A survey on high-dimensional Gaussian process modeling with application to Bayesian optimization

Mickael Binois, Nathan Wycoff

A survey on high-dimensional Gaussian process modeling with application to Bayesian optimization

Mickaël Binois[*] Nathan Wycoff[†]

May 13, 2022

Abstract

Bayesian Optimization, the application of Bayesian function approximation to finding optima of expensive functions, has exploded in popularity in recent years. In particular, much attention has been paid to improving its efficiency on problems with many parameters to optimize. This attention has trickled down to the workhorse of high dimensional BO, high dimensional Gaussian process regression, which is also of independent interest. The great flexibility that the Gaussian process prior implies is a boon when modeling complicated, low dimensional surfaces but simply says too little when dimension grows too large. A variety of structural model assumptions have been tested to tame high dimensions, from variable selection and additive decomposition to low dimensional embeddings and beyond. Most of these approaches in turn require modifications of the acquisition function optimization strategy as well. Here we review the defining structural model assumptions and discuss the benefits and drawbacks of these approaches in practice.

Keywords: Black-box optimization, low-intrinsic dimensionality, additivity, variable selection, active subspace

1 Introduction

A large number of parameters (high dimensionality) is regularly mentioned as a critical challenge for black-box optimization. The root of the issue is the so-called curse of dimensionality, as coined by Bellman (1966): the exponential dependence of complexity on input

[*] Université Côte d'Azur, Inria, CNRS, LJAD, France Corresponding author: mickael.binois@inria.fr

[†] Current institution: Massive Data Institute, McCourt School of Public Policy, Georgetown University, Washington, DC.

dimension. This is all the more true with limited evaluation budgets, where relying on a surrogate (or metamodel) is a common practice. In BO, the number of variables impacts both the GP surrogate (via the behavior of distances in high dimension, see Aggarwal et al. (2001)), as well as the search for the next designs (as it affects the acquisition function, which is employed in BO in the search for the next design(s) to query). This dimensionality-scaling difficulty is mentioned in most existing reviews, see for instance Shan and Wang (2010), Viana et al. (2014), Shahriari et al. (2016b), Frazier (2018), Ginsbourger (2018) or Stork et al. (2020). This article is thus incremental, focusing on more recent trends over several communities (e.g., engineering, operations research, and machine learning) but is by no means exhaustive.

We focus on Gaussian process (GP) surrogates here, whose popularity originates from their modeling flexibility and appealing uncertainty quantification (UQ) properties. Other alternatives could be entertained, but probably not without giving up some amount of either flexibility or small sample efficiency. For examples, see the use of tree models like random forest regression (Hutter et al., 2011), tree-structured Parzen estimators (Bergstra et al., 2011), and Bayesian additive regression trees (Chipman et al., 2012), spline models with Bayesian adaptive splines (Francom and Sansó, 2019), or neural network surrogate models like Bayesian neural networks (Snoek et al., 2015) and deep GPs (Damianou and Lawrence, 2013) for optimization, e.g., as by Hebbal et al. (2019). Radial basis function interpolation (RBF), closely related to GPs, is also popular in this setting (Regis and Shoemaker, 2013), but is comparatively lacking in UQ capabilities.

In BO, the number of variables impacts both the GP surrogate (via the behavior of distances in high dimension, see Aggarwal et al. (2001)), as well as the search for the next designs with the acquisition function (by complicating its optimization). Specifically, the number of design points required to keep the same quality of approximation grows exponentially with the number of variables and the volume concentrates on the boundary of the search space. As a result, designs are comparatively dispersed in high dimension. We refer to Sommerville (1958); Artstein-Avidan et al. (2015) for more details on high dimensional geometry, and to Zhigljavsky and Žilinskas (2021) for a discussion more oriented to BO. Optimization of the acquisition function to select the next design point benefits from (comparatively) quick evaluation and gradient evaluation, but still suffers from these same high-dimensional effects. Depending on the difficulty of the optimization problem, these issues can occur when reaching ten variables, or manifest for dozens of them. Still, dedicated methods have been shown to work for billions of variables (Wang et al., 2013)– though under rather limiting assumptions.

Indeed, stronger structural model assumptions are needed as the dimension increases, with three main categories. One idea is to reduce the dimension by removing variables with little or no impact on the output. Another is to define a few new variables based on linear or non-linear combinations of the original ones. The last direction is to assume additivity of effects of variables, or groups of them. With more structure to learn, model inference becomes harder, adding estimation risk to the difficulties above. Indisputably, estimation risk can be reduced by avoiding estimation, and this has been the approach taken by researchers relying on randomly defined structures (though obviously this is at the cost of variance and bias possibly gained from the randomization procedure). If we want to infer the structure, however, we also must face the fact that in the sequential design framework, only limited data is at first available to estimate characteristics of the function. Either the structure can be learnt and updated "online" as the data are observed, or we can break the sequential process into two consecutive stages, the first of which is focused on estimating structure and the second on exploiting it. In this latter case, the optimal balance of budget to dedicate to each stage can be quite problem dependent.

The chosen structure affects the acquisition phase in several ways. It can be exploited for the modeling effort, as in reduced dimension or additivity assumptions, or it can constrain the optimization domain to aid the acquisition function search. Independently, strategies such as deploying trust regions as in Eriksson et al. (2019) are also available to limit the effect of the curse of dimensionality by limiting the volume of the search region. The resulting optimization approach is more local, but is complemented by restarts for globalization.

There exists many additional challenges and refinements for GPs and BO that are out of scope here, for which the corresponding techniques may need to be adapted for high dimension. One such issue is the scaling in terms of number of design points. Logically, more designs are needed to learn in larger dimensions, but one can not hope to match the exponential dependence on the dimension. Techniques for coping with large data have also attracted a lot of attention, see for instance Hensman et al. (2013); Heaton et al. (2019); Wang et al. (2019); Kleijnen and van Beers (2020), where some are independent of the input dimension as is the case for, say, local models. In general we will assume here that running the black-box remains limiting compared to running the BO framework. Other refinements for GP and BO to cope with complex noise modeling (e.g., non Gaussian noise, input dependent variance) or non-stationarity could be adapted, but high-dimension exacerbates the difficulty of the learning task. While we focus on unconstrained optimization, batch (or parallel) optimization, constrained optimization and others could be entertained as well. We refer the interested reader to Garnett (2022) for a broader and more introductory overview of BO.

The remainder of this paper is as follows. First, key notions and notations are introduced in Section 2. Next, structural assumptions for high dimensional GP modeling are detailed in Section 3, before consequences and adaptations for the acquisition function optimization are presented in Section 4. Section 5 includes a list of possible test functions. Finally, some practical guidelines and a summary on promising research directions is given in Section 6.

2 Background

Let us consider an expensive-to-evaluate *black-box* simulator $f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ that we want to globally optimize:

$$\text{find } \mathbf{x}^* \in \operatorname*{argmin}_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}). \tag{1}$$

By black-box, we mean that nothing is assumed known about the functional form of f: $f(\mathbf{x})$ can only be queried at any given input point $\mathbf{x}$ (sometimes gradients are also assumed available).

2.1 The Gaussian stochastic process

Given an index set $\mathcal{X}$, in our case typically a closed and bounded subset of $\mathbb{R}^d$, a *stochastic process* is simply a rule for assigning to B members of that set $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(B)}$ a joint distribution of random variables; maybe it has a density $\delta\left(y(\mathbf{x}^{(1)}), y(\mathbf{x}^{(2)}), \ldots, y(\mathbf{x}^{(B)})\right)$ with respect to some dominating measure. If the joint distribution is assumed to be Gaussian, we simply need a way of deciding what the mean vector and covariance matrix are going to be for any possible set of B points. These are naturally referred to as the *prior mean function* μ and *covariance function* k (also *kernel function*). Such a mathematical construct taken all together is called a *Gaussian process* (GP), see, e.g., Rasmussen and Williams (2006).

If we observe $y(\mathbf{x}^*)$ (possibly corrupted by Gaussian noise), we can imagine that there is some stochastic process that maps $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(B)}$ to $\delta\left(y(\mathbf{x}^{(1)}), y(\mathbf{x}^{(2)}), \ldots, y(\mathbf{x}^{(B)}) | y(\mathbf{x}^*)\right)$, that is, that maps any set of points to its conditional distribution given $y(\mathbf{x}^*)$. Happily, this new stochastic process also happens to be a GP (conjugacy with respect to the normal likelihood), and its new mean and covariance functions are available in closed form. In particular, if we observed $y(\mathbf{x}^{(1)}), \ldots, y(\mathbf{x}^{(n)})$, then:

$$m_n(\mathbf{x}) = \mu + \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1}(\mathbf{y} - \mu\mathbf{1}),$$
$$s_n^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1}\mathbf{k}(\mathbf{x})$$

where $\mathbf{y} := (y_1, \ldots, y_n)$, $\mathbf{k}(\mathbf{x}) := (k(\mathbf{x}, \mathbf{x}^{(i)}))_{1 \leq i \leq n}$, $\mathbf{K} := (k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) + \tau^2 \mathbf{1}_{i=j})_{1 \leq i,j \leq n}$. τ^2 is the noise hyperparameter, when assuming $y_i = f(\mathbf{x}^{(i)}) + \varepsilon_i$, with $\varepsilon_i = \mathcal{N}(0, \tau^2)$.

Often, the prior mean function μ is chosen to be zero (after centering observed y): the GP's focus is on defining and exploiting spatial covariance. However, it can definitely be useful to have different μ. If the objective function is known to be nonstationary, adding in polynomial trend terms can be a way to tame this. However, if the proper trend is not *a priori* known, it pays to be cautious when using one in high dimension as it's easy to accidentally extrapolate rather severely and perniciously (Journel, 1974). Sparse trend selection might be preferable, as performed for instance by Kersaudy et al. (2015) in a combination with polynomial chaos expansion, as may be using basis elements informed via cross-validation (Liang et al., 2014) or a Bayesian framework (Joseph et al., 2008).

The choice of covariance function family determines the qualitative properties of the GP. Of course, we want to make sure that when we're done applying the kernel function to all our data, the covariance matrix that comes out is positive semi-definite. Functions which guarantee this are themselves called positive semi-definite. Attention is often further restricted to *stationary* kernels, where the kernel function k is a function of $\mathbf{x}-\mathbf{x}'$, $k(\mathbf{x},\mathbf{x}') = \tilde{k}(\mathbf{x}-\mathbf{x}')$. Within this class, the most popular kernel is probably the squared exponential one in product form, $k(\mathbf{x},\mathbf{x}') = \sigma^2 \prod_{i=1}^{d} \exp\left(-\frac{(x_i - x_i')^2}{2\theta_i}\right)$. This kernel function is defined only up to its free parameters $\theta_1, \dots, \theta_d$, the lengthscales which determines at what distance covariance begins to drop off between points, and σ^2 which converts correlation to covariance based on the scale of y. This kernel induces an infinitely differentiable metamodel, which may or may not be desirable, depending on how well that matches our understanding of the black-box. For precise control over differentiability[1], consider the Matérn class indexed by smoothness parameter ν, which is defined in terms of the modified Bessel function of the second kind (though can be written in terms of simple functions when $\nu + 0.5$ is a whole number). See Rasmussen and Williams (2006, Chapter 4) for more on the Matérn and other covariance function classes. Such GPs can be differentiated at most $\nu - 0.5$ times.

This dependence of GPs on distances between points will be problematic in high dimension. And, consequently, many of the approaches we review in this article will define a kernel function not on the original input space, but on some transformation thereof. Of course, it is possible to view the transformation and the kernel function thereon together as a novel kernel function. But we will instead take the perspective in this article of separating dimension-reducing transformations and the subsequent purely distance-based kernel, as we view this modularity as helpful in model design and comparison.

Insofar as a Gaussian process is the infinite dimensional generalization of the multivariate Gaussian distribution, it retains many of the analytical properties that have propelled Gaussianity to its prominence. Of particular use to us will be the closed form evaluation

[1] For more details on what exactly is meant by the derivative of a stochastic process see (Papoulis, 1965, Chapter 10).

of certain exceedance probabilities and related metrics which form the basis of deploying these Gaussian processes to optimization problems.

