Kernels over Sets of Finite Sets using RKHS Embeddings, with Application to Bayesian (Combinatorial) Optimization

Poompol Buathong[*,1] **David Ginsbourger**[*,2,3] **Tipaluck Krityakierne**[1,4]
[1]Department of Mathematics, Faculty of Science, Mahidol University, Thailand
[2]Uncertainty Quantification and Optimal Design group, Idiap Research Institute, Switzerland
[3]Institute of Mathematical Statistics and Actuarial Science, University of Bern, Switzerland
[4]Centre of Excellence in Mathematics, CHE, Thailand

Abstract

We focus on kernel methods for set-valued inputs and their application to Bayesian set optimization, notably combinatorial optimization. We investigate two classes of set kernels that both rely on Reproducing Kernel Hilbert Space embeddings, namely the "Double Sum" (DS) kernels recently considered in Bayesian set optimization, and a class introduced here called "Deep Embedding" (DE) kernels that essentially consists in applying a radial kernel on Hilbert space on top of the canonical distance induced by another kernel such as a DS kernel. We establish in particular that while DS kernels typically suffer from a lack of strict positive definiteness, vast subclasses of DE kernels built upon DS kernels do possess this property, enabling in turn combinatorial optimization without requiring to introduce a jitter parameter. Proofs of theoretical results about considered kernels are complemented by a few practicalities regarding hyperparameter fitting. We furthermore demonstrate the applicability of our approach in prediction and optimization tasks, relying both on toy examples and on two test cases from mechanical engineering and hydrogeology, respectively. Experimental results highlight the applicability and compared merits of the considered approaches while opening new perspectives in prediction and sequential design with set inputs.

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. * Both authors contributed equally to this work.

1 Introduction

Kernel methods (Aronszajn, 1950; Kimeldorf and Wahba, 1970; Schölkopf and Smola, 2002; Saitoh and Sawano, 2016) constitute a versatile framework for a variety of tasks in classification (Steinwart and Christmann, 2008), function approximation based on scattered data (Wendland, 2005), and probabilistic prediction (Rasmussen and Williams, 2006). One of the outstanding features of Gaussian Process (GP) prediction, in particular, is its usability to design Bayesian Optimization (BO) algorithms (Mockus et al., 1978; Jones et al., 1998; Frazier, 2018) and further sequential design strategies (Risk and Ludkovski, 2018; Binois et al., 2019; Bect et al., 2019). While in most usual GP- and BO-related contributions the focus is on continuous problems with vector-valued inputs, there has been a growing interest recently for situations involving discrete and mixed discrete-continuous inputs (Kondor and Lafferty, 2002; Gramacy and Taddy, 2010; Fortuin et al., 2018; Roustant et al., 2018; Garrido-Merchan and Hernández-Lobato, 2020; Ru et al., 2019; Griffiths and Hernández-Lobato, 2019). Here we focus specifically on kernels dedicated to finite set-valued inputs and their application to GP modelling and BO, notably (but not only) in combinatorial optimization.

A number of prediction and optimization problems from various application domains involve finite set-valued inputs, encompassing for instance sensor network design (Garnett et al., 2010), simulation-based investigation of the mechanical behaviour of bi-phasic materials depending on the positions of inclusions (Ginsbourger et al., 2016), inventory system optimization (Salemi et al., 2019), selection of starting centers in clustering algorithms (Kim et al., 2019), speaker recognition and image texture classification (as mentioned by Desobry et al. (2005)), natural language processing tasks with bags of words (Pappas and Popescu-Belis, 2017), or optimal positioning of landmarks in

shape analysis (Iwata, 2012), to cite a few. Yet, the number of available kernel methods for efficiently tackling such problems is still quite moderate, although the topic has gained interest among the machine learning and further research communities in the last few years. In particular, early investigations regarding the definition of positive definite kernels on finite sets encompass (Kondor and Jebara, 2003; Grauman and Darrell, 2007), and also indirectly (Cuturi et al., 2005) where kernels between atomic measures are introduced. Kernels on finite sets that have been used in BO include radial kernels with respect to the earth mover's distance (Garnett et al., 2010, where the question of their positive definiteness is not discussed), kernels on graphs implicitly defined via precision matrices in the context of Gaussian Markov Random Fields in (Salemi et al., 2019), and the class used in (Kim et al., 2019) and originating in (Haussler, 1999; Gärtner et al., 2002) that we refer to as *Double Sum* (DS) kernels. From the combinatorial optimization side, while an approach relying on Bayesian networks was considered already in (Larraiiaga et al., 2000), the topic has recently attracted attention in GP-based BO with respect to set inputs (see for instance Baptista and Poloczek (2018) where the emphasis is not on the employed kernels, and Oh et al. (2019) where graph representations are used), and also in GP-based BO over the latent space of a variational autoencoder (Griffiths and Hernández-Lobato, 2019).

Our approach here is to leverage the fertile framework of Reproducing Kernel Hilbert Space Embeddings (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007; Sriperumbudur et al., 2011; Muandet et al., 2017) to analyze DS kernels and the introduced *Deep Embedding* (DE) kernels, that consist in chaining radial kernels in Hilbert space with the canonical distance associated with set kernels like DS ones. As we establish, wide classes of DE kernels are strictly positive definite which contrasts with the typical case of DS kernels. We present in turn a few additional results pertaining to the parametrization of DE kernels and to related hyperparameter fitting, including geometrical considerations around the choice of hyperparameter bounds. Section 2 is mainly dedicated to the exposition and theoretical analysis of the considered classes of kernels, complemented by practicalities regarding hyperparameter fitting. In Section 3, numerical experiments are discussed that compare DS and DE kernels in prediction and optimization tasks, both on analytical and on two application test cases, namely in mechanical engineering with plasticity simulations of a bi-phasic material tackled in (Ginsbourger et al., 2016), and in hydrogeology with an original monitoring well selection problem based on the contaminant source localization test case from (Pirot et al., 2019).

