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

Poompol Buathong, David Ginsbourger, Tipaluck Krityakierne

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

Poompol Buathong*[1], David Ginsbourger*[2,3], and Tipaluck Krityakierne[1,4]

[1]Department of Mathematics, Faculty of Science, Mahidol University, Bangkok, Thailand
[2]Uncertainty Quantification and Optimal Design group, Idiap Research Institute, Centre du Parc, Rue Marconi 19, PO Box 592, CH-1920 Martigny, Switzerland.
[3]Institute of Mathematical Statistics and Actuarial Science, Department of Mathematics and Statistics, University of Bern, Alpeneggstrasse 22, CH-3012 Bern, Switzerland
[4]Centre of Excellence in Mathematics, CHE, Bangkok, Thailand

October 9, 2019

**Abstract**

We focus on kernel methods for set-valued inputs and their application to Bayesian set optimization, notably combinatorial optimization. We introduce a class of (strictly) positive definite kernels that relies on Reproducing Kernel Hilbert Space embeddings, and successfully generalizes "double sum" set kernels recently considered in Bayesian set optimization, which turn out to be unsuitable for combinatorial optimization. The proposed class of kernels, for which we provide theoretical guarantees, essentially consists in applying an outer kernel on top of the canonical distance induced by a double sum kernel. 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 illustrate the added value of the approach and open new perspectives in prediction and sequential design with set inputs.

*PB and DG contributed equally to this work and are in alphabetical order.

# 1 Introduction

Kernel methods (Aronszajn, 1950; Kimeldorf and Wahba, 1970; Schölkopf and Smola, 2002; Saitoh and Sawano, 2016) constitute a very 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 (Moćkus 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 BO and related contributions the focus is on continuous problems with vector-valued inputs, there has been a growing interest recently for GP-related modelling and BO in the presence of 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, 2018; 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 position of inclusion positions (Ginsbourger et al., 2016), inventory system optimization (Salemi et al., 2019), selection of starting centers in clustering algorithms (Kim et al., 2019), but also 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 that can also accomodate finite sets as a particular case (assuming a uniform measure, as implicitly done in the considered embedding approach). Kernels on finite sets that have been used in BO include to the best of our knowledge radial kernels with respect to the 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 a class that we refer to as "double sum" kernels (traced back to Gätner et al. (2002)) in (Kim et al., 2019). From the side of combinatorial optimization, 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 the 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 build a novel class of kernels on finite sets by chaining some "outer" kernels with the canonical (pseudo-)distances attached to the double sum kernels of (Gätner et al., 2002; Kim et al., 2019). We show that while restricting "inner" kernels to strictly positive definite ones does not lead to strictly positive definite double sum kernels, combining this assumption with another one guaranteeing strict positive definiteness in Hilbert space of the "outer" kernel (Bachoc et al., 2018) is sufficient for our proposed kernels to be strictly positive definite indeed, a crucial property in particular for combinatorial optimization. We present in turn a few additional results pertaining to the parametrization of this class of kernels and to the related hyperparameter fitting problem, including geometrical considerations around the choice of hyperparameter bounds in the embedding framework.

Section 2 is mainly dedicated to the construction and theoretical analysis of the considered classes of kernels, and complemented by practicalities regarding hyperparameter fitting. In Section 3, numerical experiments are discussed that compare double sum and proposed 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 $\mathcal{S}$, may be the whole power set $\mathcal{P}(\mathcal{X})$ or a subset thereof, e.g. $\mathcal{S}_p$ (also traditionally noted $[\mathcal{X}]^p$ in set theory) the set of subsets of $\mathcal{X}$ consisting of $p$ elements (where $p$ is assumed smaller than or equal to the cardinality of $\mathcal{X}$), or the set of all (non-void) finite subsets of $\mathcal{X}$ denoted here $\mathcal{S}_{\text{fin}} = \cup_{p \geq 1} \mathcal{S}_p$. Given a positive definite kernel $k_{\text{in}}$ over $\mathcal{X}$ and the associated Reproducing Kernel Hilbert Space $\mathcal{H}_{k_{\text{in}}}$, what we call the embedding of $\mathcal{S}_{\text{fin}}$ in $\mathcal{H}_{k_{\text{in}}}$ is the mapping