We refer to Rasmussen and Williams (2006); Gramacy (2020) for additional details on GP regression, and, e.g., to Kanagawa et al. (2018) for a discussion of connections to other kernel methods.

2.2 Bayesian optimization

Popularized with the Efficient Global Optimization method of Jones et al. (1998), Bayesian optimization, see, e.g., Mockus et al. (1978), relies on the GP probabilistic prior of f based on a few initial observations to define an acquisition function $\alpha : \mathcal{X} \to \mathbb{R}$. This is later optimized to select new points to evaluate sequentially. The initial stage, known as the Design of Experiments (DoE), typically starts by evaluating designs based on optimized Latin hypercube samples, for their space filling properties, see, e.g., Gramacy (2020). A pseudo-code is given in Algorithm 1.

Algorithm 1 Pseudo-code for BO

```
Require: $N_{max}$ (total budget), GP model trained on initial DoE of size $n = n_0$ : $(\mathbf{x}^{(i)}, y_i)_{1 \leq i \leq n}$
while $n \leq N_{max}$ do
    Choose $\mathbf{x}^{(n+1)} \in \arg\max_{\mathbf{x} \in \mathcal{X}} \alpha(\mathbf{x})$
    Update the GP model by conditioning on $(\mathbf{x}^{(n+1)}, y_{n+1})$
    $n \leftarrow n + 1$
end while
```

The expected improvement (Mockus et al., 1978, EI) and upper confidence bound (Srinivas et al., 2010, UCB) criteria are typically used due to their analytical tractability in both value and gradient. Other choices are possible, see e.g., Shahriari et al. (2016b). The generic acquisition function optimization problem is:

$$\text{find } \mathbf{x}^* \in \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}}\, \alpha(\mathbf{x}). \tag{2}$$

Some acquisition functions are natively suitable for when the black-box is noisy, while some need adaptation like EI, see e.g., Letham et al. (2018) for a discussion. A pinch of noise might be beneficial for its regularization effect, see, e.g., Gramacy and Lee (2012), but the effort of finding a solution increases drastically as the signal to noise ratio sinks. Not to mention input-varying and non-Gaussian noises that are hard to deal with even in low dimension. We refer to Forrester et al. (2008); Roustant et al. (2012); Gramacy (2020); Garnett (2022) for more details on how to apply BO in practice.

2.3 The curse of dimensionality and its effects

An increase in input dimension is first felt in distance computation, which is at the core of most common covariance kernels. The difficulty is that points are relatively further away in high dimension, such that it becomes harder to learn with a covariance based on distances. As is illustrated in Fig. 1, for large d, the squared distance between uniformly sampled points concentrates increasingly further away. In particular, the expected squared distance approaches to $\frac{d}{6}$ with standard deviation relative to the width of the interval of possible distances $[0, \sqrt{d}]$ decreasing as $\sqrt{\frac{7}{180d}}$ [2]. See Köppen (2000) for a discussion of other implications and properties of high dimensional cubes.

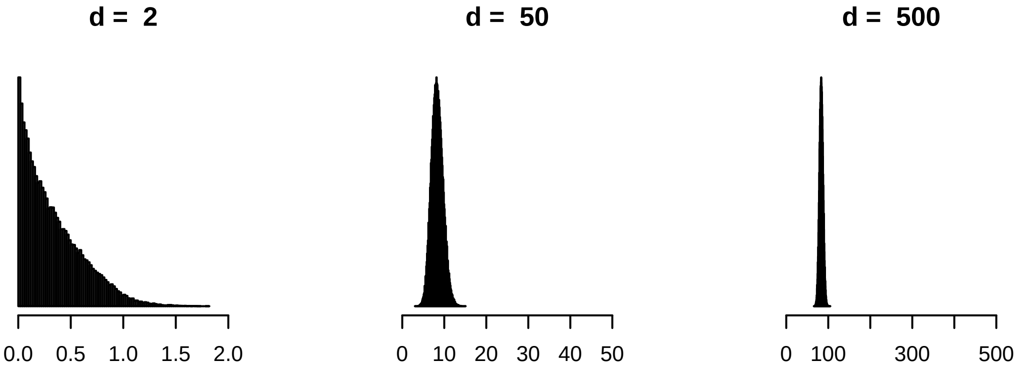

Figure 1: **Distances Concentrate in High Dimension:** Randomly sampling 10,000 points according to the uniform measure in $[0, 1]^d$ and then calculating squared inter-point distances reveals that these concentrate within the bounds of possible values in high dimension: $[0, d]$.

Except for isotropic kernels, the second effect is the increase of the number of hyperparameters of the models, typically with d many lengthscale parameters to tune for product

[2] The results follow from the fact that the density of the squared distance d between two uniform variables on $[0, 1]$ is given by $\frac{1}{\sqrt{d}} - 1$.

covariance kernels. Maximizing the (non-convex) likelihood thus becomes increasingly hard, since the curse of dimensionality also affects the gradient based optimization routines typically used. Grouping variables to share the same lengthscales alleviate this, which has been done either manually based on problem-specific considerations as done by Binois (2015) or using the Bayesian information criterion (BIC) as done by Blanchet-Scalliet et al. (2017).

Similarly, accurately optimizing the acquisition function is complex in high dimension, especially since it is a multi-modal optimization task, with flat regions. Furthermore, given that most of the volume is on the boundary of the domain and that GP-based predictive variance increases with the distance to the designs – a desirable property at least in low dimension – the downside is that the optimum is generally found on one of the exponentially many vertices or sides of the d-hypercube, tilting the balance towards blind exploration.

The curse of dimensionality also manifests itself when random sampling, required for Monte Carlo estimation of certain integral quantities, among them entropy, the optimizing point of a posterior realization (as is required in Thompson sampling), and certain global sensitivity measures. The issue is that a uniformly sampled point will be increasingly distant, on average, to any given point as dimension increases. Hence the challenge is not only to scale in terms of modeling accuracy but also to keep inference manageable and avoid pitfalls in acquisition function optimization.

3 High-dimensional Gaussian process modeling

Structural model assumptions are the only way to avoid the exponential dependence on the dimension in modeling, for which various options are summarized for instance by Bach (2017), who also gives the corresponding generalization bounds. These different structural model assumptions have been adapted in the GP context, plus others, as we detail next. One exception is to use isotropic kernels and thus always a single lengthscale, for which increasing the dimension influences only the behavior of distances.

3.1 Variable selection or screening

Whenever possible, limiting the number of variables based on expert knowledge on the optimization problem is recommended. If expert knowledge is unavailable, a simple, data-driven idea is to perform variable selection, or screening, before optimization, e.g., using the Morris technique (Morris, 1991) or via hierarchical diagonal sampling as in Chen et al. (2012). Other global sensitivity analysis techniques can be used to select variables individually and we refer to Iooss and Lemaître (2015) for an entry point to this topic.

Hence, one of the early attempts to tackle high-dimensions is to assume that most of the variables have no effect:

$$\text{model: } f(\mathbf{x}) \approx g(\mathbf{x}_I) \text{ with } I \subset \{1, \ldots, d\}, |I| \ll d \tag{3}$$

and then identify those influential variables in the set I. For the Gaussian and other stationary kernels in tensor product form, this can be performed by looking at the lengthscale values: the covariance varies less for very large values of the lengthscale (when parameterized as above), whose corresponding input variables can thus be removed. This is also known under the term automatic relevance determination (ARD), as discussed, e.g., by Rasmussen and Williams (2006, Chapter 5). Salem et al. (2018) shows that this indeed holds asymptotically. Hence the idea is to rank the variables based on their lengthscales to determine the more influential from the less. These influential variables are then used to build a GP model and optimize expected improvement.

Nevertheless, $\theta_1 > \theta_2$ does not necessarily imply that x_1 is less important than x_2, see for example Wycoff et al. (2021); Lin and Joseph (2020). Hence rather than looking only only at the lengthscales values, Linkletter et al. (2006) propose to compare the posterior distributions of lengthscales corresponding to both real and artificially added inert variables in a fully Bayesian GP framework. This is complemented by a local variable selection in Winkel et al. (2021), relying of predictions made when ignoring some dimensions. Eriksson and Jankowiak (2021) choose instead to put a Horseshoe prior (Carvalho et al., 2009) on the inverse lengthscales and perform gradient-based numerical posterior simulation in order to sample from this posterior of functions defined on subsets of variables. With a more incremental approach, Marrel et al. (2008) propose to sequentially add variables to the regression (resp. covariance) element based on the Akaike information criterion (resp. predictivity coefficient Q_2).

Rather than definitively deciding on which variables to keep, which can be time-consuming, Ulmasov et al. (2016) propose to sample a few variables at each iteration, where the weight vector is determined with principal component analysis (PCA) on the $(\mathbf{x})_{1 \leq i \leq n}$. This may not be efficient for small budget as indicated in Li et al. (2017), who prefer to select variables uniformly and to fill in values for the remaining variables.

In practice, most variables typically have a possibly limited but non-null influence on the output. Though this can mean that choosing the number of variables to keep is somewhat arbitrary, the simplicity, both in terms of implementation and interpretation, of variable selection makes it a compelling approach to dimension reduction, particularly for a first pass. Of course, the variable selection itself becomes only harder as the dimension increases, but this is to a lesser extent than some of the more sophisticated approaches discussed below. Nevertheless, the approach's strength is also its limitation, and it is common for the complicated functions which make up black-box problems to truly and strongly depend on all input parameters. In the case where all variables have the same influence, variable selection would even fail to reduce the dimension by one no matter how simple the relationship. Other structural assumptions can overcome these limitations, as well as exploit interactions between variables.

3.2 Additive and ANOVA models

One set of structural assumptions amenable to keeping all variables but limiting their interaction is that of additivity:

$$\text{model: } f(\mathbf{x}) \approx \mu + \sum_{i=1}^{d} g_i(x_i) \tag{4}$$

with univariate functions g_i. This has been transposed to the GP framework by Neal (1997); Plate (1999); Durrande et al. (2012); Duvenaud et al. (2011), initially via the summation of univariate kernels: $k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{d} \sigma_i^2 k_i(x_i, x_i')$, which results in a valid covariance function just as the product form. A useful property for interpretability and visualization is that the GP predictive mean can be decomposed into a sum of univariate components: $m_n(\mathbf{x}) = \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1}\mathbf{y} = \sum_{i=1}^{d} \mathbf{k}_i(x_i)\mathbf{K}^{-1}\mathbf{y} = \sum_{i=1}^{d} m_{n,i}(x_i)$ with $\mathbf{k}_i(x_i) := (k_i(x_i, x_i^{(j)}))_{1 \leq j \leq n}$. A more surprising property is that the covariance can become non invertible due to linear relationships appearing between distinct observations – see e.g., Durrande et al. (2012) for examples–which can be alleviated by adding a noise term. Consequently, the predictive variance can be zero at an unobserved design point, a somewhat detrimental side effect for exploration and thus optimization. Another difficulty is the estimation of the hyperparameters, with the need of an additional variance parameter σ_i per coordinate (thus approximately doubling the number of hyperparameters). Still, compared to the tensor product form whose values quickly go to zero in high dimension, the sum form scales much better. Making the black-box more additive by applying a transformation of the outputs is explored by Lin and Joseph (2020).

Higher-order models can be defined in the same way (Duvenaud et al., 2011), usually restricted to order two or selecting higher order components only if all lower orders are already selected. Plate (1999) instead proposes adding all interactions at once on top of first order ones. Directly identifying groups of variables is also possible, see e.g., Kandasamy et al. (2015); Gardner et al. (2017); Wang et al. (2017, 2018):

$$\text{model: } f(\mathbf{x}) \approx \mu + \sum_{i=1}^{M} g_i(\mathbf{x}_{A_i}) \tag{5}$$

with multivariate g_i functions acting on A_i disjoint subsets of $\{1, \dots, d\}$ such that $\bigcup_i^M A_i = \{1, \dots, d\}$. The restriction of disjoint subsets of variables have been further lifted in subsequent works, see e.g., Rolland et al. (2018); Hoang et al. (2018).