2 Set Kernels via RKHS Embeddings

2.1 Notation and Settings

We focus on positive definite kernels defined over subsets of some base set $\mathcal{X}$. Depending on the cases, $\mathcal{X}$ may be finite or infinite. The considered set of subsets of $\mathcal{X}$, denoted $\mathcal{S}(\mathcal{X})$, may be the whole power set $\mathcal{P}(\mathcal{X})$ or a subset thereof, e.g. $\mathcal{S}_p(\mathcal{X})$ (also traditionally noted $[\mathcal{X}]^p$ in set theory) the set of p-element subsets of $\mathcal{X}$ (where $p \in \mathbb{N}$, with $p \leq \#\mathcal{X}$ in case of a finite $\mathcal{X}$ with cardinality $\#\mathcal{X}$), or the set of all (non-void) finite subsets of $\mathcal{X}$ denoted here $\mathcal{S}_{\text{fin}}(\mathcal{X}) = \cup_{p\geq 1}\mathcal{S}_p(\mathcal{X})$. Given a positive definite kernel $k_{\mathcal{X}}$ over $\mathcal{X}$ and the associated Reproducing Kernel Hilbert Space $\mathcal{H}_{k_{\mathcal{X}}}$, we call here *embedding of* $\mathcal{S}_{fin}(\mathcal{X})$ *in* $\mathcal{H}_{k_{\mathcal{X}}}$ the mapping

$$\mathcal{E} : S \in \mathcal{S}_{\text{fin}}(\mathcal{X}) \to \frac{1}{\#S}\sum_{\mathbf{x}\in S} k_{\mathcal{X}}(\mathbf{x}, \cdot) \in \mathcal{H}_{k_{\mathcal{X}}}. \quad (1)$$

Note that this "set embedding" coincides with the Kernel Mean Embedding (Muandet et al., 2017) in $\mathcal{H}_{k_{\mathcal{X}}}$ of the uniform probability distribution over S.

2.2 From Linear to Deep Embedding Kernels

A natural idea to create a positive definite kernel on $\mathcal{S}_{\text{fin}}(\mathcal{X})$ from this embedding is to plainly take:

$$k_0(S, S') = \frac{1}{\#S\#S'}\sum_{\substack{\mathbf{x}\in S\\ \mathbf{x}'\in S'}} k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}'), \quad (2)$$

which is none other than the kernel used in (Kim et al., 2019) and that we refer to here as double sum kernel. As we will see in the next section and in the applications, this positive definite kernel may suffer in some settings from its lack of strict positive definiteness. Yet it appears as a crucial building block in the class of strictly positive definite kernels that we introduce here. The first step is to consider the "canonical distance" on $\mathcal{S}_{\text{fin}}(\mathcal{X})$ induced by the kernel k_0, namely

$$d_{\mathcal{E}}(S, S') = \sqrt{k_0(S, S) + k_0(S', S') - 2k_0(S, S')}. \quad (3)$$

Coming now to the proposed class of *Deep Embedding* kernels per se, these are obtained by composing what can be called a radial kernel on Hilbert space (See (Bachoc et al., 2018) for a reminder) with $d_{\mathcal{E}}$ above. We hence obtain DE kernels on $\mathcal{S}_{\text{fin}}(\mathcal{X})$ by writing

$$k_{\text{DE}}(S, S') = k_{\text{H}} \circ d_{\mathcal{E}}(S, S'), \quad (4)$$

with $k_{\text{H}} : [0, \infty) \to \mathbb{R}$ being such that $(h, h') \in \mathcal{H}^2 \to k_{\text{H}}(||h - h'||_{\mathcal{H}})$ is positive definite for any Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$. We establish next the positive definiteness of such kernels (See (Berg et al., 1984; Christmann and Steinwart, 2010) for similar constructions)

and further provide sufficient conditions for their strict positive definiteness on $\mathcal{S}_{\text{fin}}(\mathcal{X})$, a feature that k_0 is lacking, as we show too, which may lead to invertibility issues for finite $\mathcal{X}$, e.g. in combinatorial optimization.

2.3 Main Theoretical Results

Proposition 1. *Let $\mathcal{X}$ be a set, $k_{\mathcal{X}}$ be a positive definite kernel on $\mathcal{X}$ with associated reproducing kernel Hilbert space $\mathcal{H}_{k_{\mathcal{X}}}$, and $\mathcal{S}_{fin}(\mathcal{X})$ be the set of non-empty finite subsets of $\mathcal{X}$. Let $\mathcal{E} : S \in \mathcal{S}_{fin}(\mathcal{X}) \mapsto \mathcal{H}_{k_{\mathcal{X}}}$, $k_0 : \mathcal{S}_{fin}(\mathcal{X}) \times \mathcal{S}_{fin}(\mathcal{X}) \mapsto \mathbb{R}$, $d_{\mathcal{E}} : \mathcal{S}_{fin}(\mathcal{X}) \times \mathcal{S}_{fin}(\mathcal{X}) \mapsto [0, \infty)$ be defined by Equations 1,2,3, respectively. Then,*

a) *$k_0(S, S') = \langle \mathcal{E}(S), \mathcal{E}(S') \rangle_{\mathcal{H}_{k_{\mathcal{X}}}}$ for any $S, S' \in \mathcal{S}_{fin}(\mathcal{X})$, and k_0 is positive definite on $\mathcal{S}_{fin}(\mathcal{X})$ while $d_{\mathcal{E}}$ is a pseudometric on $\mathcal{S}_{fin}(\mathcal{X})$.*

Let us furthermore introduce for $n \geq 2$ the sets

$$A_n = \Bigg\{ \Bigg(\overbrace{\frac{1}{n_1}, \ldots, \frac{1}{n_1}}^{(n_1-\ell) \text{ times}}, \overbrace{\frac{n_2-n_1}{n_1 n_2}, \ldots, \frac{n_2-n_1}{n_1 n_2}}^{\ell \text{ times}}, \overbrace{\frac{-1}{n_2}, \ldots, \frac{-1}{n_2}}^{(n_2-\ell) \text{ times}} \Bigg), \\ n_1, n_2 \geq 1, \ell \geq 0 : n_1 + n_2 + \ell = n \Bigg\} \subset \mathbb{R}^n \ (n \geq 2).$$

b) *Then, the following assertions are equivalent:*

- **i)** *$k_{\mathcal{X}}$ satisfies $\sum_{i=1}^n \sum_{j=1}^n a_i a_j k_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j) > 0$ for all $n \geq 2$, pairwise distinct $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$, and $(a_1, \ldots, a_n) \in A_n$.*
- **ii)** *$\mathcal{E}$ is injective.*
- **iii)** *$d_{\mathcal{E}}$ is a metric on $\mathcal{S}_{fin}(\mathcal{X})$.*

In particular, if $k_{\mathcal{X}}$ is strictly positive definite on $\mathcal{X}$, then all three conditions above are fulfilled.