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

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

## 2.2 From Double Sum to Proposed Kernels

A natural idea to create a positive definite kernel on $\mathcal{S}_{\text{fin}}$ from this embedding is to plainly take the RKHS scalar product between embedded sets:

$$\begin{aligned} k(S, S') &= \langle \mathcal{E}(S), \mathcal{E}(S') \rangle_{\mathcal{H}_{k_{\text{in}}}} \\ &= \frac{1}{\#S} \frac{1}{\#S'} \sum_{\mathbf{x} \in S, \mathbf{x}' \in S'} k_{\text{in}}(\mathbf{x}, \mathbf{x}'), \end{aligned} \tag{2}$$

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 propose here. The first step is to consider the canonical (pseudo-)distance on $\mathcal{S}_{\text{fin}}$ induced by this $k$, namely

$$\begin{aligned} & d_{\mathcal{E}}(S, S') = ||\mathcal{E}(S) - \mathcal{E}(S')||_{\mathcal{H}_{k_{\text{in}}}} \\ & = \left( \frac{1}{(\#S)^2} \sum_{\mathbf{x} \in S} k_{\text{in}}(\mathbf{x}, \mathbf{x}) + \frac{1}{(\#S')^2} \sum_{\mathbf{x}' \in S'} k_{\text{in}}(\mathbf{x}', \mathbf{x}') \right. \\ & \left. - \frac{2}{(\#S)(\#S')} \sum_{\mathbf{x} \in S, \mathbf{x}' \in S'} k_{\text{in}}(\mathbf{x}, \mathbf{x}') \right)^{\frac{1}{2}}. \end{aligned} \tag{3}$$

Coming now to the proposed class of 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 another class of kernels on $\mathcal{S}_{\text{fin}}$ by writing

$$\begin{aligned} k(S, S') &= k_{\text{out}} \circ d_{\mathcal{E}}(S, S') \\ &= k_{\text{out}}(||\mathcal{E}(S) - \mathcal{E}(S')||_{\mathcal{H}_{k_{\text{in}}}}), \end{aligned} \tag{4}$$

with $k_{\text{out}} : r \in [0, \infty) \to \mathbb{R}$ being such that $(h, h') \in H \to k_{\text{out}}(||h - h'||_H)$ is positive definite for any Hilbert space $(H, \langle \cdot, \cdot, \rangle_H)$. The fact that kernels $k$ generated in this way are positive definite on $\mathcal{S}_{\text{fin}}$ follows directly from the positive definiteness of $k_{\text{out}}$ on the Hilbert space $\mathcal{H}_{k_{\text{in}}}$ and the representation of $d_{\mathcal{E}}(S, S')$ in terms of RKHS distance between the images of $S, S'$ by some mapping $\mathcal{E}$ (See (Berg et al., 1984; Christmann and Steinwart, 2010) for similar constructions). Furtheremore, as we develop next, we can ensure under some assumptions that such kernels will further be strictly positive definite on $\mathcal{S}_{\text{fin}}$, a feature that will turn out to be crucial in combinatorial optimization.

## 2.3 Main Theoretical Results

**Proposition 1** (Non-strict positive definiteness of double sum kernels)**.** *Let $\mathcal{X}$ be a set, $k_{in}$ be a positive definite kernel on $\mathcal{X}$, and $\mathcal{S}_{fin}$ be the set of finite subsets of $\mathcal{X}$. Then the kernel over $\mathcal{S}_{fin}$ defined by Eq. 2 is positive definite. However, even if $k_{in}$ is strictly positive definite on $\mathcal{X}$, such a $k$ will generally not be strictly positive definite on $\mathcal{S}_{fin}$ unless $\mathcal{X}$ is a singleton.*

*Proof of Prop. 1.* Let us consider here the case where the base set $\mathcal{X}$ is finite with cardinality $c$, so that $\mathcal{S}_{\text{fin}} = \mathcal{P}(\mathcal{X})\backslash\{\emptyset\}$. Since $\mathcal{X}$ is finite, a positive definite kernel $k_{\text{in}}$ on $\mathcal{X}$ boils down to a $c \times c$ Gram matrix, say $K_{\text{in}} = (k_{\text{in}}(\mathbf{x}_i, \mathbf{x}_j))_{i,j\in\{1,\dots,c\}}$ where $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_c\}$. By reformulating Eq. 2 in terms of a bilinear form with respect to a specific vector function of $S$ and $S'$, we will not only revisit the proof that $k$ is indeed positive definite but also establish that it is generally not strictly positive definite. Let us define for that purpose

$$u(S) = \frac{1}{\#S}(\mathbb{1}_{\mathbf{x}_i\in S})_{1\leq i\leq c} \in \mathbb{R}^c.$$