Similar in their form but rooted in global sensitivity analysis are additive models based on the functional analysis of variance (fANOVA, a.k.a. Sobol-Hoeffding) decomposition

(Efron and Stein, 1981; Sobol, 2001):

$$\text{model: } f(\mathbf{x}) \approx c + \sum_{i=1}^{d} g_i(x_i) + \sum_{j<k} g_{jk}(x_j, x_k) + \cdots + g_{12\ldots d}(x_1, x_2, \ldots, x_d) \tag{6}$$

with elementary functions $g_{\ldots}$ that are required to be centered and orthogonal for the uniqueness of the decomposition. The benefit is that sensitivity analysis can be conducted and interpreted as in a regular ANOVA. Muehlenstaedt et al. (2012) rely on this decomposition up to second order interaction to build their model, choosing their components on the basis of Sobol indices that appear in this formulation (and estimated by a first-pass GP with an anisotropic tensor-product kernel). These are the so-called main effects (one variable only) and total interaction effects (effect of two variables at any order). Selecting which interaction to remove to form cliques requires a thresholding scheme. Ulaganathan et al. (2016) proposes a similar approach with the addition of cut-points when gradient observations are available. Durrande et al. (2013) go further with a dedicated kernel whose form is $k_{\text{ANOVA}}(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{D}(1 + k^i(x_i, x'_i))$, as introduced by Stitson et al. (1999). There the sensitivity indices of the ANOVA representation are analytically tractable. In Ginsbourger et al. (2016), the ANOVA decomposition is performed directly on the kernel and propagated to the corresponding random field under appropriate orthogonality conditions. Looking for sparsity to avoid estimating the entirety of the 2^d components, they define projectors that allow separation of the additive components (with cross-covariance between main effects) from their complement.

The main downside of these techniques is that inference is challenging in this context, with combinatorially many terms. Hence various techniques to estimate the hyperparameters have been applied: coordinate-ascent-like (Durrande et al., 2012) or quasi-Newton (Duvenaud et al., 2011) methods for likelihood maximization. Relying on randomness is sometimes preferred to bypass the cost of a full optimization, like by Kandasamy et al. (2015) where random decompositions are sampled and the best one for the likelihood is selected (fixing the order and number of terms), or by Wang et al. (2018). Gardner et al. (2017) attempts to elicit the structure via a dedicated Metropolis-Hastings algorithm. For overlapping subsets, Hoang et al. (2018) also rely on random groups while Rolland et al. (2018) use a dependence graph and Gibbs sampling to perform inference. In this vein of methods, there is Delbridge et al. (2020) that reconstruct the high dimensional kernel by a sum of univariate ones in random directions, in the spirit of the turning band method from geostatistics (Journel, 1974).

In terms of advantages, on the other hand, this method maintains the interpretability of variable selection, particularly if the selected model includes mostly first or second order terms and is somewhat sparse. Learning which variables interact by observing which pairs are selected by an inference-driven procedure can be scientifically interesting in and of itself.

Some works are dedicated to scaling to many observations as well, such as Mutny and Krause (2018) with a basis expansion of GP kernels. Sung et al. (2019) also propose a basis expansion, further complementing it with a multi-resolution scheme with a group lasso estimation procedure. Wang et al. (2018) scale to many observations with a random partition of the input space, random additive approximation and random features decomposition of the kernel.

Here, one underlying hypothesis is that high order interaction components are negligible (because they are hard to estimate in this formalism), though this is a long-standing tradition in experimental design. Another concern is that ANOVA cannot detect non-linearity and multi-modality (Palar and Shimoyama, 2017). These downsides are mitigated by the following framework.

3.3 Linear embeddings

One approach to designing high dimensional kernels is to avoid doing so directly, and instead run the data through a dimension reducing map first. The image of these data under this map is called its embedding, and an obvious class of functions to use for dimension reduction is the class of linear functions. Indeed, as noted by Marcy (2018), this idea of using a linear mapping dates back to at least Matérn (1960). The question is then simply which linear function to use in particular.

Denote this generic mapping by $\mathbf{z} = \mathbf{A}\mathbf{x}$ with $\mathbf{A} \in \mathbb{R}^{r \times d}$. In the case of Vivarelli and

Williams (1999), $r = d$, so the mapping serves to rotate the space rather than actually reduce the dimension, but we are more interested in the $r \ll d$ case. When $r = 1$, this is a popular dimension reduction technique called the single index model:

$$\text{model: } f(\mathbf{x}) \approx g(\mathbf{a}^\top \mathbf{x}) \text{ with } \mathbf{a} \in \mathbb{R}^d. \tag{7}$$

We refer to Gramacy and Lian (2012) for the GP treatment, with $\mathbf{a}$ treated as an additional kernel hyperparameter. $\mathbf{a}$ could instead simply be randomly chosen at each iteration, as suggested by Kirschner et al. (2019). Unfortunately, extending to $r > 1$ is by no means trivial. The assumption is:

$$\text{model: } f(\mathbf{x}) \approx g(\mathbf{A}^\top \mathbf{x}) \tag{8}$$

where such functions are called *ridge functions* in the literature. It corresponds to the observation, sometimes backed by theoretical evidence (Constantine et al., 2016), that the variation of high dimensional functions can be concentrated around a few but unknown directions. There are several approaches to choosing $\mathbf{A}$, which we'll review in turn.

For GP regression, arguably the most direct way is simply to treat $\mathbf{A}$ as just another kernel hyperparameter to learn, e.g., by marginal likelihood optimization. Fixing r, Garnett et al. (2014) provide an approximately Bayesian scheme to do so, with a Laplace approximation to the likelihood followed by an approximate marginalization over hyperparameters. Tripathy et al. (2016) rely on a two stage approach: first using the likelihood to learn an orthogonal $\mathbf{A}$, then finding the rest of the hyperparameters (and repeat). BIC is used to identify r. The orthogonality constraint requires some special consideration, since these matrices lie on the Stiefel manifold. Seshadri et al. (2019) reinterpreted a similar problem, drawing connections between regression and approximation. Also building on Tripathy et al. (2016), Yenicelik (2020) observed that the likelihood may not always select the best matrix among alternatives. A full Bayesian treatment was proposed by Marcy (2018), with priors on matrix manifolds and the dimension, requiring advanced Monte Carlo methods.

Running a sensitivity analysis independent of the GP model to select $\mathbf{A}$ as a kind of preprocessing step is also an option. When gradients of the black-box are available, recovering the matrix $\mathbf{A}$ is relatively straightforward. It corresponds, up to a rotation, to the eigenvectors with non-zero eigenvalues of the matrix $\mathbf{C} := \int_{\mathcal{X}} \nabla(f(\mathbf{x}))^{\top}\nabla(f(\mathbf{x}))\lambda(d\mathbf{x})$ where λ is any well-behaved measure on the design space (typically the Lebesgue one on hypercubic domains). Then a Monte Carlo estimator can be used. Loosening the assumption that some eigenvalues will be exactly zero leads to the active subspace (AS) methodology (Constantine, 2015), analyzed for dimension reduction and visualization. In particular, the presence of an AS is hinted at by gaps in the eigenvalues of $\mathbf{C}$ (Constantine, 2015). If the gradient can be evaluated, this is a good pre-processing step before GP modeling (Eriksson et al., 2018). Without the gradient information, finite differencing is generally too costly, such that using a GP to estimate $\mathbf{C}$ might be more appropriate, e.g., as in Fukumizu and Leng (2014); Palar and Shimoyama (2017). Djolonga et al. (2013) uses directional derivatives with finite difference to recover $\mathbf{A}$ with low rank matrix recovery before BO. Other compressed sensing techniques can be used, as done initially by Carpentier and Munos (2012) and later by Groves and Pyzer-Knapp (2018). The MC estimators of $\mathbf{A}$ are applied in a two stages approach, due to the assumption of iid sampling of the design points. Wycoff et al. (2021) showed that an estimate of the $\mathbf{C}$ matrix of a GP is tractable for standard stationary kernels, relieving the need for this sampling assumption. Other works in the vein of sensitivity analysis on slice inverse regression or partial least squares to recover $\mathbf{A}$ and possibly update it during optimization (Bouhlel et al., 2016; Zhang et al., 2019; Chen et al., 2020a). Lee (2019) advocate for a modified AS matrix which exaggerates the influence of the average gradient (over the input space).

And finally, rather than learning $\mathbf{A}$ as part of inference on GP hyperparameters or as the result of a sensitivity analysis, a third option is to simply select $\mathbf{A}$ randomly, either by generating a single $\mathbf{A}$ matrix before the modeling step or using different $\mathbf{A}$s, say, at each iteration. For example, in random embedding BO (REMBO), Wang et al. (2013, 2016) use a fixed and randomly sampled $\mathbf{A}$. One justification is the stability of random projection of the L_2 norm due to the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984), as noticed, e.g., by Letham et al. (2020). In the context of BO, the rationale is further that, at least for an unbounded domain, there exists a solution to the problem on the low dimensional embedding, explaining the success of random optimization (Bergstra and Bengio, 2012) on some hyperparameter tuning problems.

Even with a candidate $\mathbf{A}$ in hand, whether through optimization, sensitivity analysis, or random sampling, we still have modeling decisions to make (not to mention acquisition decisions; Section 4.3). The seemingly innocuous choice of a hypercubic domain can give headaches when combined with a linear dimension reduction, as the possible solution space is no longer simply a cube but a polytope defined by $\mathbf{A}$. If our GP is fit to the low dimensional space, it may have an easier time finding a next candidate point to optimize, but we have to figure out which point in the original space corresponds to this low-dimensional optimum. This means that not only are preimages non-unique, but the backprojection of a point is not even guaranteed to live in the original unit hypercube, and a convex projection is possibly required to regain feasibility. This means that our backprojection is no longer linear. The remediation options proposed in Wang et al. (2013) are to ignore the corresponding non-injectivity issue, or to use a kernel defined on $\mathcal{X}$, losing the benefits of low dimensional GP modeling. Binois et al. (2015) proposed to include high-dimensional information in $\mathcal{Y}$ with a warping to address the non-injectivity issue, while Binois et al. (2020) defines an alternative mapping from $\mathcal{Y}$ to $\mathcal{X}$ to avoid it. Nayebi et al. (2019) also bypasses this problem by choosing a sparse random matrix with $\{-1, 0, 1\}$ elements only, in essence selecting from embeddings on diagonals of the hypercube. These ideas can be extended to affine embeddings, as proposed for global optimization by Cartis et al. (2020); Cartis and Otemissov (2021).

Then comes the question of how to use high dimensional data that are not on the linear embedding (unlike for REMBO). Unless $\mathbf{A}$ is perfectly recovered, introducing a noise term is necessary to account for the discrepancy. To this end, Moriconi et al. (2020) uses GP regression on quantiles for axis-aligned projections. Another question is how to select the lengthscales in this context since the product kernels are not preserved in the embedded space, for which Letham et al. (2020) show that a specific parameterization (Mahalanobis distance based kernel) is preferable and handles distortions. Still, inference is complicated whereas the REMBO program needs only to fit a low dimensional GP. What distribution to sample $\mathbf{A}$ from has also been briefly investigated in Binois (2015); Binois et al. (2020); Letham et al. (2020).

Though in general any GP can be fit in the reduced space, links to additive assumptions are natural, and Gilboa et al. (2013) combine these in the projection-pursuit style, leading to projected additive approximations. In this context, tuning the r parameter is performed by adding dimensions iteratively and halting when the accuracy stops increasing. Li et al. (2016) further discuss how to extend this approach to optimization, in a restricted-projection pursuit setup. This can be viewed as somewhere in between fitting a full low dimensional model and using first order additive model. These ideas can be combined by estimating several low-dimensional subspaces, see e.g., Yenicelik (2020); Wong et al. (2020).

The estimation risk is strong when using a linear embedding (which has $p \times r$ parameters in general), with unresolved questions: i) How well must the matrix $\mathbf{A}$ be learned before it is better than directly fitting the high dimensional problem? ii) How much of the budget should be dedicated to this task? iii) When is it better to learn $\mathbf{A}$ dynamically, leading to a noisy GP (whenever design points are not all on the same embedding as for REMBO)? Plus finding an appropriate r remains difficult unless expert information is given. Fortunately, taking larger values than necessary is not detrimental –except that it quickly reintroduces high-dimensional challenges (Wang et al., 2013; Cartis and Otemissov, 2021).