Proposition 2 (Non-strict positive definiteness of double sum kernels)**.** *Let us keep the notation of Proposition 1 and denote furthermore in the case of a finite set $\mathcal{X}$ with cardinality $c \geq 1$ and elements $\mathbf{X}_c = (\mathbf{x}_1, \ldots, \mathbf{x}_c)$ by $u : S \in \mathcal{S}_{fin}(\mathcal{X}) \to u(S) = \frac{1}{\#S}(\mathbf{1}_{\mathbf{x}_i \in S})_{1 \leq i \leq c} \in \mathbb{R}^c$ the mapping returning for any nonempty subset of $\mathcal{X}$ a vector with components $\frac{1}{\#S}$ or 0 depending whether $\mathbf{x}_i \in S$ or not. Then we have:*

a) *For $\mathcal{X}$ finite, for any $S, S' \in \mathcal{S}_{fin}(\mathcal{X})$,*

$$k_0(S, S') = u(S)^T k_{\mathcal{X}}(\mathbf{X}_c) u(S'). \tag{5}$$

Consequently, for $q \geq 1$ and $\mathbf{S} = (S_1, \ldots, S_q) \in \mathcal{S}^q$, the covariance matrix $k_0(\mathbf{S})$ associated with $k_{\mathcal{X}}$ and $\mathbf{S}$ can be compactly written as

$$k_0(\mathbf{S}) = U(\mathbf{S})^T k_{\mathcal{X}}(\mathbf{X}_c) U(\mathbf{S}), \tag{6}$$

with the notation $U(\mathbf{S}) = [u(S_1), \ldots, u(S_q)]$.

b) *For arbitrary $\mathcal{X}$, the two following assertions are mutually exclusive*

- **i)** *$\#\mathcal{X} = 1$ and $k_{\mathcal{X}}$ is non-zero.*
- **ii)** *k_0 is not strictly positive definite on $\mathcal{S}_{fin}(\mathcal{X})$.*

Proposition 3 ((Strict) positive definiteness of k_{DE})**.** *Let us consider here again the notation of Proposition 1 and consider furthermore the class of kernels $k_{DE} : (S, S') \in \mathcal{S}_{fin}(\mathcal{X}) \to k_H \circ d_{\mathcal{E}}(S, S')$ of Eq. 4, where $k_H : [0, \infty) \to \mathbb{R}$ is chosen such that $(h, h') \in \mathcal{H}^2 \to k_H(||h - h'||_{\mathcal{H}})$ is positive definite for any Hilbert space $(\mathcal{H}, \langle \cdot, \cdot, \rangle_{\mathcal{H}})$. Then,*

a) *k_{DE} is positive definite on $\mathcal{S}_{fin}(\mathcal{X})$.*

b) *If furthermore $k_{\mathcal{X}}$ satisfies **i)** of condition **b)** in Proposition 1, and $k_H : [0, \infty) \to \mathbb{R}$ is chosen such that $(h, h') \in \mathcal{H}^2 \to k_H(||h - h'||_{\mathcal{H}})$ is strictly positive definite for any Hilbert space $(\mathcal{H}, \langle \cdot, \cdot, \rangle_{\mathcal{H}})$, then k_{DE} is strictly positive definite on $\mathcal{S}_{fin}(\mathcal{X})$.*

Remark 1. *As mentioned in Bachoc et al. (2018), continuous functions inducing strictly positive definite functions on any Hilbert space can be characterized following Schoenberg's works both in terms of completely monotone functions and of infinite mixtures of squared exponential kernels (See, e.g., Wendland (2005)).*

2.4 Practicalities

In what follows and as in many practical situations, we consider "inner" (i.e., on $\mathcal{X}$) kernels of the form $k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \sigma^2_{\mathcal{X}} r_{\mathcal{X}}(\mathbf{x}, \mathbf{x}')$, where $\sigma^2_{\mathcal{X}} > 0$ and $r_{\mathcal{X}}$ is a (strictly) positive definite kernel on $\mathcal{X}$ taking the value 1 on the diagonal and parametrized by some (vector-valued or other) hyperparameter $\psi_{\mathcal{X}}$. In such a case, denoting $\mathcal{E}_{r_{\mathcal{X}}}(S) = \frac{1}{\#S} \sum_{\mathbf{x} \in S} r_{\mathcal{X}}(\mathbf{x}, \cdot)$ and $d_{\mathcal{E}_{r_{\mathcal{X}}}}$ the associated canonical distance, we immediately have that $\mathcal{E} = \sigma^2_{\mathcal{X}} \mathcal{E}_{r_{\mathcal{X}}}$ and $d_{\mathcal{E}} = \sigma_{\mathcal{X}} d_{\mathcal{E}_{r_{\mathcal{X}}}}$. As a consequence, if $k_{\text{H}}(\cdot)$ writes $\sigma^2_H r_H(\frac{\cdot}{\theta_H})$ for $\sigma^2_H, \theta_H > 0$ and $r_H(\cdot)$ defining a radial (strictly) positive definite kernel on any Hilbert space (possibly depending on some other hyperparameters ignored for simplicity) with $r_H(0) = 1$,

$$k_{\text{DE}}(S, S') = \sigma^2_H r_H \left(\frac{\sigma_{\mathcal{X}}}{\theta_H} d_{\mathcal{E}_{r_{\mathcal{X}}}}(S, S') \right),$$

and it clearly appears that having both $\sigma_{\mathcal{X}}$ and θ_H results in overparametrization of k_{DE}. For this reason, we adopt the convention that $\sigma_{\mathcal{X}} = 1$, hence remaining with the hyperparameters σ^2_H, θ_H and $\psi_{\mathcal{X}}$ to be fitted, possibly along with others such as trend and/or noise parameters. In our experiments, where noiseless settings and a constant trend are assumed, we appeal to Maximum Likelihood Estimation with concentration on the σ^2_H parameter and a genetic algorithm with

derivatives (Mebane Jr et al., 2011), in the flavour of the solution implemented in the DiceKriging R package (Roustant et al., 2012).

In the numerical experiments presented next, the base set $\mathcal{X}$ is assumed to be of the form $[0,1]^d$ (in our examples $d = 2$), and we choose for $r_{\mathcal{X}}$ an isotropic Gaussian correlation kernel solely parametrized by a "range" $\theta_{\mathcal{X}}$. As for r_H, while any kernel admissible in Hilbert space such as those of the Matérn family would be suitable, we also choose here a Gaussian for simplicity, hence ending up with a triplet of covariance hyperparameters, namely $(\sigma_H, \theta_H, \theta_{\mathcal{X}}) \in (0, +\infty)^3$. As σ_H^2 is taken care of by concentration (i.e. its optimal value for any given value of $\theta_H, \theta_{\mathcal{X}}$ can be analytically derived as a function of θ_H and $\theta_{\mathcal{X}}$), there remains to maximize the corresponding concentrated (a.k.a. profile) log-likelihood function with respect to θ_H and $\theta_{\mathcal{X}}$. For this purpose the analytical gradient of the concentrated log-likelihood with respect to these parameters has been calculated and implemented. Besides, parameter bounds need to be specified to the chosen optimization algorithm (i.e. *genoud*, here) and while it seems natural to choose bounds in terms of $\sqrt{d}$, the diameter of the unit d-dimensional hypercube, for θ_H the adequate diameter is slightly less straightforward and calls for some analysis with respect to the range of variation of $d_{\mathcal{E}_{r_{\mathcal{X}}}}$ and how it depends on $\theta_{\mathcal{X}}$. The next proposition establishes simple yet practically quite useful results regarding the diameter of $\mathcal{S}_r$ ($r > 0$) with respect to $d_{\mathcal{E}_{r_{\mathcal{X}}}}$ and its maximal value when letting $\theta_{\mathcal{X}}$ vary.