From there $k(S, S')$ can be reformulated into

$$k(S, S') = u(S)^T K_{\text{in}} u(S'), \tag{5}$$

so that, for any $q \geq 1$, $\alpha_1, \dots, \alpha_q \in \mathbb{R}$, and $\mathbf{S} = (S_1, \dots, S_q) \in \mathcal{S}^q$,

$$\begin{aligned}&\sum_{i=1}^{q}\sum_{j=1}^{q} \alpha_i\alpha_j k(S_i, S_j)\\ &= \left(\sum_{i=1}^{q} \alpha_i u(S_i)\right)^T K_{\text{in}} \left(\sum_{i=1}^{q} \alpha_i u(S_i)\right) \geq 0,\end{aligned}$$

by positive semi-definiteness of $K_{\text{in}}$, hence implying that $k$ is p.d. indeed. Yet, this representation will allow us to shed light on the fact that even if $K_{\text{in}}$ is a positive definite matrix (i.e., that $k_{\text{in}}$ is strictly p.d. on $\mathcal{X}$), the matrix $K_{\mathbf{S}} = (k(S_i, S_j))_{i,j\in\{1,\dots,q\}}$ will actually be systematically singular for $q > c$ and even sometimes in cases where $q \leq c$. Exploiting Eq. 5 and defining $U_{\mathbf{S}} = [u(S_1), \dots, u(S_q)]$, we get $K_{\mathbf{S}}$ in product form as follows

$$K_{\mathbf{S}} = U_{\mathbf{S}}^T K_{\text{in}} U_{\mathbf{S}}.$$

From there we get that $K_{\mathbf{S}} = M_{\mathbf{S}}^T M_{\mathbf{S}}$ with $M_{\mathbf{S}} = K_{\text{in}}^{\frac{1}{2}} U_{\mathbf{S}}$ and so $\text{rank}(K_{\mathbf{S}}) = \text{rank}(M_{\mathbf{S}})$. Now, the rank of a matrix being invariant under pre (or post) multiplication by a non singular matrix, we get in the case of a non-singular $K_{\text{in}}$ that $\text{rank}(K_{\mathbf{S}}) = \text{rank}(U_{\mathbf{S}}) \leq \min(q, c)$. Hence for $q > c$, $\text{rank}(K_{\mathbf{S}}) \leq c < q$ and the matrix of interest is singular. To see that $\text{rank}(K_{\mathbf{S}})$ can be singular when $q \leq c$ even in the case of an invertible $K_{\text{in}}$, one can think for instance of the situation where $c = 5, q = 4$ and $U_{\mathbf{S}} \propto \begin{pmatrix} 1 & 1 & 0 & 0 & 1\\ 0 & 0 & 1 & 1 & 1\\ 1 & 0 & 0 & 1 & 1\\ 0 & 1 & 1 & 0 & 1 \end{pmatrix}$. □

**Proposition 2** (Distance over $\mathcal{S}_{\text{fin}}$ induced by RKHS embedding)**.** *If $k_{in}$ is strictly positive definite over $\mathcal{X}$, then $\mathcal{E}$ is injective and $d_{\mathcal{E}} : \mathcal{S}_{fin} \times \mathcal{S}_{fin} \to [0, \infty)$ defined by Eq.3 defines a distance over $\mathcal{S}_{fin}$.*

*Proof of Prop. 2.* $d_{\mathcal{E}}(S, S') = ||\mathcal{E}(S) - \mathcal{E}(S')||_{\mathcal{H}_{k_{\text{in}}}}$ straightforwardly inherits from the triangle inequality associated with $\mathcal{H}_{k_{\text{in}}}$'s distance. The only point to check is the definiteness, i.e. that $d_{\mathcal{E}}(S, S') \neq 0$ for $S \neq S'$. Yet, assuming

$d_{\mathcal{E}}(S, S') = 0$ with $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ and $S' = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_{m'}\}$ for some $m, m' \geq 1$ and $\mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{x}'_1, \ldots, \mathbf{x}'_m \in \mathcal{X}$ would imply that

$$\left|\left| \frac{1}{m} \sum_{i=1}^{m} k_{\text{in}}(\mathbf{x}_i, \cdot) - \frac{1}{m'} \sum_{i=1}^{m'} k_{\text{in}}(\mathbf{x}'_i, \cdot) \right|\right|^2_{\mathcal{H}_{k_{\text{in}}}} = 0.$$