Already, the linear embeddings approach jettison much of the interpretability that selection and ANOVA approaches offer (it will get worse when we consider nonlinear embeddings in the next section). In the case when $r \in \{1, 2\}$, it is possible to create visualizations of the function, which can be great sources of intuition. But in larger dimensional cases, we are reduced to squinting at the loadings of each variable in the retained directions. On the other hand, there are indeed functions which can be reduced to much smaller dimension when using linear embeddings compared to the axis-aligned variable selection [3]. Which of these two approaches will be more appropriate is quite problem-dependent.

[3] A toy example is $\sin(\frac{4\pi}{d}\mathbf{1}^\top \mathbf{x})$.

3.4 Non-linear embeddings and structured spaces

Dropping the linearity assumption adds even more flexibility to the model at the cost of requiring even more data to fit it. Recovering such a manifold suitable for regression has been proposed e.g., by Guhaniyogi and Dunson (2016), when the data is on a low dimensional manifold. A simple way to extend linear dimension reduction to the nonlinear case is to use a locally linear approach, as does Wycoff (2021b) in the context of AS. Also defined by analogy to AS is Bridges et al. (2019)'s 1D active manifold, not yet applied to GPs. In this line of thought, it is also possible to include generative topographic mapping (Viswanath et al., 2011), GP latent variable models, e.g., (Lawrence, 2005; Titsias and Lawrence, 2010), or deep GPs (Damianou and Lawrence, 2013; Hebbal et al., 2019; Sauer et al., 2020). Deep Gaussian processes refer to the modeling strategy of assuming that the input is linked to the output via a chain of Gaussian processes, the output of one serving as the input for the next. This is not to be confused with the concept of deep *kernels* (Wilson et al., 2016; Huang et al., 2015), which involves only a single Gaussian process, but whose kernel function is parameterized by a neural network. Typically, inference is conducted on the GP and neural net weights via optimization of (an estimate of) the log marginal likelihood, but in the event of structured inputs or partial labeling the neural network can be initialized as an autoencoder (i.e., initialized to reconstruct the input). An orthogonal direction to avoid the larger optimization budgets is via multi-fidelity, when cheaper but less accurate version(s) of the black-box are available, as exploited e.g., by Ginsbourger et al. (2012); Falkner et al. (2018).

A setup where these highly flexible models may be more amenable is in the case where some additional information is available about the structure of the problem. One such case is the optimization of geometric shapes, typically airfoils. AS is performed in this context by Lukaczyk et al. (2014). As many options to parameterize these shapes are available, some are more adapted to optimization. Independently of this choice, Gaudrie et al. (2020) propose to work in the shape space, defined with shape eigenvectors and their values, which is not costly since the computation of the geometry is in general negligible next to the subsequent simulation. A non-linear computation of the shape basis is used in

Chen et al. (2020b), with generative adversarial networks used to learn a manifold from real data.

A related approach uses variational autoencoders (Kingma and Welling, 2013) as a deep kernel (Gómez-Bombarelli et al., 2018), which ensures that the latent state upon which the Gaussian process applies its kernel can approximately recover the original, unencoded input. This approach is popular when optimizing over structured non-Euclidean inputs, such as molecules (Gómez-Bombarelli et al., 2018; Tripp et al., 2020; Maus et al., 2022; Deshwal and Doppa, 2021; Grosnit et al., 2021; Eismann et al., 2018), for which standard BO is unavailable. Additional examples on functional data indexed on graphs include Espinasse et al. (2014).

Jaquier and Rozo (2020) perform geometry-aware BO on Riemannian manifolds for robotics, transposing some of the ideas from Section 3.3 to non-Euclidean spaces. A somewhat related assumption by Oh et al. (2018) is to use cylindrical coordinates instead of the original Cartesian ones (and rather than polar coordinates as Padonou and Roustant (2016), that do not scale well). The underlying assumption is that the solution is close to the center of the domain, if suitably chosen. Transforming the coordinates amounts to separating radial and angular components:

$$T(\mathbf{x}) = \begin{cases} (||\mathbf{x}||_2, \mathbf{x}/||\mathbf{x}||_2) & \text{for } ||\mathbf{x}||_2 \neq 0 \\ (0, \mathbf{a}_{arbitrary}) & \text{for } ||\mathbf{x}||_2 = 0 \end{cases},$$

while the inverse transformation is $T^{-1}(r, \mathbf{a}) = r\mathbf{a}$. The corresponding covariance kernel is $k_{cyl}(\mathbf{x}, \mathbf{x}') = k_r(r, r') \times k_{\mathbf{a}}(\mathbf{a}, \mathbf{a}')$. The 1D radius component $k_r(r, r') = k(|(1 - r^\alpha)^\beta - (1 - r'^\alpha)^\beta|)$, $\alpha, \beta > 0$ is chosen to further focus on the center (Oh et al., 2018) while the angular component is a continuous radial kernel $K_{\mathbf{a}}(\mathbf{a}, \mathbf{a}') = \sum_{p=0}^{P} c_p(\mathbf{a}^\top \mathbf{a}')^p$, $c_p > 0$, $\forall p$ (Jayasumana et al., 2014) with a user defined P.

The nonlinear approaches have seen great success in high dimension, particularly when the space is structured/non-Euclidean. However, this comes at the cost of added complexity, both in computation (e.g., when using deep neural networks) and interpretation. Therefore, our recommendation is to reserve these techniques for the toughest problems, and consider simpler solutions to simpler ones, particularly if insight into function, and not only an optimal solution, is desired.

There are many hoops and hurdles to construct a satisfying GP model in high dimension. Yet, this only gets one half the way, since it remains to optimize the acquisition function.

4 High-dimensional acquisition function optimization

A proposed high-dimensional GP model is usually combined with a strategy for acquisition function optimization, adapted to the specific model structure and the curse of dimensionality. Many of these approaches can be extricated from the modeling framework in which they were proposed, leaving many possible combinations unexplored. We now discuss them on their own merits.

Faster evaluation time and the availability of gradients provide limited help towards globally optimizing the multi-modal acquisition function. Global optimality guarantees are out of reach, since branch-and-bound methods or DIRECT do not scale (Jones et al., 1998). Aside from local minima, another pitfall is the presence of large plateaus where the acquisition function value is flat, while local optima may be peaked. To decrease these two effects, Rana et al. (2017) suggest to artificially inflate the gradient by taking a larger lengthscale within an isotropic GP. Then the lengthscale is successively decreased while tracking subsequent optima (warm started by previous values). Applying the random linear embedding of REMBO (Section 3.3) at the level of acquisition function optimization, Tran-The et al. (2019) optimize over a finite set of subspaces.

Regardless of the quality of the fitted GP model, the expanding boundary volume as a function of the number of variables has a large impact on the acquisition's optimization (Equation 2). If based on uniform sampling, e.g., with multi-start gradient optimization or even evolutionary algorithms, the search will focus in these boundary areas. This is reinforced by the larger variance on the boundary that would locate the acquisition function's optimum there, with no hope of evaluating all the faces or vertices before focusing on the interior. Commenting on and discussing these effects, Oh et al. (2018) proposed the use of cylindrical coordinates with BOCK to upweight the volume of the interior, imparting the prior knowledge that the optimum is close to the center. Adding virtual derivative observations on the boundary like in Siivola et al. (2018) is hardly feasible in high dimension, as derivative information is only analytically enforceable at finitely many collocation points in GPs (and indeed computation scales cubically in the same). An infinite version could possibly be entertained via spectral methods, e.g., based on Gauthier and Bay (2012). In such a case, the use of a trust-region (TR) can limit the size of the search space drastically. Trust region methods focus the optimization locally within the neighborhood of a current best solution (TR center), whose size increases if the new candidate point improves sufficiently over the TR center, or decreases instead, see e.g., Larson et al. (2019) for a review. Combined with BO, this has been shown to be quite beneficial in high-dimension (Regis, 2016; Eriksson et al., 2019; Diouane et al., 2021; Zhou et al., 2021; Daulton et al., 2021), perhaps at the cost of a less global search (possibly compensated for with restarts or parallel TR). To further avoid the attraction of the boundary of the TR, (2) is only optimized over a discrete set in Eriksson et al. (2018), with some coordinates randomly kept at the TR center.

While the strategies above are mostly independent of the GP model (e.g., used with isotropic or anisotropic product kernels), we next detail strategies for GPs with structural assumptions. Note that these assumptions, never perfectly fulfilled or estimated in practice, require the introduction of a noise component to account for the introduced approximations, with a few exceptions. Hence this has to be taken into account in the optimization of (2).

4.1 Additive case

Additivity provides the opportunity to rely on the same decomposition as the GP for solving (2), hence reducing the search to several lower-dimensional searches, possibly in parallel. That is, rather than focusing on the posterior of f, the idea is to look at those of the additive components instead, the g_i from models (4, 5, 6): $\mathcal{N}(m_{n,i}(\mathbf{x}_I), s^2_{n,i}(\mathbf{x}_I))$ where $s^2_{n,i}(\mathbf{x}_I) = k_i(\mathbf{x}_I, \mathbf{x}_I) - \mathbf{k}_i(\mathbf{x}_i)^\top \mathbf{K}^{-1} \mathbf{k}_i(\mathbf{x}_i)$ for a general index I. Then partial acquisition functions are defined on each g_i model, as a sum, e.g., as in Kandasamy et al. (2015). More care is needed for overlapping subsets, say relying on message passing (Rolland et al., 2018;

Hoang et al., 2018) but it remains more efficient than optimizing in the original space. The major advantage of this approach is that it allows for solving d many one dimensional optimization problems instead of one d dimensional problem, which *ceteris paribus* is many times easier in the nonconvex case. Nevertheless, the extent to which the zero variance at unobserved locations with additive models affects the globality of the search remains unknown. Adding a kernel a component that helps mitigate incorrect assumptions as for variable selection or AS as below may be interesting. Additive GPs should be considered for problems where the high dimensionality challenges the traditional acquisition function optimization and the budget is small enough that we cannot expect to learn a high-fidelity representation of the function.

4.2 Variable selection

If variables are completely removed beforehand in a preliminary stage, then a fixed value may be used, returning to a low dimensional problem. Observations with other values are generally discarded to keep the problem deterministic, the alternative being to add some noise if the inert variables can vary. Otherwise, after optimizing (2) on few variables, values for the screened variables must be determined to evaluate f. Alternatives include fixing these to a constant value, taking the values at the best design sampled so far for these coordinates, a random sampling, or a combination of these, see e.g., Li et al. (2017); Spagnol et al. (2019). In Salem et al. (2018), alternative lengthscales are estimated for the remaining variables by finding the most different values still passing a likelihood ratio test. Then, values for these variables are selected where the difference between the predictive means are the most different between the two sets of hyperparameters, to challenge the initial split.

When considering the acquisition function, the main advantage of variable selection approaches is to reduce the dimension of the search space, which lessens the burden of the acquisition optimization. On the other hand, the problem doesn't decompose, as in the additive case, and may be using suboptimal values for the “inactive” variables.

