Supplementary Material

High-Dimensional Bayesian Optimization via Additive Models with Overlapping Groups (AISTATS 2018)

Paul Rolland, Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher

All citations below are to the reference list in the main document.

A Derivation of posterior mean and variance

In this section, we prove equation (4) for the posterior mean and variance conditioned on the observations. The derivation is similar to that of Kandasamy *et al.* [19], which in turn is similar to the standard Gaussian process posterior derivation [35].

Recall that instead of directly computing the posterior mean and variance on the high-dimensional function, we are considering the terms $f^{(j)}$ in the additive decomposition of f separately. We are thus interested in the distributions of $f_*^{(j)} = f^{(j)}(x_*^{(j)})$, $j = 1, ..., M$ conditioned on the noisy samples $\mathbf{y} = y_1, ..., y_n$ at points $\mathbf{x} = x_1, ..., x_n$, for some query points $x_*^{(j)}$. We claim that the joint distribution of $f_*^{(j)}$ and $\mathbf{y}$ can be written as

$$\begin{pmatrix} f_*^{(j)} \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \kappa^{(j)}(x_*^{(j)}, x_*^{(j)}) & \kappa^{(j)}(x_*^{(j)}, \mathbf{x}^{(j)}) \\ \kappa^{(j)}(\mathbf{x}^{(j)}, x_*^{(j)}) & \kappa(\mathbf{x}, \mathbf{x}) + \eta^2 I_n \end{pmatrix}\right) \tag{9}$$

To see this, we recall that distinct functions in the additive decomposition (1) are independent given $\mathbf{x}$. Hence, for any observation $y_p = f(x_p) + \epsilon$ and $j = 1, ..., M$, we have

$$\begin{aligned} \mathrm{Cov}(f_*^{(j)}, y_p) &= \mathrm{Cov}\left(f_*^{(j)}, \sum_{i=1}^{M} f^{(i)}(x_p^{(i)}) + \epsilon\right) \\ &= \mathrm{Cov}\left(f_*^{(j)}, f^{(j)}(x_p^{(j)})\right) \\ &= \kappa^{(j)}(x_*^{(j)}, x_p^{(j)}), \end{aligned}$$

which establishes (9).

With (9) in place, we can use a standard conditional Gaussian formula (as used in standard GP posterior derivations [35], as well as the non-overlapping setting of [19]) to derive the posterior mean and variance. Specifically, defining the matrix $\Delta = \kappa(\mathbf{x}, \mathbf{x}) + \eta^2 I_n \in \mathbb{R}^{n \times n}$, we have for past query points $\mathbf{x}$ and next query point $x_*^{(j)}$ that

$$\begin{aligned} (f_*^{(j)} | \mathbf{y}) \sim \mathcal{N}\Big(&\kappa^{(j)}(x_*^{(j)}, \mathbf{x}^{(j)})\Delta^{-1}\mathbf{y}, \\ &\kappa^{(j)}(x_*^{(j)}, x_*^{(j)}) - \kappa^{(j)}(x_*^{(j)}, \mathbf{x}^{(j)})\Delta^{-1}\kappa^{(j)}(\mathbf{x}^{(j)}, x_*^{(j)})\Big) \end{aligned} \tag{10}$$

under the notation in (4). This concludes the derivation.

B Mathematical analysis and theoretical challenges

B.1 Discussion on existing theory

Guarantees of GP-UCB. A notable early work providing theoretical guarantees on Bayesian optimization (without the high-dimensional aspects) is that of Srinivas *et al.* [33], who considered the *cumulative regret* $R_T = \sum_{t=1}^{T} \big(f(x_{\mathrm{opt}}) - f(x_t)\big)$ with $x_{\mathrm{opt}} = \arg\max_{x \in \mathcal{X}} f(x)$. In the case of a finite domain $\mathcal{X}$, it was shown that the GP-UCB algorithm with exploration parameter $\beta_t = 2\log\big(\frac{|\mathcal{X}|t^2\pi^2}{6\delta}\big)$ achieves

$$R_T \leq \sqrt{\frac{8}{\log(1+\sigma^{-2})} T \beta_T \gamma_T} \tag{11}$$

with probability at least $1-\delta$. Here the kernel-dependent quantity γ_T is known as the *information gain*, and is defined as the maximum of a mutual information quantity:

$$\gamma_T = \max_{A\,:\,|A|=T} I(y_A; f_A) \tag{12}$$

for $A = \{x_1, ..., x_T\}$ with corresponding function values f_A and observations y_A. Analogous results were presented for the continuous setting in [33] under mild technical assumptions, and explicit bounds on γ_T were provided for the squared exponential and Matérn kernels.