After regrouping terms when necessary, one would arrive at an equality of the kind $||\sum_{i=1}^{\widetilde{m}} \widetilde{a}_i k_{\text{in}}(\widetilde{\mathbf{x}}_i, \cdot)||^2_{\mathcal{H}_{k_{\text{in}}}} = 0$ (with, for instance in the case $S \cap S' = \emptyset$, $\widetilde{m} = m + m'$, $\widetilde{\mathbf{x}}_i = \mathbf{x}_i$ and $\widetilde{a}_i = \frac{1}{m}$ for $1 \leq i \leq m$ while $\widetilde{\mathbf{x}}_i = \mathbf{x}_{m+i}$ and $\widetilde{a}_i = -\frac{1}{m'}$ for $m + 1 \leq i \leq m + m'$), which would be in contradiction with $k_{\text{in}}$'s strict positive definiteness as soon as $S \neq S'$, proving in turn that $\mathcal{E}$ is injective. □

**Proposition 3** (Strict positive definiteness of $k$ )**.** *If $k_{in}$ is strictly positive definite over $\mathcal{X}$ and $k_{out} : [0, +\infty) \to \mathbb{R}$ is such that $(h, h') \in H^2 \to k_{out}(||h - h'||_H)$ is strictly positive definite for any Hilbert space $H$, then $k$ of Eq. 4 is strictly positive definite over $\mathcal{S}_{fin}$.*

*Proof of Prop. 3.* $(h, h') \in H^2 \to k_{\text{out}}(||h - h'||_{\mathcal{H}_{k_{\text{in}}}})$ is strictly positive definite on $\mathcal{H}_{\text{in}}$ by assumption on $k_{\text{out}}$ and the fact that $\mathcal{H}_{\text{in}}$ is a particular Hilbert space. The strict positive definiteness of $k$ on $\mathcal{S}_{\text{fin}}$ then follows from the injectivity of $\mathcal{E}$ implied by the strict positive definiteness of $k_{\text{in}}$ (as established in Prop. 2). □

**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 kernels of the form $k_{\text{in}}(\mathbf{x}, \mathbf{x}') = \sigma^2_{\text{in}} r_{\text{in}}(\mathbf{x}, \mathbf{x}')$, where $\sigma^2_{\text{in}} > 0$ and $r_{\text{in}}$ is a (strictly) positive definite kernel on $\mathcal{S}_{\text{fin}}$ taking the value 1 on the diagonal and parametrized by some (vector-valued or other) hyperparameter $\psi_{\text{in}}$. In such a case, denoting $\mathcal{E}_{r_{\text{in}}}(S) = \frac{1}{\#S} \sum_{\mathbf{x} \in S} r_{\text{in}}(\mathbf{x}, \cdot)$ and $d_{\mathcal{E}_{r_{\text{in}}}}$ the associated canonical (pseudo-)distance, we immediately have that $\mathcal{E} = \sigma^2_{\text{in}} \mathcal{E}_{r_{\text{in}}}$ and $d_{\mathcal{E}} = \sigma_{\text{in}} d_{\mathcal{E}_{r_{\text{in}}}}$. As a consequence, if $k_{\text{out}}(\cdot)$ writes $\sigma^2_{\text{out}} r_{\text{out}}(\frac{\cdot}{\theta_{\text{out}}})$ for $\sigma^2_{\text{out}}, \theta_{\text{out}} > 0$ and $r_{\text{out}}(\cdot)$ defining a radial (strictly) positive definite kernel on Hilbert space (possibly depending on some other hyperparameters ignored for simplicity) with $r_{\text{out}}(0) = 1$,

$$k(S, S') = \sigma^2_{\text{out}} r_{\text{out}} \left( \frac{\sigma_{\text{in}}}{\theta_{\text{out}}} d_{\mathcal{E}_{r_{\text{in}}}}(S, S') \right),$$