4.3 Embedding case

With embeddings, additional difficulties arise with bounded domains. Let $\mathbf{W} = [\mathbf{A}\ \mathbf{W}_2]$ be a basis of $\mathbb{R}^d$. Splitting between active and inactive (or less active) variables: $\forall \mathbf{x} \in \mathbb{R}^d$, $\mathbf{x} = \mathbf{W}\mathbf{W}^\top \mathbf{x} = \mathbf{A}\mathbf{A}^\top \mathbf{x} + \mathbf{W}_2 \mathbf{W}_2^\top \mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{W}_2 \mathbf{z}$, $\mathbf{y} \in \mathbb{R}^r$, $\mathbf{z} \in \mathbb{R}^{d-r}$. If f has a true active subspace, the problem becomes

$$\text{find } \mathbf{y}^* \in \underset{\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^r}{\operatorname{argmin}} f(p_{\mathcal{X}}(\mathbf{A}\mathbf{y})) \tag{9}$$

with $p_{\mathcal{X}}$ the convex projection onto $\mathcal{X}$; otherwise, the problem is:

$$\text{find } (\mathbf{y}^*, \mathbf{z}^*) \in \underset{\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^r, \mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^{d-r}}{\operatorname{argmin}} f(\mathbf{A}\mathbf{y} + \mathbf{W}_2 \mathbf{z}) \text{ s.t. } \mathbf{A}\mathbf{y} + \mathbf{W}_2 \mathbf{z} \in \mathcal{X}.$$

As discussed by Constantine (2015), perhaps the optimization over $\mathbf{z}$ is secondary, which can be rewritten in the form

$$\text{find } \mathbf{y}^* \in \underset{\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^r}{\operatorname{argmin}} \min_{\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^{d-r}} f(\mathbf{A}\mathbf{y} + \mathbf{W}_2 \mathbf{z}) \text{ s.t. } \mathbf{A}\mathbf{y} \in \mathcal{X}$$

where less effort can be put on solving the high-dimensional subproblem involving $\mathbf{z}$. This can be performed by random sampling (Constantine, 2015). Alternatively, we can take a page from optimization in the context of variable selection and fix the inactive directions, see e.g., Cartis et al. (2020). Innocuous for unbounded domains[4], the constraint to belong to the domain for evaluation is complex for a compact domain $\mathcal{X}$. The optimization problem is noisy whenever $\mathbf{A}$ changes over iterations as design points need to be projected on the embedding (that is, when a $\mathbf{z}$ component is present but ignored in the modeling of f).

Discussing solely the (centered) $[-1, 1]^d$ hypercubic case, the intersection with a linear embedding is a convex polytope, defined by $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^r \text{ s.t. } -1 \leq \mathbf{A}\mathbf{y} \leq 1\}$. This is the domain advocated by Letham et al. (2020) to solve (9), at the extra cost of handling the linear constraints in optimizing α. With random $\mathbf{A}$, a simpler option is to optimize $\mathbf{y}$ within a hypercubic domain $\mathcal{Y} = [-l, l]^r$, with l chosen based on the probability of finding

[4] At least in the formulation, unbounded domains are another topic in BO, see e.g., Shahriari et al. (2016a).

a solution, as derived in Wang et al. (2013, 2016); Qian et al. (2016); Cartis et al. (2020). These results are simpler if the true active subspace is axis aligned, and depend on the difference between r and the true low dimension (if it exists). Nonetheless, these choices for $\mathcal{Y}$ may not even contain $\mathbf{y}^*$. The smallest compact set for this is a star-shaped polygon described in Binois et al. (2020). The reason to focus on smaller $\mathcal{Y}$ is two-fold. First, most of the difference corresponds to points on the boundary of $\mathcal{X}$, with increasingly large volume compared to the intersection, bringing back the curse of dimensionality. Second is that the corresponding observations are distorted with the convex projection and are thus harder to model. These issues are by-passed by Nayebi et al. (2019); Moriconi et al. (2020), at the cost of the stricter diagonal or axis-aligned embedding assumptions. Changing the domain, as is implicitly done by choosing l relates to another strategy not explored much in BO: to reduce the optimization space, as discussed by Stork et al. (2020) or employed with TRs.

A weighted-PCA estimation of $\mathbf{A}$ is proposed in Raponi et al. (2020), and they handle the domain issue with a penalty for infeasibility. Chen et al. (2020a) use a semi-supervised version of sliced inverse regression (SIR) to find important input directions, using both labeled (evaluated) and unlabeled (unevaluated) designs. Namely, they collect points which have high acquisition function values but were not chosen for evaluation into an "unlabeled dataset", which they incorporate into the estimation of the embedding. The "inverse" in SIR indicates the goal of swapping the role of $\mathbf{x}$ and y, finding which values of the designs lead to given values of the outputs. Selecting a domain suffers the same difficulties as above, plus those related to updating the SIR model.

Most existing work fixes the embedding, or does not propagate the uncertainty when estimated iteratively. Exploring the interplay between acquisition functions defined over AS functions, e.g., in Garnett et al. (2014); Wycoff (2021b) and those defined for optimization is a promising direction to enhance both of them.

So when it comes to embeddings, they more often complicate than simplify the acquisition function search. This can still be worth the extra effort when a sufficient evaluation budget exists to learn the embedding.

These problems of domain selection are amplified for non-linear dimension reduction as illustrated by Siivola et al. (2021). Li et al. (2016) faces both additivity and embedding issues, restricting the extent of the projection term to remain mostly in the domain. Next we provide pointers to benchmark problems to empirically assess the various methods detailed above.

5 Synthetic test optimization problems

Due to the multiplicity of possible methods and combinations, a thorough empirical comparison is out of scope in this work. Besides, implementations are not always available or even compatible. Inference techniques may differ as well, complicating comparisons. A less obvious difficulty is the absence of standard benchmark functions for high-dimensional surrogate-based optimization. We thus list here options we have come across.

First, some of the standard benchmark functions from global optimization, say as in Hansen et al. (2021), can be scaled to high dimension. These synthetic test functions may be grouped into those which are separable, unimodal with moderate (resp. high) conditioning and finally multi-modal with global structure. Only the last one is in the realm of BO, as stated by Diouane et al. (2021). But even these are often pathological, for instance with exponentially many local optima as d increases, and consequently demand a greater budget than could be reasonably afforded to find the global solution, regardless of how quickly the simulator can be evaluated. Defining reasonable, sub-globally-optimal targets for these scalable problems under limited budgets could be an option. Table 3 of the Appendix enumerates these global optimization functions. Additional ones, created for BO, are tested by Siivola et al. (2018). The five physically-motivated analytical functions from `https://www.sfu.ca/~ssurjano/optimization.html` have been the subject of BO benchmarking too: Borehole ($8d$), OTL circuit ($6d$), piston ($7d$), robot arm ($8d$), and wingweight ($10d$). The power circuit function ($13d$) (Lee, 2019) makes one additional such example. Several functions designed for sensitivity analysis have been experimented on as well: the Sobol g-function (any d), the Ishigami function ($3d$), and the $8d$ flood model in Iooss and Lemaître (2015). Assessing the adequacy of these analytical examples with the structural model assumptions is interesting future work.

Another option is to add artificial, inactive variables to classical low dimensional test functions, making the problem nominally high dimensional yet intrinsically low dimensional. While this scales easily to billions of variables as in Wang et al. (2013), it may not be realistic. Repeated versions of these low dimensional toy functions are also proposed, e.g., by Oh et al. (2018). This approach simply sums many low-dimensional versions of the same function together, allowing for more active variables. Another option is to rotate the low-dimensional function via a linear transformation, with two snags: the initial hypercubic domain may not fill the entire high-dimensional one, or parts may be mapped outside, including known optima. Analytical functions can be extended easily to fill the gaps, but the difficulty of the problem becomes rotation dependent. A set of fixed rotations could be used for ease of comparison, or embedding matrices like the class proposed in Nayebi et al. (2019) that do not have the same domain issues.

GP realizations are an easy option to test the estimation difficulty of a given structure, and may also be used for comparisons between structures as in Ginsbourger et al. (2016). The extent to which these results apply to real applications remains a topic of research. To this end, a list of problems which may be reproduced using publicly available software is provided in Table 1. Some of these problems arise from hyperparameter tuning applications that are popular in the machine learning community, and consist of tuning various neural networks properties with BO. The choice of the neural networks structure seems to often be hand tuned, the results are noisy and platform dependent, but the code is sometimes available to reproduce results. The most realistic examples comes from engineering and simulation, but the pipeline to reproduce high-fidelity simulation is rarely available in these cases. Besides the ones in Table 1 for which this is the case, examples include the design of airfoils, wings or fans (Chen et al., 2020b; Gaudrie et al., 2020; Palar and Shimoyama, 2017; Viswanath et al., 2011; Lukaczyk et al., 2014; Seshadri et al., 2019), automotive industry test cases (Binois, 2015), alloy design (Li et al., 2017; Rana et al., 2017), biology (Ulmasov et al., 2016), physics (Kirschner et al., 2019; Mutny and Krause, 2018), and electronics (Jones et al., 1998).

Table 1: Examples of reproducible, realistic test problems.

Name(s)	Field/Domain	d	Example reference
Bayesian neural net	Hyperparameter tuning	5	Falkner et al. (2018)
Cartpole swing-up task	Hyperparameter tuning	8	Falkner et al. (2018)
Covid-19 model	Epidemiology	8	Wycoff et al. (2021)
Elliptic PDE	Simulation	100	Tripathy et al. (2016)
Groundwater remediation	Environment	6	Gramacy (2020)
Neural net	Hyperparameter tuning	6	Falkner et al. (2018)
Robot pushing	Robotics	14	Wang et al. (2018)
Rover trajectory	Robotics	60	Wang et al. (2018)
Satellite drag	Astrodynamics	7	Gramacy (2020)
SDSS	Cosmology	9	Kandasamy et al. (2015)
Shallow Net Weights	Hyperparameter tuning	100-500	Oh et al. (2018)
Three-link walk robot	Robotics	25	Zhang et al. (2019)
VJ Cascade classifier	Hyperparameter tuning	22	Kandasamy et al. (2015)

High dimensional data sets are another proper option to test high-dimensional GP regression, and can be extended to BO by allowing the BO algorithm to select not arbitrary points in the input space, but only those points corresponding to observed data. However, this discounts the effect of continuous optimization of the acquisition function, a fully non-neglible component of BO on real black-boxes. Still, some common data sets for regression tasks that have been used are listed in Table 2. One option, following the example of (Jones, 2008) presenting at MOPTA, is to fit an interpolant to a given dataset and then treat that interpolant as the black-box to be optimized. We should remark that if the interpolant used to generate the "black-box" is the same as that used for BO, the results may be unrealistically rosy.

From the set of examples listed above, we see that the definition of "high dimensional" in BO varies extensively, from the low tens to several thousands. The optimization budgets share the same range of variation, but each analysis is conducted only for a fixed budget. There is thus room for a thorough study on which high-dimensional structure is the most adapted given the dimension and budget, or even better, enabling more complexity as more evaluations become available.

Table 2: Example of data sets for regression used for GPs and BO.

Name(s)	d	n	Example reference
Airfoil	5	1,503	Francom and Sansó (2019)
Amazon commerce reviews	10,000	1,500	Fukumizu and Leng (2014)
Autos	24	159	Delbridge et al. (2020)
Communities and crime	96	1,993	Garnett et al. (2014)
Concrete	8	1,030	Garnett et al. (2014)
Crash	9	20	Gramacy (2020)
CT slices	318	3,071	Garnett et al. (2014)
Elevators	17	8,752	Gilboa et al. (2013)
Engine block and head joint sealing	8	27	Joseph et al. (2008)
Heat exchanger	4	64	Lin and Joseph (2020)
Housing	13	506	Delbridge et al. (2020)
Isomap faces	4,096	698	Guhaniyogi and Dunson (2016)
Kink40k	8	10,000	Gilboa et al. (2013)
LGBB	3	3,167	Gramacy (2020)
MARTHE	20	300	Marrel et al. (2008)
Olivetti faces data set (modified)	1,935	400	Delbridge et al. (2020)
PIC	9	45,730	Hoang et al. (2018)
Pumadyn	8	7,168	Gilboa et al. (2013)
Pumadyn	32	7,168	Gilboa et al. (2013)
Sarcos robot	21	44,484	Winkel et al. (2021)
Servo	3	167	Duvenaud et al. (2011)
Tecator	100	215	Plate (1999)
Temperature	106	10,675	Garnett et al. (2014)
Yacht	6	308	Garnett et al. (2014)

6 Conclusion and perspectives

Often rightly presented as one of the top challenges for GP-based BO, high-dimensionality has generated many different ideas to address each aspect of the manifestation of the curse of dimensionality. While the structural assumptions seem to cluster into one or more of variable selection, additive decomposition or linear embeddings, no consensus is present on the best way to tackle the problem. Even if this is largely a problem-dependent issue, the inference methods used also restrict which structures are possible to assume, whether relying on a random structure or actually trying to infer it. Even besides these specifics, the particular strategy used for acquisition function optimization can overshadow the rest. More systematic comparisons are necessary, as is the definition of suitable benchmarks.