Proposition 4. *Let $r_{\mathcal{X}}$ be an isotropic positive definite kernel on $\mathcal{X} = [0,1]^d$ assumed to be monotonically decreasing to 0 with respect to the Euclidean distance between elements of $\mathcal{X}$, with range parameter $\theta_{\mathcal{X}} > 0$. Then the $d_{\mathcal{E}_{r_{\mathcal{X}}}}$-diameter of $\mathcal{S}_p(\mathcal{X})$ ($p \geq 1$), i.e. $\sup_{S,S' \in \mathcal{S}_p} d_{\mathcal{E}_{r_{\mathcal{X}}}}(S, S')$, is reached with arguments $\{\mathbf{0}_d, \ldots, \mathbf{0}_d\}$ and $\{\mathbf{1}_d, \ldots, \mathbf{1}_d\}$, where $\mathbf{0}_d = (0, \ldots, 0), \mathbf{1}_d = (1, \ldots, 1) \in \mathcal{X}$. Furthermore, the supremum of this diameter with respect to $\theta_{\mathcal{X}} \in (0, +\infty)$ is given by $\sqrt{2}$.*

3 Applications

We now demonstrate the applicability of the class of DE kernels for both prediction and optimization purposes, with comparisons when applicable to similar methods based on DS kernels, and also to random search in the optimization case. In all examples, both inner and outer kernels (resp. $k_{\mathcal{X}}$ and k_{H}) are assumed Gaussian. The three hyperparameters $(\sigma_H, \theta_H, \theta_{\mathcal{X}})$ are estimated by Maximum Likelihood with concentration on σ_H^2, as detailed in Section 2.4. Three synthetic test functions and two application test cases are considered, respectively in mechanical engineering (CASTEM) and in hydrogeology (Contaminant source localization), all presented below. In the CASTEM case, the available data set consists of a fixed number (404) of (set input)-output instances, while in the other test cases one may boil down to a similar situation by restricting the scope to finitely many such instances. Yet, the hydrogeology test case is the only one where $\mathcal{X}$ is structurally restricted to remain finite, here a set of 25 possible well locations, hence leading to a combinatorial optimization problem.

3.1 Presentation of Test Functions and Cases

3.1.1 Synthetic Functions

Our three synthetic test functions consist of extensions of the Branin-Hoo test function (See, e.g., Roustant et al., 2012), denoted below by g, for set-valued inputs. These extensions are based respectively on the maximum (MAX), minimum (MIN), and mean (MEAN) of g values associated with each of $p = 10$ evaluation points in $\mathcal{X} = [0,1]^2$, leading to

$$f(S) = \max_{\mathbf{x} \in S} g(\mathbf{x}) \tag{7}$$

$$f(S) = \min_{\mathbf{x} \in S} g(\mathbf{x}) \tag{8}$$

$$f(S) = \frac{1}{\#S} \sum_{\mathbf{x} \in S} g(\mathbf{x}), \tag{9}$$

where $S \in \mathcal{S}_p = ([0,1]^2)^{10}$. Let us remark that by design, the f of Eq. 9 is well-suited to be approximated using the double sum kernel of Eq. 2. Indeed, if g is assumed to be a draw of a GP with kernel $k_{\mathcal{X}}$, then f is a draw of a GP with kernel $\frac{1}{\#S}\frac{1}{\#S'}\sum_{\mathbf{x} \in S, \mathbf{x} \in S'} k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}')$, as numerical results of Sections 3.2 and 3.3 do reflect.

3.1.2 CASTEM Simulations

The CASTEM dataset, inherited from (Ginsbourger et al., 2016), was originally generated from mechanical simulations performed using the Cast3m code (Castem, 2016) to compute equivalent stress values on biphasic material subjected to uni-axial traction. The unit-square represents a matrix material containing 10 circular inclusions with identical radius of $R = 0.056419$. The dataset consists of 404 point-sets along with their corresponding stress levels. Fig. 1 illustrates two (set input)-output instances from it. While the goal pursued in (Ginsbourger et al., 2016) was rather in uncertainty propagation, we consider this data set here also from an optimization perspective.

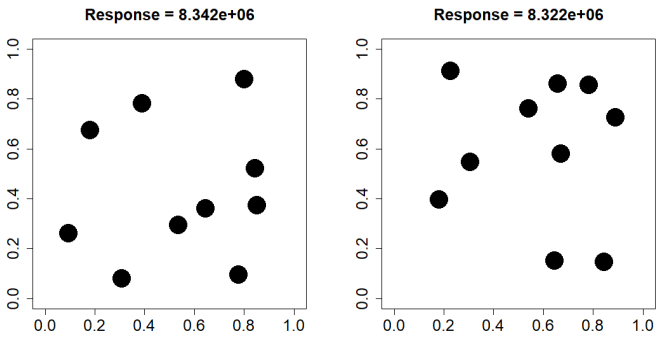

Figure 1: Two CASTEM (set input)-output instances

3.1.3 Selection of Monitoring Wells for Contaminant Source Localization

This test case relies on a benchmark generator of groundwater contaminant source localization problems from (Pirot et al., 2019). The original problems consisted in finding among given candidate source localizations $\mathbf{x}_i \in \mathbb{R}^2$ $(1 \leq i \leq 2601)$ which globally minimizes some measures of misfit between "reference" (or "observed") and "simulated" contaminant concentrations at fixed times and monitoring wells such as

$$g(\mathbf{x}, S) = \left(\sum_{i \in S} \sum_{t=1}^{T} |c_{\text{obs}}(i,t) - c_{\text{sim}}(\mathbf{x}, i, t)|^2 \right)^{\frac{1}{2}}, \quad (10)$$

where $c_{\text{obs}}(i,t)$ is the reference concentration at well i and time step t, $c_{\text{sim}}(\mathbf{x}, i, t)$ is the corresponding simulated concentration when the source of contaminant is at $\mathbf{x}$, and $S \subset S_{\text{full}} := \mathcal{X} = \{1, 2, \ldots, 25\}$ is a given subset from 25 fixed monitoring wells.