and it clearly appears that having both $\sigma_{\text{in}}$ and $\theta_{\text{out}}$ creates some overparametrization of $k$. For this reason, we adopt the convention that $\sigma_{\text{in}} = 1$, hence remaining with the hyperparameters $\sigma^2_{\text{out}}$, $\theta_{\text{out}}$ and $\psi_{\text{in}}$ 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 $\sigma^2_{\text{out}}$ 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 $r_{\text{in}}$ an isotropic Gaussian correlation kernel with a unique range parameter denoted $\theta_{\text{in}}$. As for $r_{\text{out}}$, while any kernel admissible in Hilbert space such as those of the Matérn family would be suitable, we also choose here a Gaussian one for simplicity, and we hence end up with a triplet of covariance hyperparameters, namely $(\sigma_{\text{out}}, \theta_{\text{out}}, \theta_{\text{in}}) \in (0, +\infty)^3$. As $\sigma^2_{\text{out}}$ is taken care of by concentration (i.e. its optimal value for any given value of $\theta_{\text{out}}, \theta_{\text{in}}$ can be analytically derived as a function of $\theta_{\text{out}}$ and $\theta_{\text{in}}$), there remains to maximize the corresponding concentrated (a.k.a. profile) log-likelihood function with respect to $\theta_{\text{out}}$ and $\theta_{\text{in}}$. 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 (recall that *genoud* is used here), and while it seems natural to choose bounds in terms of $\sqrt{d}$, the diameter of the unit $d$-dimensional hypercube, for $\theta_{\text{out}}$ the adequate diameter is slightly less straightforward and calls for some analysis with respect to the range of variation of $d_{\mathcal{E}_{r_{\text{in}}}}$ and how it depends on $\theta_{\text{in}}$. The next proposition establishes simple yet practically quite useful results regarding the diameter of $\mathcal{S}_r$ $(r > 0)$ with respect to $d_{\mathcal{E}_{r_{\text{in}}}}$ and its maximal value when letting $\theta_{\text{in}}$ vary.

**Proposition 4.** *Let $r_{in}$ 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 $\mathcal{X}$, and with a range parameter $\theta_{in} > 0$. Then the $d_{\mathcal{E}_{r_{in}}}$-diameter of $\mathcal{S}_p$ $(p > 0)$, i.e. $\sup_{S,S' \in \mathcal{S}_p} d_{\mathcal{E}_{r_{in}}}(S, S')$, 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 $\theta_{in} \in (0, +\infty)$ is given by $\sqrt{2}$.*

*Proof.* Let us consider two sets $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}, S' = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_p\} \in \mathcal{S}_p$. Then, from the fact that a correlation kernel is upper-bounded by 1, we get

$$
\begin{aligned}
d^2_{\mathcal{E}_{r_{\text{in}}}}(S, S') &= \frac{1}{p^2}\left(\sum_{i=1}^{p}\sum_{j=1}^{p} r_{\text{in}}(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{p}\sum_{j=1}^{p} r_{\text{in}}(\mathbf{x}'_i, \mathbf{x}'_j)\right. \\
&\qquad \left. -2\sum_{i=1}^{p}\sum_{j=1}^{p} r_{\text{in}}(\mathbf{x}_i, \mathbf{x}'_j)\right) \\
&\leq \frac{1}{p^2}\left(2p^2 - 2\sum_{i=1}^{p}\sum_{j=1}^{p} r_{\text{in}}(\mathbf{x}_i, \mathbf{x}'_j)\right) \\
&\leq \frac{1}{p^2}\left(2p^2 - 2\sum_{i=1}^{p}\sum_{j=1}^{p} r_{\text{in}}(\mathbf{0}_d, \mathbf{1}_d)\right),
\end{aligned}
$$

where the last inequality follows from the assumed monotonicity of $r_{\text{in}}$ with respect to the Euclidean distance between elements of $\mathcal{X}$ and the fact that the maximal distance between two points of $\mathcal{X}$, i.e. the Euclidean diameter of $[0,1]^d$, is precisely attained for $\mathbf{x} = \mathbf{0}_d$ and $\mathbf{x}' = \mathbf{1}_d$. Finally, by assumption

again, $r_{\text{in}}(\mathbf{0}_d, \mathbf{1}_d)$ is monotonically decreasing to 0 when $\theta_{\text{in}}$ decreases to 0, and so the upper bound of $d^2_{\mathcal{E}_{r_{\text{in}}}}$ tends to $\frac{1}{p^2}\left(2p^2 - 0\right) = 2$, showing that upper bound of the $d_{\mathcal{E}_{r_{\text{in}}}}$-diameter of $\mathcal{S}_p$ with respect to $\theta_{\text{in}} \in (0, +\infty)$ is $\sqrt{2}$ indeed, independently of the dimension. □