Extension to non-overlapping additive models. Kandasamy *et al.* [19] attempted to upper bound the cumulative regret of their algorithm Add-GP-UCB in the high-dimensional setting with additive models. In particular, they sought a bound whose complexity is only exponential in the maximal dimension d of the low-dimensional kernels, instead of the full dimension D as in Srinivas *et al.* [33]. However, they subsequently stated in an updated version of their paper that the proof contains an error [36]. In our understanding, the error is due to the fact that the approximated standard deviation $\sum_{i=1}^M \sigma_{t-1}^{(i)}$ is different from the true standard deviation σ_{t-1}, and that the ratio $\frac{\sum_{i=1}^M \sigma_{t-1}^{(i)}(x)}{\sigma_{t-1}(x)}$ cannot be upper bounded for all x (see below).

In the parallel independent work of Hoang *et al.* [25], it was shown that a sufficient condition that leads to sub-linear regret bounds in high-dimensional additive models is as follows [25, Assumption 4]: The posterior variance $\sigma_t^{(i)}(x^{(i)})$ of each component i given the observations can be upper bounded by a constant times $\widehat{\sigma}_t^{(i)}(x^{(i)})$, defined to be the posterior variance as if the corresponding function $f^{(i)}$ had been sampled directly instead. However, it remains an open problem to determine specific models and kernels for which this assumption is true.

Outline of this appendix. We further discuss the relation between the true and approximate posterior standard deviations in Section B.2, and then provide a novel bound on the information gain for our setting in Section B.3. While the latter is only one step towards attaining a regret bound for G-Add-GP-UCB, it also quantifies the regret bound (11) when GP-UCB is applied to the high-dimensional setting. Unfortunately, GP-UCB is not computationally feasible in high dimensions, so establishing a similar regret bound G-Add-GP-UCB remains an important direction for future research.

B.2 Relation between true posterior variance and its approximation

Our algorithm G-Add-GP-UCB is based on an acquisition function that can be computed efficiently in high dimensions. This property comes from the fact that it that can be decomposed into the sum of low-dimensional components (see (6)). Each term in the sum consists of a mean and standard deviation corresponding to a low-dimensional function.

We observe that $\tilde{\mu}_{t-1}(x) = \sum_{j=1}^M \mu_{t-1}^{(j)}(x^{(j)}) = \mu_{t-1}(x)$. Therefore, this way of splitting the posterior mean into several lower dimensional components does not involve any approximation. However, $\tilde{\sigma}_{t-1}(x) = \sum_{j=1}^M \sigma_{t-1}^{(j)}(x^{(j)}) \neq \sigma_{t-1}(x)$ in general; this can be viewed as being due to the non-linearity of the quadratic term $\kappa(x_*, \mathbf{x})\Delta^{-1}\kappa(\mathbf{x}, x_*)$ in the posterior variance (4).

Our analysis below reveals that

$$\sum_{j=1}^M \sigma_{t-1}^{(j)}(x^{(j)}) \geq \sigma_{t-1}(x). \tag{13}$$

Thus, this splitting of the posterior standard deviation into low-dimensional components generally over-estimates the true variance. An example where the inequality is strict is as follows: In the case of zero noise, the true posterior standard deviation at a point $x_{\text{evaluated}}$ that has already been evaluated is $\sigma_{t-1}(x_{\text{evaluated}}) = 0$. However in general, $\sum_{j=1}^M \sigma_{t-1}^{(j)}(x_{\text{evaluated}}) > 0$. Therefore, the ratio $\frac{\sum_{j=1}^M \sigma_{t-1}^{(j)}(x)}{\sigma_{t-1}(x)}$ can sometimes diverge, and it is thus not possible to upper bound it for all x.