6.1 Some general guidelines

As mentioned at the beginning of Section 3, most structural model assumptions can be seen as instances of the more general model:

$$\text{model: } f(\mathbf{x}) \approx \sum_{i=1}^{\kappa} g_i(\mathbf{A}_i \mathbf{x}) \tag{10}$$

with $\mathbf{A}_i \in \mathbb{R}^{r \times d}$, $\kappa \in \mathbb{N}^*$, which allows for sums of low dimensional effects, This includes at one end the additive Gaussian process of Durrande et al. (2012) which separates each variable individually into its own bin, at another end linear embeddings of the form $f(\mathbf{x}) = h(\mathbf{A}\mathbf{x})$ which ignore certain input directions, and beyond lie the functional ANOVA kernels of Muehlenstaedt et al. (2012) or in combining several random low dimensional subspaces such as in Delbridge et al. (2020). There are still promising research directions at the intersections of these methods, and we argue in particular that “default” kernel components, for instance a simple isotropic kernel, could be useful in complementing methods which search a particular part of the input space. The various structural modeling options are presented in Figure 2, highlighting the links between options.

Rather than starting directly from the most general model, whose direct inference would make a very steep initial step, we recommend to try the simpler models first. That is, begin

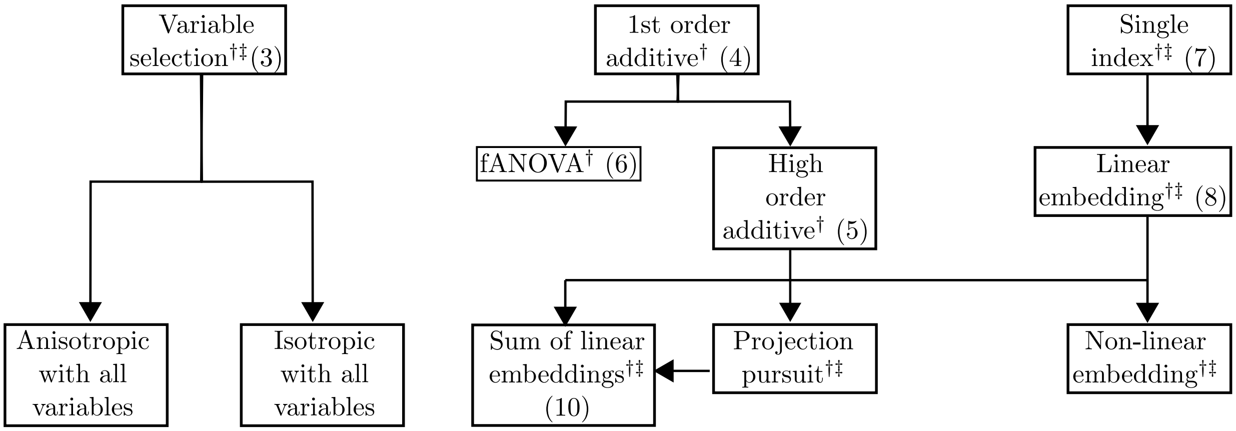

Figure 2: A taxonomy of structural model assumptions, going from simpler (top) to more general ones (bottom). Arrows marks models that can be generalized, (†) indicates that the estimation of a noise term is needed and (‡) cases for which a filling strategy is needed unless optimization is conducted only on a low dimensional manifold.

with a standard GP to build an alternative one based on the ARD principle, or to compute sensitivity indices like Sobol indices. Then build a fully additive GP model or a single index one as the basic models of the additive and linear embedding families. From comparing the prediction given by these simple models, building more advanced models from the best performing one is possible: functional ANOVA or block additive from the additive one; linear embedding on top of single index. If none is performing well, projection pursuit or non-linear embeddings could be entertained, if n is sufficiently large to allow a reliable inference of the corresponding model.

More precisely for the inference, depending on d, the budget n and the application dependent complexity of f, inference by maximum likelihood (or, even better, a Bayesian alternative) is preferred. If too difficult or slow when d reaches the hundreds, relying on a random structure to bypass the full-dimensional model inference remains possible. The simplest models scale well in computation with increasing d, while the more complex ones may benefit from recent advances brought by the increasing use of GPUs and automatic differentiation frameworks. This is especially true as more complex models require larger values of n to show improvements.

Alternatively, if the budget of optimization is limited, a simple filling strategy for the “inactive” variables combined with optimization on some restricted subspace could be preferable. Finally the use of the trust region framework has shown promising results for optimization in high dimensional problems, and is amenable to parallelization.

6.2 Future work

There is potential for hybrids between models, inference methods and mathematical optimization techniques. For global optimization, a research avenue pointed out by Spagnol et al. (2019) is to focus on variables that are important to reach low function values, rather than evaluating their influence on the whole range of f. Chen et al. (2020a) go some way to focusing on low values, as do Guhaniyogi and Dunson (2016) on manifold-bound inputs. The ability to doubt the current structural assumption may also be quite helpful in order to avoid negative feedback loop effects appearing in adaptive design, see e.g., Gramacy (2020).

Whether the ability to keep a kernel defined on the whole initial domain is superior to defining one directly on a low dimensional embedding (which while appealing is associated with domain selection issues) remains to be seen. If a kernel defined on an embedding is indeed preferred, keeping kernels sufficiently expressive (e.g., not restricted to be first order additive like in Durrande et al. (2013)) by adding an isotropic component may be improved by considering ortho-additivity (Ginsbourger et al., 2016) instead. These orthogonality constraints are likely to help inference by reducing identifiability issues.

Additionally, the success of trust region BO approaches which restrict the search space to some subset of the input space motivates dimension reduction approaches which are themselves local. For instance Wycoff (2021a) propose to define an active subspace with respect to probability measures with non-uniform densities in order to emphasize certain regions of interest. Namely, the focus is restricted to a trust region, but other schemes could be considered. Of course, a dimension reduction which is *locally* linear is in fact globally nonlinear, and such local linear models may prove to be a tractable approach to developing nonlinear dimension reduction for BO.

Furthermore, there have already been useful results born of the influence of mathematical programming on approaches to high dimensional Bayesian optimization. In particular, use of a rectangular trust region seems to be of benefit in Eriksson et al. (2019) and a Gaussian process analogue of the augmented Lagrangian method is developed in Gramacy et al. (2016); Picheny et al. (2016). Helping to develop new algorithms which combine benefits from low dimensional modeling with trust regions or other approaches successful in high dimensional mathematical programming broadly could help BO finally make the leap to practical, widespread usage on complex, high dimensional problems.

A Synthetic benchmark functions

Some global optimization benchmark functions are used for testing high-dimensional BO, of which a list is provided in Table 3.

Table 3: Synthetic problems for global optimization benchmarks tested on for high-dimensional BO. See e.g., `https://www.sfu.ca/~ssurjano/optimization.html` for further details, expressions and codes, or Cartis and Otemissov (2021). Cola is used by Binois et al. (2020) while the Beach, Dream and Simba functions can be found in Winkel et al. (2021).

Name(s)	d	Name(s)	d
Ackley	Any	Hartmann 3	3
Beach	6	Hartmann 6	6
Beale	2	Levy	Any
Branin	2	Michalewicz	Any
Brent	2	Perm 4, 0.5	4
Bukin N.6	2	Quadratic Form	Any
Camel	2	Rastrigin	any
Cola	17	Rosenbrock	Any
Colville	4	Simba	6
Dream	6	Shekel 5,7,10	4
Easom	2	Shubert	2
Franke	2	Styblinski Tang	Any
Friedman	5	Trid	any
Giunta	2	Welsh	20
Goldstein-Price	2	Zakharov	any
Griewank	Any	Zettl	2

References

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). "On the surprising behavior of distance metrics in high dimensional space." In *Proceedings of the 8th International Conference on Database Theory, 2001*, 420–434.

Artstein-Avidan, S., Giannopoulos, A., and Milman, V. D. (2015). *Asymptotic geometric analysis, Part I*, vol. 202. American Mathematical Soc.

Bach, F. (2017). "Breaking the curse of dimensionality with convex neural networks." *The Journal of Machine Learning Research*, 18, 1, 629–681.

Bellman, R. (1966). "Dynamic programming." *Science*, 153, 3731, 34–37.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). "Algorithms for Hyper-Parameter Optimization." In *Advances in Neural Information Processing Systems*, eds. J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, vol. 24. Curran Associates, Inc.

Bergstra, J. and Bengio, Y. (2012). "Random search for hyper-parameter optimization." *The Journal of Machine Learning Research*, 13, 281–305.

Binois, M. (2015). "Uncertainty quantification on Pareto fronts and high-dimensional strategies in Bayesian optimization, with applications in multi-objective automotive design." Ph.D. thesis, Saint-Etienne, EMSE.

Binois, M., Ginsbourger, D., and Roustant, O. (2015). "A Warped Kernel Improving Robustness in Bayesian Optimization Via Random Embeddings." In *Learning and Intelligent Optimization*, eds. C. Dhaenens, L. Jourdan, and M.-E. Marmion, vol. 8994 of *Lecture Notes in Computer Science*, 281–286. Springer International Publishing.

— (2020). "On the choice of the low-dimensional domain for global optimization via random embeddings." *Journal of Global optimization*, 76, 1, 69–90.

Blanchet-Scalliet, C., Helbert, C., Ribaud, M., and Vial, C. (2017). "A specific kriging kernel for dimensionality reduction: Isotropic by group kernel." *preprint hal-01496521v1*.

Bouhlel, M. A., Bartoli, N., Otsmane, A., and Morlier, J. (2016). “Improving kriging surrogates of high-dimensional design models by Partial Least Squares dimension reduction.” *Structural and Multidisciplinary Optimization*, 53, 5, 935–952.

Bridges, R., Gruber, A., Felder, C., Verma, M., and Hoff, C. (2019). “Active Manifolds: A non-linear analogue to Active Subspaces.” In *International Conference on Machine Learning*, 764–772. PMLR.

Carpentier, A. and Munos, R. (2012). “Bandit theory meets compressed sensing for high dimensional stochastic linear bandit.” In *Artificial Intelligence and Statistics*, 190–198.

Cartis, C., Massart, E., and Otemissov, A. (2020). “Constrained global optimization of functions with low effective dimensionality using multiple random embeddings.” *arXiv preprint arXiv:2009.10446*.

Cartis, C. and Otemissov, A. (2021). “A dimensionality reduction technique for unconstrained global optimization of functions with low effective dimensionality.” *Information and Inference: A Journal of the IMA*.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). “Handling sparsity via the horseshoe.” *Journal of Machine Learning Research*.

Chen, B., Castro, R., and Krause, A. (2012). “Joint optimization and variable selection of high-dimensional Gaussian processes.” In *Proc. International Conference on Machine Learning (ICML)*.

Chen, G. and Tuo, R. (2020). “Projection Pursuit Gaussian Process Regression.” *arXiv preprint arXiv:2004.00667*.

Chen, J., Zhu, G., Yuan, C., and Huang, Y. (2020a). “Semi-supervised Embedding Learning for High-dimensional Bayesian Optimization.” *arXiv preprint arXiv:2005.14601*.

Chen, W., Chiu, K., and Fuge, M. D. (2020b). “Airfoil Design Parameterization and Optimization Using Bézier Generative Adversarial Networks.” *AIAA Journal*, 58, 11, 4723–4735.

Chipman, H., Ranjan, P., and Wang, W. (2012). "Sequential design for computer experiments with a flexible Bayesian additive model." *Canadian Journal of Statistics*, 40, 4, 663–678.

Constantine, P. G. (2015). *Active subspaces: Emerging ideas for dimension reduction in parameter studies*. SIAM.

Constantine, P. G., del Rosario, Z., and Iaccarino, G. (2016). "Many physical laws are ridge functions." *arXiv preprint arXiv:1605.07974*.

Damianou, A. and Lawrence, N. D. (2013). "Deep Gaussian processes." In *Artificial intelligence and statistics*, 207–215. PMLR.

Daulton, S., Eriksson, D., Balandat, M., and Bakshy, E. (2021). "Multi-Objective Bayesian Optimization over High-Dimensional Search Spaces."

Delbridge, I., Bindel, D., and Wilson, A. G. (2020). "Randomly projected additive Gaussian processes for regression." In *International Conference on Machine Learning*, 2453–2463. PMLR.