# 3 Applications

We now demonstrate the applicability of the proposed class of kernels for both prediction and optimization purposes, with comparisons when applicable to similar methods based on double sum kernels, and also to random search in the optimization case. In all examples, both inner and outer kernels are assumed Gaussian. The three hyperparameters $(\sigma_{\text{out}}, \theta_{\text{out}}, \theta_{\text{in}})$ are estimated by Maximum Likelihood with concentration on $\sigma^2_{\text{out}}$, 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 given number (404) of simulation input/outputs, while in the other test cases one may boil down to a similar situation by studying finite sets of subsets. Yet, the hydrogeology test case is the only one where the points/elements of subsets are structurally restricted to remain in a finite $\mathcal{X}$, 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 rescaled Branin-Hoo test function (See Picheny et al., 2013), 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

$$f(S) = \max_{\mathbf{x} \in S} g(\mathbf{x}) \tag{6}$$

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

$$f(S) = \frac{1}{\#S} \sum_{\mathbf{x} \in S} g(\mathbf{x}), \tag{8}$$

where $S \in \mathcal{S}_p = ([0, 1]^2)^{10}$. Let us remark that by design, the $f$ of Eq. 8 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 $k_{\text{in}}$, then $f$ is a draw of a GP with kernel $\frac{1}{\#S}\frac{1}{\#S'}\sum_{\mathbf{x} \in S, \mathbf{x} \in S'} k_{\text{in}}(\mathbf{x}, \mathbf{x}')$, 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 example input/response from this dataset. 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.

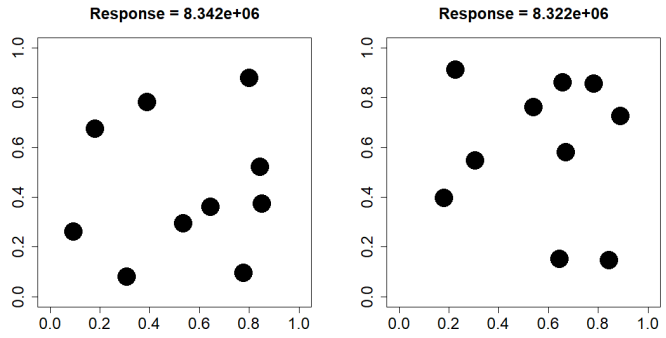


Figure 1: Examples of CASTEM data

### 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

$$g(\mathbf{x}, S) = \left( \sum_{i \in S} \sum_{t=1}^{T} |c_{\text{obs}}(i,t) - c_{\text{sim}}(\mathbf{x}, i, t)|^2 \right)^{\frac{1}{2}}, \tag{9}$$

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

$$f(S) = \sum_{i=1}^{2601} (g(\mathbf{x}_i, S_{\text{full}}) - g(\mathbf{x}_i, S))^2 \tag{10}$$

over the set $\mathcal{S}_p$ 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 $\mathcal{S}_5$ is

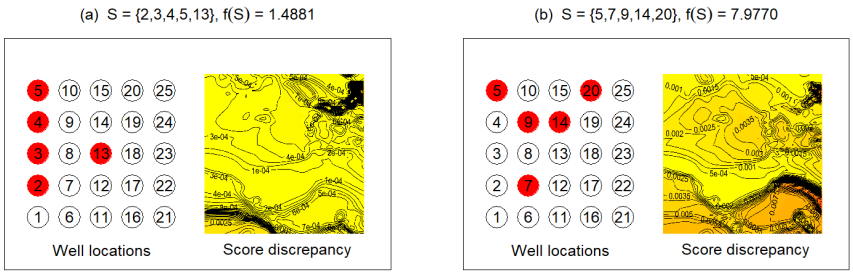


Figure 2: Score discrepancy map: location of selected wells (input $S$), score discrepancy landscape, and the spatial sum of score discrepancy objective function value $f(S)$.

$\binom{25}{5} = 53,130$. 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.