Derivation of the upper bound (13)**.** The true posterior variance based on observations at the points $\mathbf{x} = (x_1, ..., x_t)$ is given by

$$\sigma_t(x)^2 = \kappa(x, x) - \kappa(x, \mathbf{x})\Delta^{-1}\kappa(\mathbf{x}, x), \tag{14}$$

where κ is the full dimensional kernel $\kappa(x, x') = \sum_{i=1}^M \kappa^{(i)}(x^{(i)}, x'^{(i)})$, $k(x, \mathbf{x})$ and $k(\mathbf{x}, x)$ are the corresponding vectors of kernel values, and Δ is a matrix such that $\Delta_{ij} = \kappa(x_i, x_j)$ for $i, j = 1, ..., t$. The approximated posterior

variance based on the same points is as follows:

$$\sum_{i=1}^{M} \sigma_t(x^{(i)})^2 = \sum_{i=1}^{M} \kappa^{(i)}(x^{(i)}, x^{(i)}) - \kappa^{(i)}(x^{(i)}, \mathbf{x}^{(i)})\Delta^{-1}\kappa^{(i)}(\mathbf{x}^{(i)}, x^{(i)}) \tag{15}$$

$$= \kappa(x, x) - \sum_{i=1}^{M} \kappa^{(i)}(x^{(i)}, \mathbf{x}^{(i)})\Delta^{-1}\kappa^{(i)}(\mathbf{x}^{(i)}, x^{(i)}) \tag{16}$$

under the notation following (4).

By definition, the matrix Δ is symmetric and positive definite, and hence so is the matrix Δ^{-1}. We can thus define a norm induced by this matrix on the space $\mathbb{R}^t$; for $\vec{k} \in \mathbb{R}^t$, we have

$$\|\vec{k}\|^2_{\Delta^{-1}} = \vec{k}^T \Delta^{-1} \vec{k}. \tag{17}$$

The fact that Δ^{-1} is symmetric positive definite implies that this has all the properties of a norm. For any $x \in \mathbb{R}^D$, we define the t-dimensional vector $\vec{k}^{(i)}(x)$ as $\vec{k}^{(i)}(x)_j = \kappa^{(i)}(x^{(i)}, x_j^{(i)})$. We also recall that $\kappa(x, x) = 1$ for all $x \in \mathbb{R}^D$. Using this notation, we can rewrite the expressions for the true and approximated posterior variances as

$$\sigma_t(x)^2 = 1 - \Big\| \sum_{i=1}^{M} \vec{k}^{(i)}(x) \Big\|^2_{\Delta^{-1}} \tag{18}$$

and

$$\sum_{i=1}^{M} \sigma_t(x^{(i)})^2 = 1 - \sum_{i=1}^{M} \|\vec{k}^{(i)}(x)\|^2_{\Delta^{-1}}. \tag{19}$$

By the triangle inequality, we have

$$\begin{aligned}
\sigma_t(x)^2 &= 1 - \Big\| \sum_{i=1}^{M} \vec{k}^{(i)}(x) \Big\|^2_{\Delta^{-1}} \\
&\leq 1 - \Big(\sum_{i=1}^{M} \|\vec{k}^{(i)}(x)\|_{\Delta^{-1}} \Big)^2 \\
&\leq 1 - \sum_{i=1}^{M} \|\vec{k}^{(i)}(x)\|^2_{\Delta^{-1}} \\
&= \sum_{i=1}^{M} \sigma_t(x^{(i)})^2
\end{aligned}$$

As $\sigma_t(x^{(i)}) \geq 0 \ \forall x \in \mathbb{R}^D$, we have that $\sum_{i=1} \sigma_t(x^{(i)})^2 \leq \left(\sum_{i=1} \sigma_t(x^{(i)})\right)^2$, which implies

$$\sigma_t(x) \leq \sum_{i=1}^{M} \sigma_t(x^{(i)}) \tag{20}$$

as desired.

Numerical evaluation. In order to observe the difference between the true posterior variance σ_t^2 and the approximated variance $\left(\sum_{j=1}^{M} \sigma_{t-1}^{(j)}\right)^2$, we generate a 10 dimensional synthetic function via Gaussian processes with the star dependency graph shown in Figure 2. We first evaluate this function at 200 randomly-selected points. Based on these observations, we then evaluate the true and approximated posterior variances at 1000 randomly selected points. We then compare the results obtained with these two different methods (Figure 8).