Here, instead of fixing the subset of well locations S and looking for the optimal $\mathbf{x}$, we consider instead the maps of score discrepancies $g(\cdot, S_{\text{full}}) - g(\cdot, S)$ as a function of S. From there, the considered combinatorial optimization problem consists in minimizing

$$f(S) = \sum_{i=1}^{2601} (g(\mathbf{x}_i, S_{\text{full}}) - g(\mathbf{x}_i, S))^2 \quad (11)$$

over the set $\mathcal{S}_p(\mathcal{X})$ of subsets of $p < 25$ wells from $\mathcal{X}$. In the numerical experiments, we fix $p = 5$, and hence the cardinality of the considered set of subsets $\mathcal{S}_5(\mathcal{X})$ is $\binom{25}{5} = 53,130$. To test the efficiency of our approach on this application, the two contaminant source locations (A and B) and two geological geometries of (Pirot et al., 2019) are considered, leading to four cases (denoted (Src A, Geo 1), (Src A, Geo 2), (Src B, Geo 1), (Src B, Geo 2), respectively).

Since the base set $\mathcal{X} = \{1, 2, \ldots, 25\}$ is itself finite here, it follows from Prop. 2 that resulting double sum kernels are not strictly positive definite so that BO with those kernels fails after few iterations, as found in numerical experiments. Two subsets of five well locations are plotted in Fig. 2 along with contours of corresponding score discrepancy maps $g(\cdot, S_{\text{full}}) - g(\cdot, S)$ and values of objective function f derived from them.

The first combination (left subfigure) better represents the misfit function $g(\cdot, S_{\text{full}})$ overall with a lower f value. This subset is in fact the optimal one, obtained by exhaustive search over all $53,130$ candidates. Our goal is precisely to efficiently locate by BO these optimal well locations whose contributions minimize the spatial sum of score discrepancies. The reader is referred to (Pirot et al., 2019) for further details and visualization of the misfit function, location of the contaminant source, and coordinates of well locations.

3.2 Prediction: Settings and Results

To assess the predictive ability of the considered GP models under the considered settings of data sets split into learning and test parts, we appeal to the so-called Q^2 or "predictive coefficient" (Marrel et al., 2008),

$$Q^2 = 1 - \frac{\sum_{i=1}^{n_{\text{test}}} (f(S_i^{(\text{test})}) - m_n(S_i^{(\text{test})}))^2}{\sum_{i=1}^{n_{\text{test}}} (f(S_i^{(\text{test})}) - \bar{\mathbf{f}})^2}, \quad (12)$$

where n_{test} is the number of test point-sets, $f(S_i^{(\text{test})})$ and $m_n(S_i^{(\text{test})})$ are the actual response and the mean values predicted by the GP model, respectively. $\bar{\mathbf{f}}$ is the mean of $f(S_i^{(\text{test})})$'s. The closer to 1 the value of Q^2, the more efficient the predictor is. In addition, we also look at visual diagnostics based on the comparison of standardized residuals (i.e. divided by GP prediction standard deviations) with the normal distribution, both in cross- and external validation.

As a result of Prop. 2, the DS kernel is not readily applicable for the contaminant source localization test case, due to singularity issues with covariance matrices. One way around this is to add a small positive jitter to their diagonal (inspired by Ranjan et al., 2011). This approach will be referred to hereafter as DS+j whenever it is used in place of the original DS. More detail on the procedure used for jitter tuning and additional results can be found in supplementary material.

The total size of datasets used to assess prediction performances for the three synthetic test problems, CASTEM, and the contamination applications are 1000, 404, and 200, respectively. Each dataset is further partitioned into training and testing sub-datasets with percentages (80:20), (50:50) and (20:80). Average Q^2 values over 20 replications are provided in Table 1. First, we observe that Q^2 tends to increase with the proportion of the full data set used for training, except

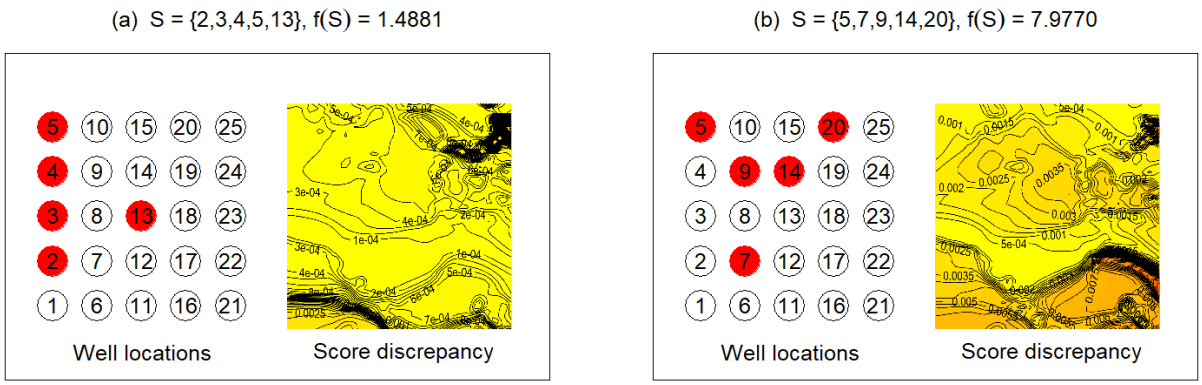

Figure 2: Score discrepancy map: location of selected wells (input S), score discrepancy landscape, and the spatial sum of score discrepancy objective function value $f(S)$.

in one case with CASTEM. We see that the proposed approach with the DE kernel gives higher value of Q^2 than that with the DS kernel on all problems except for the MEAN function. We hypothesize the latter to be due to the adequacy between the MEAN function's nature and the DS kernel, as remarked earlier.

Finally, Fig. 3 shows leave-one-out (left panel) and out-of-sample diagnostics (right panel) for the source localization application (Src A, Geo 1) with DE kernel. The results show relatively moderate departures from the normality assumptions. Complete residual analysis for all scenarios as well as for DS kernels (with jitter) can be found in supplementary material.

3.3 Optimization: Settings and Results

In this section, the efficiency of DE versus DS kernels (possibly with jitter) are evaluated within the BO framework, using the Expected Improvement (EI) (Mockus et al., 1978) as infill sampling criterion. To assess optimization performances, the same datasets as those used in previous section are used for the three synthetic problems and CASTEM. As for the contaminant source application, the whole dataset of size $53,130$ is employed. Optimization performances are assessed on 50 repetitions of EI algorithms with 10 initial design point-sets. For each repetition, all algorithms start with the same initial design, and are allocated 40 additional objective function evaluations. The hyperparameters are iteratively re-determined in every iteration using MLE (See Section 2.4 and supplementary material).