Since the base set $\mathcal{X} = \{1, 2, \ldots, 25\}$ is itself finite here, it follows from Prop. 1 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. In fact, this subset is indeed the optimal one, obtained by exhaustive search over all $53,130$ candidates. Our goal is precisely to locate these optimal well locations whose contributions minimize the spatial sum of score discrepancies without involving exhaustive enumeration. The reader is referred to (Pirot et al., 2019) for further details and visualization of the misfit objective function, location of the contaminant source, and the 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),

$$Q^2 = 1 - \frac{\sum_{i=1}^{n_{\text{test}}} (f(S_i^{(\text{test})}) - m_n(S_i^{(\text{test})}))^2}{\sum_{i=1}^{n_{\text{test}}} (f(S_i^{(\text{test})}) - \bar{\mathbf{f}})^2}, \tag{11}$$

where $n_{\text{test}}$ 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.

Table 1: Prediction performance: $Q^2$ values for models with proposed versus double sum kernels

| $Q^2$ | GP (proposed $k$) | | | GP (double sum $k$) | | |
|---|---|---|---|---|---|---|
| | **20:80** | **50:50** | **80:20** | **20:80** | **50:50** | **80:20** |
| **MAX** | 0.6926 | 0.8001 | 0.8525 | 0.6189 | 0.7429 | 0.7725 |
| **MIN** | 0.3309 | 0.4582 | 0.4929 | 0.1406 | 0.2163 | 0.2538 |
| **MEAN** | 0.9996 | 0.9999 | $\sim$1 | $\sim$1 | $\sim$1 | $\sim$1 |
| **CASTEM** | 0.5806 | 0.6641 | 0.6543 | 0.5067 | 0.5326 | 0.5107 |
| **Cont (Src A,Geo 1)** | 0.7616 | 0.8790 | 0.9115 | n.a. | n.a. | n.a. |
| **Cont (Src A,Geo 2)** | 0.7228 | 0.8569 | 0.9048 | n.a. | n.a. | n.a. |
| **Cont (Src B,Geo 1)** | 0.7937 | 0.9029 | 0.9309 | n.a. | n.a. | n.a. |
| **Cont (Src B,Geo 2)** | 0.7958 | 0.8755 | 0.8968 | n.a. | n.a. | n.a. |

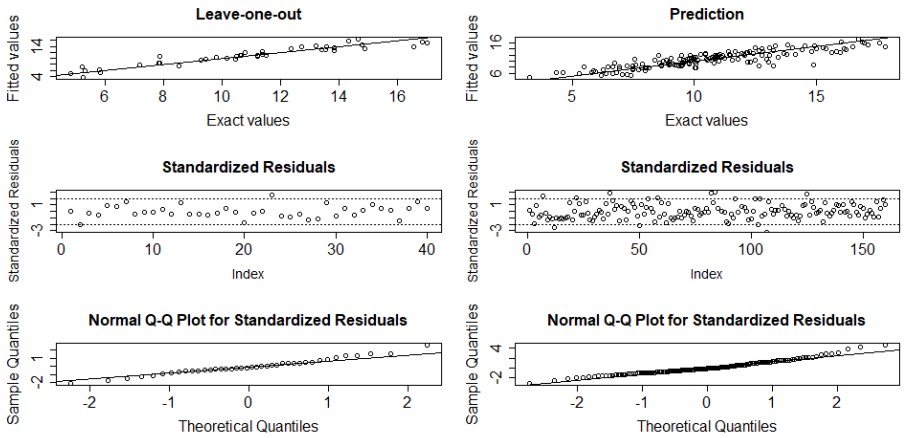


Figure 3: Residual analysis on the contaminant source localization problem (Source A, Geology 1) with ratio (20:80). (a) Internal errors (left); (b) External errors (right).

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. We see that the proposed approach gives higher value of $Q^2$ than that with double sum kernel on all problems except for the MEAN function. Moreover, $Q^2$ tends to increase with the proportion of the full data set used for training, except in one case with CASTEM.

Finally, to highlight the predictive performance of our method, Fig. 3 shows the leave-one-out diagnostic (left panel) and the out-of-sample validation (right panel) for the source localization application (Source A, Geology 1). The results show relatively moderate departures from the normality assumptions.