From the figure on the right, we can observe that the obtained values can be significantly different. However, there is a clear correlation between these two quantities, in the sense that points with higher true variance tend to also have a higher approximated variance. The figure on the left shows that the ratio between approximated and true variance increases as the true variance becomes smaller. This approximation error strongly depends on the decomposition, and in particular on M. The higher the value of M, the higher the approximation error.

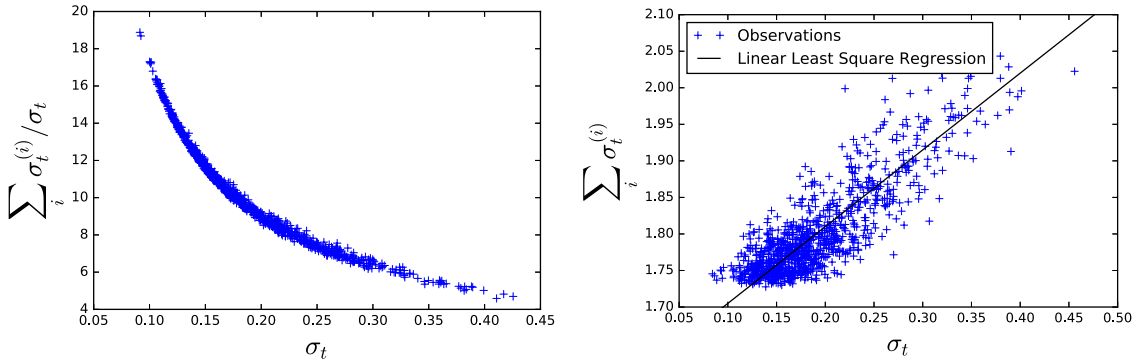

Figure 8: Evaluation of the posterior standard deviation from the full dimensional kernel (true posterior variance), vs. separating the kernel into lower dimensional additive components (approximated posterior variance). Left: Ratio between the two computed posterior standard variations as a function of the true posterior standard variation. Right: Approximated posterior standard deviation versus true posterior standard deviation.

B.3 Bounding the information gain with overlapping groups

As outlined above, the information gain γ_T in (12) plays a crucial role in the regret bounds for Bayesian optimization. While we do not claim any regret bounds for G-Add-GP-UCB, bounding γ_T may provide an initial step towards this, and also allows us to understand the performance of GP-UCB with our structured kernels (*cf.*, (11)). We provide such a bound for our setting, focusing on the squared exponential kernel, and ultimately showing that $\gamma_T = O(Dd^d(\log T)^{d+1})$ analogously to [19]. Here d is the highest dimension of any low-dimensional function in the additive decomposition.

We follow the high-level steps of [19] with suitable modifications for our setting. It was shown in [26] that under some mild assumptions on the target function f, the maximal information gain can be bounded as

$$\gamma_T \leq \inf_{\tau} \left(\frac{1/2}{1-e^{-1}} \max_{r \in \{1,\dots,T\}} (T_* \log(rn_T/\eta^2) + C_9 \eta^2 (1 - \frac{r}{T})(T^{r+1} B_\kappa(T_*) + 1) \log T) + O(T^{1-\tau/D}) \right), \tag{21}$$

for any $T_* \in \{1, \dots, \min(T, n_T)\}$, where $C_9 = 4D + 2$, $n_T = C_9 T^\tau \log T$, and $B_\kappa(T_*) = \sum_{s > T_*} \lambda_s$. Here $\{\lambda_n\}_{n \in \mathbb{N}}$ are the eigenvalues of κ with respect to the uniform distribution.

In order to bound γ_T, it therefore suffices to bound $B_\kappa(T_*)$, i.e., to bound the sum of the eigenvalues of κ at the tail. Unlike the setting of [19], the eigenfunctions corresponding to different kernels $\kappa^{(i)}$ and $\kappa^{(j)}$ are not necessarily orthogonal, since overlaps between kernel variables are possible. To circumvent this difficulty, we can make use of Weyl's inequality.