Concerning EI maximization, EI values are computed at all point-sets and the one attaining the highest value is selected (no ties occurred). The performance is measured by (1) counting the number of trials (out of 50) for which the algorithm could find the best point from the considered dataset; and (2) monitoring the distribution of best found responses over iterations. A random sampling method is used as baseline. Table 2 summarizes the number of trials that the minimum is found and Fig. 4 represents progress curves in terms of median and 95th percentile values of current best objective function values over 50 trials.

EI algorithms with any of the two considered kernel classes clearly appear here superior to random sampling. Experiments on synthetic problems show that within the two considered EI algorithm settings, DE kernels outperform DS ones on the MAX problem both in terms of the number of trials that the true minimum is found and of the final best responses. On the MEAN problem, though, while both approaches lead to locate the minimum for all 50 replications, DS kernels lead to a fewer number of iterations as anticipated due to adequacy between this kernel class and the test function. EI algorithms with both kernels did not perform well on the MIN problem which may be explained by the fact that the underlying Branin-Hoo function has the large portion of the search space being quite flat. For the CASTEM dataset, EI-k_{DE} and EI-k_0 methods could locate the minimum for 28 and 10 trials, respectively, against 5 for random sampling.

As for the source localization application, the obtained EI-k_0 results are all involving the use of a jitter. Overall, EI algorithms coupled with either of the two kernel classes appeared by far better than random sampling. Comparing performances between the two EI algorithms, EI-k_{DE} method could locate the global optimum more frequently (as indicated in Table 2). In particular, with the DE kernel, the EI algorithm found

Table 1: Q^2 values for GP predictions on all test cases with DE versus DS kernels (k_{DE} versus k_0(+j))

Problem	k_{DE}			k_0		
	20:80	**50:50**	**80:20**	**20:80**	**50:50**	**80:20**
(a) MAX	0.6926	0.8011	0.8559	0.5644	0.7429	0.7725
(b) MEAN	0.9996	0.9999	∼1	∼1	∼1	∼1
(c) MIN	0.3309	0.4582	0.4929	0.1080	0.2245	0.2749
(d) CASTEM	0.5799	0.6641	0.6543	0.5067	0.5410	0.5056
Problem	k_{DE}			k_0+j		
	20:80	**50:50**	**80:20**	**20:80**	**50:50**	**80:20**
(e) (Src A, Geo 1)	0.7607	0.9133	0.9352	0.7437	0.8445	0.8804
(f) (Src A, Geo 2)	0.7239	0.8855	0.9240	0.7130	0.8485	0.8729
(g) (Src B, Geo 1)	0.7977	0.9190	0.9447	0.7901	0.8746	0.8904
(h) ()Src B, Geo 2)	0.8486	0.9151	0.9439	0.8389	0.8944	0.9252

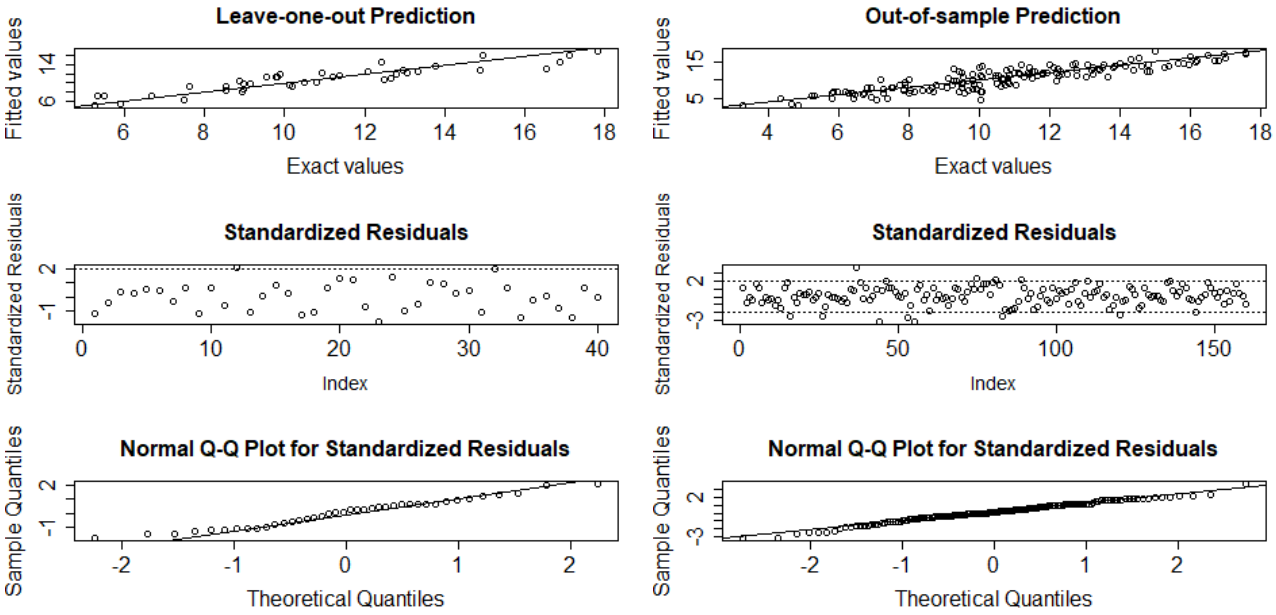

Figure 3: GP prediction residual analysis on the contaminant source localization problem (Src A, Geo 1) with kernel k_{DE} and ratio (20:80). (a) Internal errors (left); (b) External errors (right).

Table 2: Numbers of trials (out of 50) for which the minimum is found for EI algorithms based on GP models with DE versus DS kernels, as well as for Random Sampling.

Problem	**EI-k_{DE}**	**EI-k_0**	**RANDOM**
(a) MAX	36	8	6
(b) MEAN	50	50	4
(c) MIN	9	8	3
(d) CASTEM	28	10	5
Problem	**EI-k_{DE}**	**EI-k_0+j**	**RANDOM**
(e) (Src A, Geo 1)	50	48	0
(f) (Src A, Geo 2)	34	25	0
(g) (Src B, Geo 1)	50	47	0
(h) (Src B, Geo 2)	43	44	0

the global optimum in every trial run on two scenarios of contaminant source localization problems (i.e. (Src A, Geo 1) and (Src B, Geo 1)).