Table 2: The number of trials (out of 50) for which the minimum is found for EI algorithms based on GP models with proposed versus double sum kernels, as well as for Random Sampling

| **Problems** | **Number of trials** | | |
|---|---|---|---|
| | **EI (proposed $k$)** | **EI (double sum $k$)** | **RANDOM** |
| **MAX** | 38 | 8 | 3 |
| **MIN** | 10 | 9 | 3 |
| **MEAN** | 50 | 50 | 2 |
| **CASTEM** | 33 | 10 | 6 |
| **Cont (Src A,Geo 1)** | 43 | n.a. | 0 |
| **Cont (Src A,Geo 2)** | 27 | n.a. | 0 |
| **Cont (Src B,Geo 1)** | 39 | n.a. | 0 |
| **Cont (Src B,Geo 2)** | 29 | n.a. | 0 |

## 3.3 Optimization: Settings and Results

In this section, the efficiency of proposed kernels against double sum kernels are evaluated within the BO framework, using the Expected Improvement (EI) (Moćkus et al., 1978) as infill sampling criterion.

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. Concerning EI maximization, in all three synthetic problems and in the CASTEM case, as the problem sizes are relatively small, it is feasible to compute EI values at all point-sets and select the one attaining the highest value. However, for our contaminant source application, since the problem size is $> 5 \times 10^5$, EI maximization is surrogated by taking the best among 500 generated candidate point-sets using 2 strategies in the flavour of (Garnett et al., 2010). The first one focuses on exploitation by considering candidate point-sets departing by only one element from the current best subset. The second one promotes exploration by randomly generating candidate subsets.

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 (or median/selected quantiles) 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 the median value of current best $f$ values over 50 trials along with the 25th and 75th percentiles.

It is clear that EI algorithms with any of the two considered kernels are superior to random sampling. Experiments on synthetic problems show that within the two considered EI algorithm settings, the proposed kernels outperform the double sum ones on the MAX and MIN problems in terms of the number of trials that the minimum is found. On the MEAN problem, though, while both methods could locate the minimum for all 50 replications, using a double sum kernel used fewer number of iterations as can be seen in Fig. 4.

For the CASTEM dataset, EI algorithms with the proposed versus double sum kernels could locate the minimum for 33 and 10 trials, respectively (against 6 for random sampling) confirming the superior performance of the proposed method.

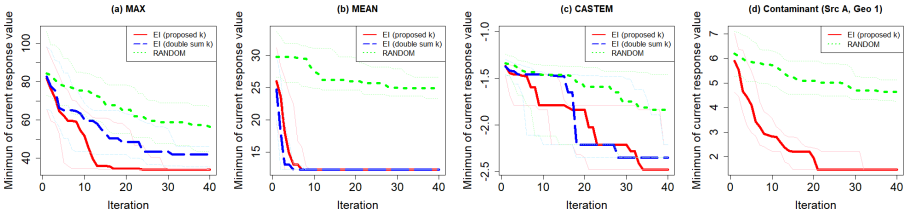


Figure 4: Progress curves of the median value of the current best response on problems (a) MAX, (b) MEAN, (c) CASTEM and (d) Contaminant Source Localization (Source A, Geology 1)

Again due to lack of strict positive definiteness, the double sum kernel is not applicable for the contaminant source applications. In this application, the EI algorithm coupled with proposed kernel is by far better than the random sampling on all four considered cases, being able to locate within 40 iterations the global optimum of the considered combinatorial optimization problem respectively in 43, 27, 39, and 29 out of 50 trials. These results certainly illustrate the potential of our proposed class of kernels to efficiently address expensive combinatorial optimization problems in a Bayesian Optimization framework.

# 4 Discussion

Experimental results obtained on the analytical objective functions and application test cases clearly confirm the added value of the proposed approach 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, and this not only to handle bigger matrices but also to tackle the optimization of infill criteria.

From the test case perspective, future work may also include tackling further prediction and subset selection problems (be it in continuous or combinatorial settings), not only for optimization purposes but also with more general goals around uncertainty quantification and reduction. Besides this, a nice feature of the propose approach is that it 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 accomodating 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, opening the door in turn to numerous non-conventional extensions of BO and related algorithms.

**Acknowledgements**

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 several colleagues including notably Fabrice Gamboa, Athénaïs Gautier, Luc Pronzato, and Henry Wynn 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, Center 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 where 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 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. (2018). Dealing with categorical and integer-valued variables in bayesian optimization with gaussian processes. arXiv:1805.03463.

Gätner, 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.

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*. arXiv:1905.09780.

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.

Moćkus, J., Tiesis, V., and Źilinskas, 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.

Picheny, V., Wagner, T., and Ginsbourger, D. (2013). A benchmark of kriging-based infill criteria for noisy optimization. *Structural and Multidisciplinary Optimization*, 48(3):607–626.

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.

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.

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.