Lemma 1. (Weyl's inequality) *Let $H, P \in \mathbb{R}^{n \times n}$ be two Hermitian matrices, and define $M = H + P$. Let μ_i, ν_i, ρ_i, $i = 1, \dots, n$ be the eigenvalues of M, H and P respectively such that $\mu_1 \geq \dots \geq \mu_n$, $\nu_1 \geq \dots \geq \nu_n$ and $\rho_1 \geq \dots \geq \rho_n$. Then for all $i \geq r + s - 1$, we have*

$$\mu_i \leq \nu_r + \rho_s \tag{22}$$

This result immediately generalizes to a sum with an arbitrary number of matrices. In particular, we will use the following corollary.

Corollary 1. *Let $K_i \in \mathbb{R}^{n \times n}$, $i = 1, \dots, M$ be Hermitian matrices, and define $K = \sum_{i=1}^M K_i$. Let $\{\lambda_j^{(i)}\}_{j=1,\dots,n}$, be the eigenvalues of K_i such that $\lambda_1^{(i)} \geq \dots \geq \lambda_n^{(i)}$ $\forall i = 1, \dots, M$, and let $\{\lambda_i\}_{i=1,\dots,n}$ be the eigenvalues of K such that $\lambda_1 \geq \dots \geq \lambda_n$. Then for all $k \in \mathbb{N}$ such that $kM + 1 \leq n$, we have*

$$\lambda_{kM+1} \leq \sum_{i=1}^{M} \lambda_{k+1}^{(i)}. \tag{23}$$

Let $\{\lambda_s\}_{s \in \mathbb{N}}$, $\lambda_1 \geq \lambda_2 \geq \dots$ denote the eigenvalues of κ, and for all $j = 1, \dots, M$, let $\{\lambda_s^{(j)}\}_{s \in \mathbb{N}}$, $\lambda_1^{(j)} \geq \lambda_2^{(j)} \geq \dots$ denote the eigenvalues of $\kappa^{(j)}$. It was shown by Seeger *et al.* [37] that the eigenvalues $\lambda_s^{(j)}$ for the square

exponential kernel $\kappa^{(j)}$ satisfy $\lambda_s^{(j)} \leq c^d B^{s^{1/d}}$, $B < 1$, where each $\kappa^{(j)}$ is a d_j-dimensional kernel, and $d_j \leq d$. Defining $T_+ = \left\lfloor \frac{T_*}{M} \right\rfloor$ we have the following:

$$B_\kappa(T_*) = \sum_{s>T_*} \lambda_s \tag{24}$$

$$\leq \sum_{k>T_+} \sum_{i=1}^{M} \lambda_{(k-1)M+l} \tag{25}$$

$$\leq \sum_{k>T_+} \sum_{i=1}^{M} \sum_{j=1}^{M} \lambda_k^{(j)} \tag{26}$$

$$\leq M^2 c^d \sum_{k>T_+} B^{k^{1/d}}, \tag{27}$$

where the second line uses the fact that the eigenvalues are increasingly ordered, the third line follows from Weyl's inequality, and the final line follows from the bound on the tail eigenvalues given in [37].

The rest of the proof follows via a similar analysis to [19]. One difference is that we get an extra M term in our bound for B_κ compared to the setting of [19]. However, this does not affect the bound for γ_T, since the leading term on the right hand side of (21) is $T_* \log(rn_T/\eta^2)$ which does not involve B_κ. We thus obtain the same bound for γ_T as in [19], namely,

$$\gamma_T = O(Dd^d(\log T)^{d+1}). \tag{28}$$

Note that this bound only has linear dependence on the dimension D, while being exponential in the maximal group size d.

C Astrophysical data experiment

In this appendix, we consider an additional experiment on real-world data that aims at estimate a set of 9 cosmological parameters (e.g., Hubble's constant, proportion of helium in the universe, etc) in order to best match reality. These constants are involved in the theoretical model of physics, but are estimated experimentally. To do so, programs model the dynamics of the universe given these parameters, and compare the results of the simulations with the observed data.