Deshwal, A. and Doppa, J. (2021). "Combining Latent Space and Structured Kernels for Bayesian Optimization over Combinatorial Spaces." *Advances in Neural Information Processing Systems*, 34.

Diouane, Y., Picheny, V., Riche, R. L., and Di Perrotolo, A. S. (2021). "TREGO: a Trust-Region Framework for Efficient Global Optimization." *arXiv preprint arXiv:2101.06808*.

Djolonga, J., Krause, A., and Cevher, V. (2013). "High-Dimensional Gaussian Process Bandits." In *Neural Information Processing Systems*, 1025–1033.

Durrande, N., Ginsbourger, D., and Roustant, O. (2012). "Additive Kernels for Gaussian Process Modeling." *Annales de la Facultée de Sciences de Toulouse*, 17.

Durrande, N., Ginsbourger, D., Roustant, O., and Carraro, L. (2013). "ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis." *Journal of Multivariate Analysis*, 115, 57–67.

Duvenaud, D. K., Nickisch, H., and Rasmussen, C. E. (2011). "Additive Gaussian processes." In *Advances in neural information processing systems*, 226–234.

Efron, B. and Stein, C. (1981). "The jackknife estimate of variance." *The Annals of Statistics*, 586–596.

Eismann, S., Levy, D., Shu, R., Bartzsch, S., and Ermon, S. (2018). "Bayesian optimization and attribute adjustment." In *UAI*.

Eriksson, D., Dong, K., Lee, E., Bindel, D., and Wilson, A. G. (2018). "Scaling Gaussian Process Regression with Derivatives." In *Advances in Neural Information Processing Systems*, 6866–6876.

Eriksson, D. and Jankowiak, M. (2021). "High-Dimensional Bayesian Optimization with Sparse Axis-Aligned Subspaces." *CoRR*, abs/2103.00349.

Eriksson, D., Pearce, M., Gardner, J., Turner, R. D., and Poloczek, M. (2019). "Scalable Global Optimization via Local Bayesian Optimization." In *Advances in Neural Information Processing Systems*, 5497–5508.

Espinasse, T., Gamboa, F., and Loubes, J.-M. (2014). "Parametric estimation for Gaussian fields indexed by graphs." *Probability Theory and Related Fields*, 159, 1-2, 117–155.

Falkner, S., Klein, A., and Hutter, F. (2018). "BOHB: Robust and efficient hyperparameter optimization at scale." In *International Conference on Machine Learning*, 1437–1446. PMLR.

Forrester, A., Sobester, A., and Keane, A. (2008). *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons.

Francom, D. and Sansó, B. (2019). "Bass: An R package for fitting and performing sensitivity analysis of Bayesian adaptive spline surfaces." *Journal of Statistical Software*, 2.

Frazier, P. I. (2018). "Bayesian optimization." In *Recent Advances in Optimization and Modeling of Contemporary Problems*, 255–278. INFORMS.

Fukumizu, K. and Leng, C. (2014). "Gradient-based kernel dimension reduction for regression." *Journal of the American Statistical Association*, 109, 505, 359–370.

Gardner, J., Guo, C., Weinberger, K., Garnett, R., and Grosse, R. (2017). "Discovering and Exploiting Additive Structure for Bayesian Optimization." In *Artificial Intelligence and Statistics*, 1311–1319.

Garnett, R. (2022). "Bayesian Optimization."

Garnett, R., Osborne, M. A., and Hennig, P. (2014). "Active learning of linear embeddings for Gaussian processes." In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, 230–239. AUAI Press.

Gaudrie, D., Le Riche, R., Picheny, V., Enaux, B., and Herbert, V. (2020). "Modeling and optimization with Gaussian processes in reduced eigenbases." *Structural and Multidisciplinary Optimization*, 1–19.

Gauthier, B. and Bay, X. (2012). "Spectral approach for kernel-based interpolation." *Annales de la Faculté des sciences de Toulouse: Mathématiques*, 21, 3, 439–479.

Gilboa, E., Saatçi, Y., and Cunningham, J. (2013). "Scaling multidimensional Gaussian processes using projected additive approximations." In *International Conference on Machine Learning*, 454–461. PMLR.

Ginsbourger, D. (2018). *Sequential Design of Computer Experiments*, 1–9. American Cancer Society.

Ginsbourger, D., Rosspopoff, B., Pirot, G., Durrande, N., and Renard, P. (2012). "Distance-based kriging relying on proxy simulations for inverse conditioning." *Advances in Water Resources*.

Ginsbourger, D., Roustant, O., Schuhmacher, D., Durrande, N., and Lenz, N. (2016). "On ANOVA decompositions of kernels and Gaussian random field paths." In *Monte Carlo and Quasi-Monte Carlo Methods*, 315–330. Springer.

Gramacy, R. B. (2020). *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. CRC Press.

Gramacy, R. B., Gray, G. A., Digabel, S. L., Lee, H. K. H., Ranjan, P., Wells, G., and Wild, S. M. (2016). "Modeling an Augmented Lagrangian for Blackbox Constrained Optimization." *Technometrics*, 58, 1, 1–11.

Gramacy, R. B. and Lee, H. K. (2012). "Cases for the nugget in modeling computer experiments." *Statistics and Computing*, 22, 3, 713–722.

Gramacy, R. B. and Lian, H. (2012). "Gaussian process single-index models as emulators for computer experiments." *Technometrics*, 54, 1, 30–41.

Grosnit, A., Tutunov, R., Maraval, A. M., Griffiths, R.-R., Cowen-Rivers, A. I., Yang, L., Zhu, L., Lyu, W., Chen, Z., Wang, J., et al. (2021). "High-dimensional Bayesian optimisation with variational autoencoders and deep metric learning." *arXiv preprint arXiv:2106.03609*.

Groves, M. and Pyzer-Knapp, E. O. (2018). "Efficient and scalable batch Bayesian optimization using K-means." *arXiv preprint arXiv:1806.01159*.

Guhaniyogi, R. and Dunson, D. B. (2016). "Compressed Gaussian process for manifold regression." *Journal of Machine Learning Research*, 17, 69, 1–26.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2018). "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules." *ACS Central Science*, 4, 2, 268–276. PMID: 29532027.

Hansen, N., Auger, A., Ros, R., Mersmann, O., Tušar, T., and Brockhoff, D. (2021). "COCO: A platform for comparing continuous optimizers in a black-box setting." *Optimization Methods and Software*, 36, 1, 114–144.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., et al. (2019). "A case study competition among methods for analyzing large spatial data." *Journal of Agricultural, Biological and Environmental Statistics*, 24, 3, 398–425.

Hebbal, A., Brevault, L., Balesdent, M., Talbi, E.-G., and Melab, N. (2019). "Multi-objective optimization using Deep Gaussian Processes: Application to aerospace vehicle design." In *AIAA Scitech 2019 Forum*, 1973.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). "Gaussian processes for Big data." In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 282–290.

Hoang, T. N., Hoang, Q. M., Ouyang, R., and Low, K. H. (2018). "Decentralized high-dimensional Bayesian optimization with factor graphs." In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Huang, W., Zhao, D., Sun, F., Liu, H., and Chang, E. (2015). "Scalable Gaussian process regression using deep neural networks." In *Twenty-fourth international joint conference on artificial intelligence*.

Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2011). "Sequential model-based optimization for general algorithm configuration." In *International conference on learning and intelligent optimization*, 507–523. Springer.

Iooss, B. and Lemaître, P. (2015). "A review on global sensitivity analysis methods." In *Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications*, eds. C. Meloni and G. Dellino. Springer.

Jaquier, N. and Rozo, L. (2020). "High-Dimensional Bayesian Optimization via Nested Riemannian Manifolds." *Advances in Neural Information Processing Systems*, 33.

Jayasumana, S., Hartley, R., Salzmann, M., Li, H., and Harandi, M. (2014). "Optimizing over radial kernels on compact manifolds." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3802–3809.

Johnson, W. B. and Lindenstrauss, J. (1984). "Extensions of Lipschitz mappings into a Hilbert space 26." *Contemporary mathematics*, 26.

Jones, D., Schonlau, M., and Welch, W. (1998). "Efficient global optimization of expensive black-box functions." *Journal of Global Optimization*, 13, 4, 455–492.

Jones, D. R. (2008). "Large-scale multi-disciplinary mass optimization in the auto industry." FORTRAN code for this simulator was retrieved from `https://www.miguelanjos.com/jones-benchmark`.

Joseph, V., Hung, Y., and Sudjianto, A. (2008). "Blind Kriging: a new method for developing metamodels." *Journal of Mechanical Design*, 130, 3, 31102.

Journel, A. G. (1974). "Geostatistics for conditional simulation of ore bodies." *Economic Geology*, 69, 5, 673–687.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018). "Gaussian processes and kernel methods: A review on connections and equivalences." *arXiv preprint arXiv:1807.02582*.

Kandasamy, K., Schneider, J., and Póczos, B. (2015). "High dimensional Bayesian optimisation and bandits via additive models." In *International conference on machine learning*, 295–304. PMLR.

Kersaudy, P., Sudret, B., Varsier, N., Picon, O., and Wiart, J. (2015). "A new surrogate modeling technique combining Kriging and polynomial chaos expansions–Application to uncertainty analysis in computational dosimetry." *Journal of Computational Physics*, 286, 103–117.

Kingma, D. P. and Welling, M. (2013). "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114*.

Kirschner, J., Mutnỳ, M., Hiller, N., Ischebeck, R., and Krause, A. (2019). "Adaptive and safe Bayesian optimization in high dimensions via one-dimensional subspaces." *arXiv preprint arXiv:1902.03229*.

Kleijnen, J. P. and van Beers, W. C. (2020). "Prediction for big data through Kriging: small sequential and one-shot designs." *American Journal of Mathematical and Management Sciences*, 39, 3, 199–213.

Köppen, M. (2000). "The Curse of Dimensionality." In *5th online world conference on soft computing in industrial applications (WSC5)*, vol. 1, 4–8.

Larson, J., Menickelly, M., and Wild, S. M. (2019). "Derivative-free optimization methods." *Acta Numerica*, 28, 287–404.

Lawrence, N. (2005). "Probabilistic non-linear principal component analysis with Gaussian process latent variable models." *The Journal of Machine Learning Research*, 6, 1783–1816.

Lee, M. R. (2019). "Modified Active Subspaces Using the Average of Gradients." *SIAM/ASA Journal on Uncertainty Quantification*, 7, 1, 53–66.

Lenz, N. (2013). "Additivity and Ortho-Additivity in Gaussian Random Fields." Master's thesis, University of Bern, Switzerland.

Letham, B., Calandra, R., Rai, A., and Bakshy, E. (2020). "Re-Examining Linear Embeddings for High-Dimensional Bayesian Optimization." *Advances in Neural Information Processing Systems*, 33.

Letham, B., Karrer, B., Ottoni, G., and Bakshy, E. (2018). "Constrained Bayesian optimization with noisy experiments." *Bayesian Analysis*.

Li, C., Gupta, S., Rana, S., Nguyen, V., Venkatesh, S., and Shilton, A. (2017). "High Dimensional Bayesian Optimization using Dropout." In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2096–2102.

Li, C.-L., Kandasamy, K., Póczos, B., and Schneider, J. (2016). "High Dimensional Bayesian Optimization via Restricted Projection Pursuit Models." In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 884–892.

Liang, H., Zhu, M., and Wu, Z. (2014). "Using Cross-Validation to Design Trend Function in Kriging Surrogate Modeling." *AIAA Journal*, 52, 10, 2313–2327.

Lin, L.-H. and Joseph, R. V. (2020). "Transformation and additivity in Gaussian processes." *Technometrics*, 62, 4, 525–535.

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D., and Ye, K. Q. (2006). "Variable selection for Gaussian process models in computer experiments." *Technometrics*, 48, 4, 478–490.

Lukaczyk, T. W., Constantine, P., Palacios, F., and Alonso, J. J. (2014). "Active subspaces for shape optimization." In *10th AIAA multidisciplinary design optimization conference*, 1171.

Marcy, P. (2018). "Bayesian Gaussian Process Models for Dimension Reduction Uncertainties." ASA Joint research conference.

