates this issue, and demonstrated its effectiveness. However, this meta-combining may be perceived

as adding a complex workaround to BMA rather than addressing the underlying problems. Further research could explore the incorporation of other Bayesian combining methods, such as Bayesian (hierarchical) stacking.

While we provide guidance on initializing ensembles given a set of basis expansions, determining which basis expansions to use is an important open topic. A naive approach would be to expand on the existing use of the marginal likelihood for model selection, but this may be "unsafe" when using different basis expansions and therefore requires caution. We additionally presented several ideas for inference with non-Gaussian likelihoods, for example, for classification tasks. Determining which, if any, of these tasks is superior to the Laplace approximation is another interesting topic for future study.

Finally, it could be beneficial to modify or add new basis expansions in the online setting. Indeed, recent progress in GPs has worked towards selecting and adapting kernels online to great benefit. If such techniques could be adapted to DOEBE, it could eliminate the pre-training period and allow for adapting the domain of approximations when new data arrives.

6 Tables

Table 1: Dataset statistics, including the number of samples, the number of features, and the original source. In addition to the original sources above, several of these datasets were curated by the UCI Machine Learning Repository or LibSVM.

Dataset Name	Number of Samples	Dimensionality d
Friedman 1	40,000	10
Friedman 2	40,000	4
Elevators	16,599	17
SARCOS	44,484	21
Kuka 1	197,920	21
CaData	20,640	8
CPU Small	8,192	12

Table 2: Predictive log-likelihood of DOEBE and E-DOEBE models in Experiment 2 (higher is better).

Method	Predictive Log-Likelihood
DOEBE	-403.41
E-DOEBE	0.55

Table 3: Predictive likelihood (higher is better) and normalized MSE (lower is better) of type-II MLE and Laplace-approximated initialization, plus/minus one standard deviation over 100 trials. Bolded entries denote superior performance significant at the p = 0.05 level according to a one-sided Wilcoxon rank-sum test.

2*Method	Predictive Log-Likelihood			Normalized Mean Square Error		
	Elevators	SARCOS	CaData	Elevators	SARCOS	CaData
DOE-HSGP-MLE	-0.753 $\pm$ 0.000	0.421 $\pm$ 0.000	0.081 $\pm$ 0.000	0.221 $\pm$ 0.000	0.017 $\pm$ 0.000	0.055 $\pm$ 0.000
DOE-HSGP-Sample	**-0.748 $\pm$ 0.003**	**0.466 $\pm$ 0.010**	**0.120 $\pm$ 0.010**	**0.219 $\pm$ 0.001**	**0.018 $\pm$ 0.000**	**0.052 $\pm$ 0.001**
DOE-RFF-MLE	-0.640 $\pm$ 0.007	0.756 $\pm$ 0.018	0.243 $\pm$ 0.009	0.178 $\pm$ 0.003	0.018 $\pm$ 0.001	0.040 $\pm$ 0.001
DOE-RFF-Sample	**-0.639 $\pm$ 0.007**	**0.766 $\pm$ 0.019**	**0.247 $\pm$ 0.009**	**0.177 $\pm$ 0.004**	**0.018 $\pm$ 0.001**	**0.040 $\pm$ 0.002**

Table 4: Dataset statistics, including the number of samples and the number of features for datasets used in Delbridge et al. (2020). All datasets are available on the UCI Machine Learning Repository.

Dataset Name	Number of Samples	Dimensionality d
autos	159	25
servo	167	4
machine	209	7
yacht	308	6
autompg	392	7
housing	506	13
stock	536	11
energy	768	8
concrete	1,030	8
airfoil	1,503	5
gas	2,565	128
skillcraft	3,338	19
sml	4,137	26
pol	15,000	26
bike	17,379	17
kin40k	40,000	8