For each set of parameters, we can compute the likelihood that the chosen parameters match the reality. The aim is thus to find the set of parameters that maximize this likelihood, or equivalently that minimize the negative log-likelihood. We use the LRG DR7 Likelihood Software released by NASA[1] in order to compute likelihoods given these cosmological parameters based on experimental data released by the Sloan Digital Sky Survey.

We note that this data was used by both Kandasamy *et al.* [19] and Gardner *et al.* [22] for testing high-dimensional BO algorithms, but it was used in somewhat different ways. We adopt the approach of [22], and we avoid adding additional "dummy dimensions" as in [19].

The software provides a set of parameters which achieves a negative log-likelihood of 23.7. We thus apply the two Bayesian algorithms "Overlap" and "No Overlap" in a range of $75\% - 125\%$ of this default set of parameters. Unlike the previous real world experiment, we do not set a fixed dependency graph for the "Overlap" model and learn it throughout the algorithm using Gibbs sampling (*cf.*, Section 4). Similarly, for the "No Overlap" model, we use the Gibbs sampling approach of [21], placing no "hard constraints" (i.e., choices of M and d in [19]).

The remaining parameters for the Bayesian optimization algorithms are the same as for the first real-world experiment (Table 2), except that we run the simulation for 1000 iterations. The results are shown in Figure 9. We observe that both algorithms achieve a higher likelihood than with the default parameters, and that the "Overlap" algorithm achieves a higher likelihood than the "No Overlap" one across the entire time horizon. In particular, by the final iteration, the gap to "random" has increased from approximately 0.09 to 0.15, i.e., an increase of over 60%.

[1] http://lambda.gsfc.nasa.gov/toolbox/lrgdr/

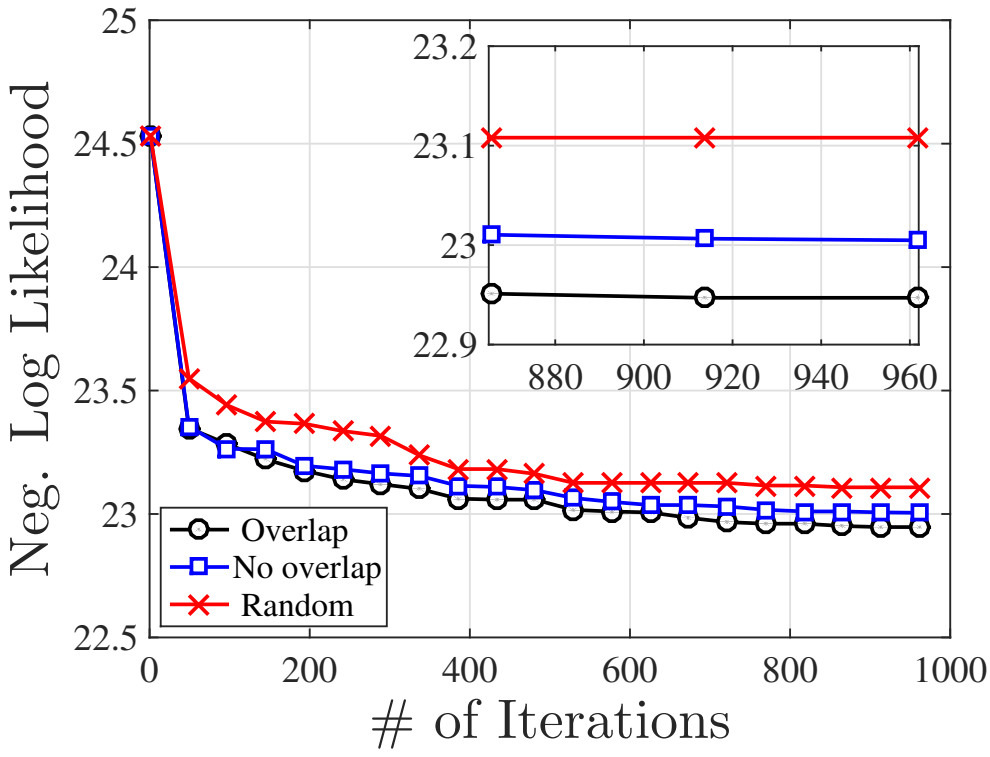

Figure 9: Results on the astrophysical experiment. The lower the vertical axis value, the more likely it is that the chosen constants match the observed data.