Marrel, A., Iooss, B., Van Dorpe, F., and Volkova, E. (2008). "An efficient methodology for modeling complex computer codes with Gaussian processes." *Computational Statistics & Data Analysis*, 52, 10, 4731–4744.

Matérn, B. (1960). "Spatial variation. Stochastic models and their application to some problems in forest surveys and other sampling investigations." *Meddelanden fran Statens Skogsforskningsinstitut*, 49, 5.

Maus, N., Jones, H. T., Moore, J. S., Kusner, M. J., Bradshaw, J., and Gardner, J. R. (2022). "Local Latent Space Bayesian Optimization over Structured Inputs."

Mockus, J., Tiesis, V., and Zilinskas, A. (1978). "The application of Bayesian methods for seeking the extremum." *Towards Global Optimization*, 2, 117-129, 2.

Moriconi, R., Kumar, K. S., and Deisenroth, M. P. (2020). "High-dimensional Bayesian optimization with projections using quantile Gaussian processes." *Optimization Letters*, 14, 1, 51–64.

Morris, M. D. (1991). "Factorial sampling plans for preliminary computational experiments." *Technometrics*, 33, 2, 161–174.

Muehlenstaedt, T., Roustant, O., Carraro, L., and Kuhnt, S. (2012). "Data-driven Kriging models based on FANOVA-decomposition." *Statistics and Computing*, 22, 3, 723–738.

Mutny, M. and Krause, A. (2018). "Efficient high dimensional Bayesian optimization with additivity and quadrature Fourier features." In *Advances in Neural Information Processing Systems*, 9005–9016.

Nayebi, A., Munteanu, A., and Poloczek, M. (2019). "A framework for Bayesian optimization in embedded subspaces." In *International Conference on Machine Learning*, 4752–4761.

Neal, R. M. (1997). "Monte Carlo implementation of Gaussian process models for Bayesian regression and classification." *arXiv preprint physics/9701026*.

Oh, C., Gavves, E., and Welling, M. (2018). "BOCK: Bayesian Optimization with Cylindrical Kernels." In *International Conference on Machine Learning*, 3868–3877.

Padonou, E. and Roustant, O. (2016). "Polar Gaussian processes and experimental designs in circular domains." *SIAM/ASA Journal on Uncertainty Quantification*, 4, 1, 1014–1033.

Palar, P. S. and Shimoyama, K. (2017). "Exploiting active subspaces in global optimization: how complex is your problem?" In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 1487–1494.

Papoulis, A. (1965). *Random Variables and Stochastic Processes*. McGraw-Hill.

Picheny, V., Gramacy, R. B., Wild, S., and Le Digabel, S. (2016). "Bayesian optimization under mixed constraints with a slack-variable augmented Lagrangian." *Advances in neural information processing systems*, 29.

Plate, T. A. (1999). "Accuracy versus interpretability in flexible modeling: Implementing a tradeoff using Gaussian process models." *Behaviormetrika*, 26, 1, 29–50.

Qian, H., Hu, Y.-Q., and Yu, Y. (2016). "Derivative-Free Optimization of High-Dimensional Non-Convex Functions by Sequential Random Embeddings." In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, 1946–1952. AAAI Press.

Qian, H. and Yu, Y. (2017). "Solving high-dimensional multi-objective optimization problems with low effective dimensions." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31.

Rana, S., Li, C., Gupta, S., Nguyen, V., and Venkatesh, S. (2017). "High dimensional Bayesian optimization with elastic Gaussian process." In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2883–2891. JMLR. org.

Raponi, E., Wang, H., Bujny, M., Boria, S., and Doerr, C. (2020). "High dimensional Bayesian optimization assisted by principal component analysis." In *International Conference on Parallel Problem Solving from Nature*, 169–183. Springer.

Rasmussen, C. E. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Regis, R. G. (2016). "Trust regions in Kriging-based optimization with expected improvement." *Engineering optimization*, 48, 6, 1037–1059.

Regis, R. G. and Shoemaker, C. A. (2013). "Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization." *Engineering Optimization*, 45, 5, 529–555.

Rolland, P., Scarlett, J., Bogunovic, I., and Cevher, V. (2018). "High-Dimensional Bayesian Optimization via Additive Models with Overlapping Groups." In *International Conference on Artificial Intelligence and Statistics*, 298–307.

Roustant, O., Ginsbourger, D., and Deville, Y. (2012). "DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization." *Journal of Statistical Software*, 51, 1, 1–55.

Salem, M. B., Bachoc, F., Roustant, O., Gamboa, F., and Tomaso, L. (2018). “Sequential dimension reduction for learning features of expensive black-box functions.” *preprint hal-01688329v2*.

Sauer, A., Gramacy, R. B., and Higdon, D. (2020). “Active Learning for Deep Gaussian Process Surrogates.” *arXiv preprint arXiv:2012.08015*.

Seshadri, P., Yuchi, S., and Parks, G. T. (2019). “Dimension reduction via Gaussian ridge functions.” *SIAM/ASA Journal on Uncertainty Quantification*, 7, 4, 1301–1322.

Shahriari, B., Bouchard-Côté, A., and Freitas, N. (2016a). “Unbounded Bayesian optimization via regularization.” In *Artificial intelligence and statistics*, 1168–1176. PMLR.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N. (2016b). “Taking the human out of the loop: A review of Bayesian optimization.” *Proceedings of the IEEE*, 104, 1, 148–175.

Shan, S. and Wang, G. G. (2010). “Survey of modeling and optimization strategies to solve high-dimensional design problems with computationally-expensive black-box functions.” *Structural and Multidisciplinary Optimization*, 41, 2, 219–241.

Siivola, E., Paleyes, A., González, J., and Vehtari, A. (2021). “Good practices for Bayesian optimization of high dimensional structured spaces.” *Applied AI Letters*, 2, 2, e24.

Siivola, E., Vehtari, A., Vanhatalo, J., González, J., and Andersen, M. R. (2018). “Correcting boundary over-exploration deficiencies in Bayesian optimization with virtual derivative sign observations.” In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. IEEE.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, M., and Adams, R. (2015). “Scalable Bayesian optimization using deep neural networks.” In *International conference on machine learning*, 2171–2180. PMLR.

Sobol, I. M. (2001). “Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates.” *Mathematics and computers in simulation*, 55, 1-3, 271–280.

Sommerville, D. M. Y. (1958). *An Introduction to the Geometry of n Dimensions*, vol. 512. Dover New York.

Spagnol, A., Riche, R. L., and Veiga, S. D. (2019). "Global sensitivity analysis for optimization with variable selection." *SIAM/ASA Journal on uncertainty quantification*, 7, 2, 417–443.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). "Gaussian process optimization in the bandit setting: no regret and experimental design." In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 1015–1022.

Stitson, M., Gammerman, A., Vapnik, V., Vovk, V., Watkins, C., and Weston, J. (1999). "Support vector regression with ANOVA decomposition kernels." *Advances in kernel methods—Support vector learning*, 285–292.

Stork, J., Friese, M., Zaefferer, M., Bartz-Beielstein, T., Fischbach, A., Breiderhoff, B., Naujoks, B., and Tušar, T. (2020). "Open issues in surrogate-assisted optimization." In *High-performance simulation-based optimization*, 225–244. Springer.

Sung, C.-L., Wang, W., Plumlee, M., and Haaland, B. (2019). "Multiresolution functional anova for large-scale, many-input computer experiments." *Journal of the American Statistical Association*.

Titsias, M. and Lawrence, N. D. (2010). "Bayesian Gaussian process latent variable model." In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 844–851. JMLR Workshop and Conference Proceedings.

Tran-The, H., Gupta, S., Rana, S., and Venkatesh, S. (2019). "Trading Convergence Rate with Computational Budget in High Dimensional Bayesian Optimization." *arXiv preprint arXiv:1911.11950*.

Tripathy, R., Bilionis, I., and Gonzalez, M. (2016). "Gaussian processes with built-in dimensionality reduction: Applications to high-dimensional uncertainty propagation." *Journal of Computational Physics*, 321, 191 – 223.

Tripp, A., Daxberger, E., and Hernández-Lobato, J. M. (2020). "Sample-efficient optimization in the latent space of deep generative models via weighted retraining." *Advances in Neural Information Processing Systems*, 33, 11259–11272.

Ulaganathan, S., Couckuyt, I., Dhaene, T., Degroote, J., and Laermans, E. (2016). "High dimensional Kriging metamodelling utilising gradient information." *Applied Mathematical Modelling*, 40, 9-10, 5256–5270.

Ulmasov, D., Baroukh, C., Chachuat, B., Deisenroth, M. P., and Misener, R. (2016). "Bayesian optimization with dimension scheduling: Application to biological systems." In *Computer Aided Chemical Engineering*, vol. 38, 1051–1056. Elsevier.

Viana, F. A., Simpson, T. W., Balabanov, V., and Toropov, V. (2014). "Metamodeling in Multidisciplinary Design Optimization: How Far Have We Really Come?" *AIAA Journal*, 52, 4, 670–690.

Viswanath, A., J. Forrester, A., and Keane, A. (2011). "Dimension reduction for aerodynamic design optimization." *AIAA journal*, 49, 6, 1256–1266.

Vivarelli, F. and Williams, C. K. (1999). "Discovering hidden features with Gaussian processes regression." In *Advances in Neural Information Processing Systems*, 613–619.

Wang, K., Pleiss, G., Gardner, J., Tyree, S., Weinberger, K. Q., and Wilson, A. G. (2019). "Exact Gaussian processes on a million data points." *Advances in Neural Information Processing Systems*, 32.

Wang, Z., Gehring, C., Kohli, P., and Jegelka, S. (2018). "Batched Large-scale Bayesian Optimization in High-dimensional Spaces." In *International Conference on Artificial Intelligence and Statistics*.

Wang, Z., Hutter, F., Zoghi, M., Matheson, D., and de Feitas, N. (2016). "Bayesian Optimization in a Billion Dimensions via Random Embeddings." *Journal of Artificial Intelligence Research (JAIR)*, 55, 361–387.

Wang, Z., Li, C., Jegelka, S., and Kohli, P. (2017). "Batched High-dimensional Bayesian Optimization via Structural Kernel Learning." In *International Conference on Machine Learning (ICML)*.

Wang, Z., Zoghi, M., Hutter, F., Matheson, D., and de Freitas, N. (2013). "Bayesian Optimization in a Billion Dimensions via Random Embeddings." *Proceedings of the international joint conference on Artificial Intelligence*.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. (2016). "Deep kernel learning." In *Artificial intelligence and statistics*, 370–378. PMLR.

Winkel, M. A., Stallrich, J. W., Storlie, C. B., and Reich, B. J. (2021). "Sequential Optimization in Locally Important Dimensions." *Technometrics*, 63, 2, 236–248.

Wong, C. Y., Seshadri, P., Parks, G. T., and Girolami, M. (2020). "Embedded ridge approximations." *Computer Methods in Applied Mechanics and Engineering*, 372, 113383.

Wycoff, N. (2021a). "Gradient-Based Sensitivity Analysis with Kernels." Ph.D. thesis, Virginia Tech, Blacksburg, VA 24061.

Wycoff, N., Binois, M., and Wild, S. M. (2021). "Sequential Learning of Active Subspaces." *Journal of Computational and Graphical Statistics*, 30, 4, 1224–1237.

Wycoff, N. B. (2021b). "Gradient-Based Sensitivity Analysis with Kernels." Ph.D. thesis, Virginia Tech.

Yenicelik, D. (2020). "Parameter Optimization using high-dimensional Bayesian Optimization." *arXiv preprint arXiv:2010.03955*.

Zhang, M., Li, H., and Su, S. (2019). "High dimensional Bayesian optimization via supervised dimension reduction." In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 4292–4298. AAAI Press.

Zhigljavsky, A. and Žilinskas, A. (2021). *Bayesian and High-Dimensional Global Optimization*. Springer Nature.

Zhou, J., Yang, Z., Si, Y., Kang, L., Li, H., Wang, M., and Zhang, Z. (2021). “A Trust-Region Parallel Bayesian Optimization Method for Simulation-Driven Antenna Design.” *IEEE Transactions on Antennas and Propagation*, 69, 7, 3966–3981.