The median progress curves (bottom panel of Fig. 4) illustrate on the other hand that the DS kernel seem quite well-suited for the contaminant problems, as highlighted in particular by the fast initial decrease in best response value. The 95% quantile curves suggest however that in the worst situations, EI-k_{DE} per-

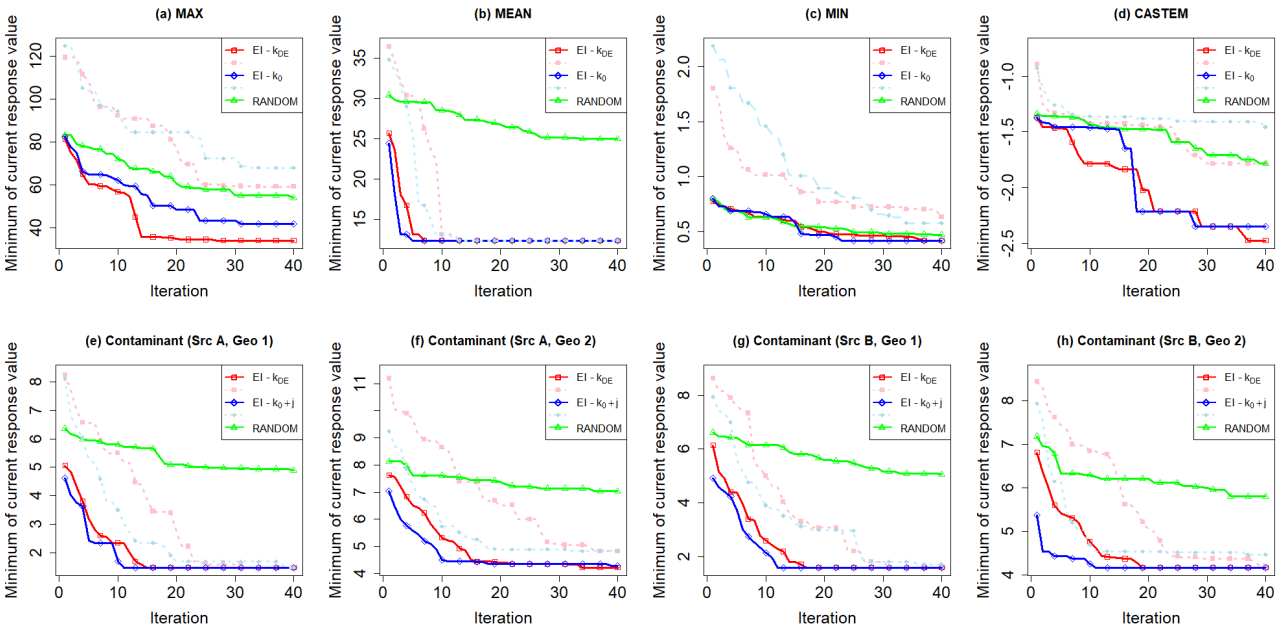

Figure 4: BO progress curves showing the median (solid lines) and 95th percentile (dotted lines) values of the current best response of problems (a) MAX, (b) MEAN, (c) MIN, (d) CASTEM, and contaminant problems (e) (Src A, Geo 1), (f) (Src A, Geo 2), (g) (Src B, Geo 1) and (h) (Src B, Geo 2).

forms relatively better and seems to be more robust especially toward the end of the course when the jitter was needed to make EI-k_0 work. It is worth noting that determining an appropriate jitter level to add to the DS kernel is not a straightforward task. While one would want to add a smallest possible value of jitter, oftentimes, a too small jitter is not enough to fix conditioning issues. Additional results, with a large number of trials, revealing the effect of a poor choice of jitter level on DS kernel model's accuracy as well as optimization results are given in supplementary material. Overall, the strict positive definiteness of considered DE kernels (and the fact that no jitter is required) make them appear as a relatively robust option to efficiently address expensive combinatorial optimization problems in a "black-box" Bayesian Optimization framework (i.e., without requiring much prior knowledge about the problem structure).

4 Discussion

Experimental results obtained on the analytical objective functions and application test cases confirm the added value of the considered approaches for set-function prediction and (combinatorial) optimization.

Yet a number of challenges and potential extensions remain to be addressed in future work. This includes computational difficulties that will arise when working with larger numbers of subsets and/or subset cardinalities, not only to handle bigger matrices but also to tackle the optimization of infill criteria. These criteria include the Expected Improvement as well as adaptations of further families of BO acquisition functions from frameworks such as Predictive Entropy Search (Hernández-Lobato et al., 2014), Knowledge Gradient (Frazier, 2018), and others.

From the test case perspective, future work may also include tackling further prediction and subset selection problems (be it in continuous or combinatorial settings, with problem structures of various levels of complexity), not only for optimization purposes but also with more general goals around uncertainty quantification and reduction (Bect et al., 2019). Besides this, a nice feature of the considered approaches is that they would naturally extend to cases with varying subset cardinalities and also with "marked" point sets (in the vein of (Cuturi et al., 2005)'s molecular measures), hence accommodating applications such as CASTEM but with varying inclusion numbers and radii. Furthermore, the conceptual approach of chaining an embedding and a kernel in Hilbert space (also in the flavour of (Christmann and Steinwart, 2010)) could apply to a variety of other input types provided that relevant mappings to Hilbert space can be found, opening the door to numerous non-conventional extensions of GP-based prediction, BO, and related kernel methods.

Acknowledgements

The authors would like to thank the anonymous referees for constructive comments having lead to substantial improvements of the paper. P.B. would like to thank DPST scholarship project granted by IPST, Ministry of Education, Thailand for providing financial support during his master study. D.G.'s contributions have taken place within the Swiss National Science Foundation project number 178858. Furthermore, D.G. would like to thank colleagues including notably Fabrice Gamboa, Athénaïs Gautier, Luc Pronzato, Henry Wynn, and Anatoly Zhigljavsky for enriching discussions in recent years around ideas presented in this paper. T.K. would like to acknowledge the support of Thailand Research Fund under Grant No.: MRG6080208, Centre of Excellence in Mathematics, CHE, Thailand, and the Faculty of Science, Mahidol University. The authors would like to acknowledge the support of Idiap Research Institute. In particular, most numerical experiments presented here were run on Idiap's grid. The authors also thank Drs. Jean Baccou and Frédéric Perales (Institut de Radioprotection et de Sûreté Nucléaire, Saint-Paul-lès-Durance, France) for the CASTEM data, and Dr. Clément Chevalier who has been involved in investigations on this data in the framework of the ReDICE consortium.

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Transaction of the American Mathematical Society*, 68 (3):337 – 404.

Bachoc, F., Suvorikova, A., Ginsbourger, D., Loubes, J.-M., and Spokoiny, V. (2018). Gaussian processes with multidimensional distribution inputs via optimal transport and hilbertian embedding. *arXiv preprint arXiv:1805.00753*.

Baptista, R. and Poloczek, M. (2018). Bayesian optimization of combinatorial structures. In *Proceedings of the 35th International Conference on Machine Learning*.

Bect, J., Bachoc, F., and Ginsbourger, D. (2019). A supermartingale approach to Gaussian process based sequential design of experiments. *Bernoulli*, 25(4A):2883–2919.

Berg, C., Christensen, J., and Ressel, P. (1984). *Harmonic Analysis on Semigroups*. Springer.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers.

Binois, M., Huang, J., Gramacy, R., and Ludkovski, M. (2019). Replication or exploration? Sequential design for stochastic simulation experiments. *Technometrics*, 61(1):7–23.

Castem (2016). Cast3m software, http://www-cast3m.cea.fr.

Christmann, A. and Steinwart, I. (2010). Universal kernels on non-standard input spaces. In *Advances in neural information processing systems*, pages 406–414.

Cuturi, M., Fukumizu, K., and Vert, J. (2005). Semigroup kernels on measures. *Journal of Machine Learning Research*, 6:1169–1198.

Desobry, F., Davy, M., and Fitzgerald, W. (2005). A class of kernels for sets of vectors. In *In Proceedings of the 13th European Symposium on Artificial Neural Networks*.

Fortuin, V., Dresdner, G. Strathmann, H., and Rätsch, G. (2018). Scalable gaussian processes on discrete domains. arXiv:1810.10368.

Frazier, P. (2018). A tutorial on bayesian optimization. arXiv:1807.02811.

Garnett, R., Osborne, M. A., and Roberts, S. J. (2010). Bayesian optimization for sensor set selection. In *Proceedings of the 9th ACM/IEEE international conference on information processing in sensor networks*, pages 209–219. ACM.

Garrido-Merchan, E. and Hernández-Lobato, D. (2020). Dealing with categorical and integer-valued variables in bayesian optimization with gaussian processes. *Neurocomputing*, 380. arXiv:1805.03463.

Gärtner, T., Flach, P. A., Kowalczyk, A., and Smola, A. J. (2002). Multi-instance kernels. In *Proceedings of the International Conference on Machine Learning*.

Ginsbourger, D., Baccou, J., Chevalier, C., and Perales, F. (2016). Design of computer experiments using competing distances between set-valued inputs. In *mODa 11-Advances in Model-Oriented Design and Analysis*, pages 123–131. Springer.

Gramacy, R. B. and Taddy, M. A. (2010). Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp version 2, an r package for treed gaussian process models. *Journal of Statistical Software*, 33(6).

Grauman, K. and Darrell, T. (2007). The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, 8:725–760.

Griffiths, R.-R. and Hernández-Lobato, J. M. (2019). Constrained bayesian optimization for automatic chemical design. arXiv:1709.05501.

Haussler, D. (1999). Convolution kernels on discrete structures. Technical report, University of California at Santa Cruz, Department of Computer Science.

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. In *Neural Information Processing Systems*.

Iwata, K. (2012). Placing landmarks suitably for shape analysis by optimization. In *21st International Conference on Pattern Recognition*.

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

Kim, J., McCourt, M., You, T., Kim, S., and Choi, S. (2019). Bayesian optimization over sets. In *6th ICML Workshop on Automated Machine Learning*.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41(2):495–502.

Kondor, R. and Jebara, T. (2003). A kernel between sets of vectors. In *Proceedings of the Twentieth International Conference on Machine Learning*.

Kondor, R. and Lafferty, J. (2002). Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th International Conference on Machine Learning*, page 315–322.

Larraiiaga, P., Etxeberria, R., Lozano, J., and Peiia, J. (2000). Combinatorial optimization by learning and simulation of bayesian networks. In *Uncertainty in Artificial Intelligence Proceedings*.

Marrel, A., Iooss, B., van Dorpe, F., and Volkova, E. (2008). An efficient methodology for modeling complex computer codes with gaussian processes. *Computational Statistics and Data Analysis*.

Mebane Jr, W. R., Sekhon, J. S., et al. (2011). Genetic optimization using derivatives: the rgenoud package for r. *Journal of Statistical Software*, 42(11):1–26.

Mockus, J., Tiesis, V., and Zilinskas, A. (1978). The application of bayesian methods for seeking the extremum. vol. 2.

Muandet, K., Fukumizu, K., and B., S. (2017). Kernel mean embedding of distributions : A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141.

Oh, C., Tomczak, J., Gavves, E., and Welling, M. (2019). Combo: Combinatorial bayesian optimization using graph representations. In *ICML Workshop on Learning and Reasoning with Graph-Structured Data*.

Pappas, N. and Popescu-Belis, A. (2017). Explicit document modeling through weighted multiple-instance learning. *Journal of Artificial Intelligence Research*, 58.

Pirot, G., Krityakierne, T., Ginsbourger, D., and Renard, P. (2019). Contaminant source localization via bayesian global optimization. *Hydrology and Earth System Sciences*, 23(1):351–369.

Ranjan, P., Haynes, R., and Karsten, R. (2011). A computationally stable approach to gaussian process interpolation of deterministic computer simulation data. *Technometrics*, 53(4):366–378.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian process for machine learning*. MIT press.

Risk, J. and Ludkovski, M. (2018). Sequential design and spatial modeling for portfolio tail risk measurement. *SIAM Journal on Financial Mathematics*, 9(4):1137–1174.

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.

Roustant, O., Padonou, E., Deville, Y., Clémet, A., Perrin, G., Giorla, J., and Wynn, H. (2018). Group kernels for gaussian process metamodels with categorical inputs. arXiv:1802.02368.

Ru, B., Alvi, A., Nguyen, V., Osborne, M. A., and Roberts, S. (2019). Bayesian optimisation over multiple continuous and categorical inputs. In *3rd Workshop on Meta-Learning at NeurIPS 2019, Vancouver, Canada*.

Saitoh, S. and Sawano, Y. (2016). *Theory of Reproducing Kernels and Applications*. Springer.

Salemi, P. L., Song, E., Nelson, B., and Staum, J. (2019). Gaussian markov random fields for discrete optimization via simulation: Framework and algorithms. *Operations Research*, 67:250–266.

Schölkopf, B. and Smola, A. (2002). *Learning with kernels*. MIT Press.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A hilbert space embedding for distributions. In *Algorithmic Learning Theory: 18th International Conference*, page 13–31. Springer.

Sriperumbudur, B., Fukumizu, K., and Lanckriet, G. (2011). Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, (12):2389–2410.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer.

Wendland, H. (2005). *Scattered Data Approximation*. Cambridge University Press.