

A WEIGHTED KERNEL METHOD FOR APPROXIMATION THAT ADAPTS TO LEARNED MULTIVARIABLE STRUCTURE

JOHN E. DARGES[✉] AND LAURA WEIDENSAGER[✉]

ABSTRACT. Approximating the input-output behavior of a multivariable black-box function from limited data is challenging when blind to the importance of its inputs and their interactions. We introduce total sensitivity kernels (TSKs), a method based on families of weighted ANOVA kernels that learn and adapt to this multivariable structure. TSKs parameterize the weights on each multivariable component of the target function by factors for each input. We propose learning these factors directly from function evaluations by selecting the reproducing kernel Hilbert space (RKHS) in which the target function has minimum norm. Under suitable conditions, we show that this norm-minimization problem admits a unique solution, and we establish consistency of a finite-data formulation based on minimum-norm interpolation. The learned TSK factors characterize the participation of individual inputs across interactions and main effects, providing a kernel-dependent notion of input sensitivity related to total Sobol’ indices. Numerical experiments demonstrate that adapting the kernel to learned multivariable structure can substantially improve approximation accuracy over a standard product kernel and ANOVA kernel.

Keywords: Adaptive kernel methods, multivariable approximation, ANOVA kernels, Weighted RKHS

1. INTRODUCTION

Consider approximating a black box function $f : \mathbb{R}^d \rightarrow \mathbb{C}$ from a limited number of function evaluations. When f is high-dimensional, accurate approximation requires many function evaluations if the approximation method does not adapt to multivariable structure. For many functions in applications, some inputs are more influential than others and only certain groups of inputs interact strongly. Effective methods must exploit, within the budget, multivariable structure not known a priori.

We describe multivariable structure through the functional analysis of variance (ANOVA) decomposition [21, 48, 2, 27]. With respect to a product measure on the input domain, f can be decomposed into mutually orthogonal components

$$f(\mathbf{x}) = \sum_{\mathbf{u} \subseteq \{1, \dots, d\}} f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}). \quad (1.1)$$

Each term $f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}})$ represents the contribution of input groupings indexed by $\mathbf{u} \subseteq \{1, \dots, d\}$. With 2^d subsets, the number of potentially relevant components grows exponentially with the input dimension.

Many functions in applications, however, have a low effective dimension [8, 32]. Functions with low effective dimension have most of their variation explained by a few ANOVA components. High-dimensional approximation is tractable if we can concentrate on the most relevant portion of the ANOVA components rather than treating them all equally.

As high-dimensional functions may possess exploitable low-dimensional structure, researchers have developed methods for structured approximation. High-dimensional model representations (HDMR) construct approximations from low-dimensional component functions by assuming that high-order interactions make small contributions [39]. Another approach estimates the effective dimension or variable importance to adapt anchored-ANOVA approximations [52]. Related ANOVA-based methods include Fourier approximation [37], random feature expansions [13, 38], and tensor train approximation [10]. These methods screen for or require tuning to important variables and interaction terms. The statistical learning community has used sparse learning and variable selection to identify the most relevant variables and interaction terms [4, 17]. Broadly, these methods impose, infer, or exploit a reduced structural representation of the target function.

Date: September 24, 2026.

John E. Darges: Department of Mathematics, Emory University. Email: jdarges@emory.edu.

Laura Weidensager: Department of Mathematics, Simon Fraser University. Email: laura.weidensager@sfu.ca.

Another perspective is to encode multivariable structure into the approximation space. One does this by constructing a norm that places weights on variables and interaction terms. Choice of weights imposes a geometry that aligns to a particular multivariable structure. Weighted function spaces appear in the analysis of high-dimensional integration to account for unequal importance of inputs and interactions [45]. Weighted ANOVA spaces use collections of weights $\{\gamma_{\mathbf{u}}\}_{\mathbf{u} \subseteq [d]}$ to rescale contributions of a function’s multivariable components. Weights determine how strongly each component is penalized by the norm, so that the same target function has a very different norm depending on how well the weights align with its multivariable structure [36, 24]. Appropriate choice of weights explains tractability of quasi-Monte Carlo (QMC) integration in high dimensions [45, 15, 36]. Structured, parameterized weights, like product weights and product-order-dependent (POD) weights, assign weight values based on a shared set of weights on input variables and interaction order [26, 15]. Parameterized weights avoid having to assign 2^d individual values for each weight. There is no all-encompassing method for selecting appropriate weights for QMC integration [29, 14].

In our setting, the goal is to select, using the limited function evaluations at hand, a weighted space suitable for high-dimensional approximation. We frame our approach in weighted reproducing kernel Hilbert spaces (RKHS), where the weights that determine the norm also determine the expression of the associated kernel function. Specifically, ANOVA kernels are our foundation. ANOVA kernels decompose into components associated with subsets of input variables and mirror the ANOVA decomposition (1.1)

$$\kappa_{\text{ANOVA}}(\mathbf{x}, \mathbf{y}) = \prod_{k=1}^d \left(1 + \kappa_k(x_k, y_k)\right) = \sum_{\mathbf{u} \subseteq \{1, \dots, d\}} \kappa_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}, \mathbf{y}_{\mathbf{u}}), \quad (1.2)$$

see also [16]. Standard ANOVA kernels have unweighted components, but current statistical methods learn a sparse set of weights for the components [4, 1]. Weighted RKHSs have seen use in deterministic kernel-based approximation for partial differential equations (PDE), with a set of structured weights chosen based on analytical properties of the PDE [23]. We aim to learn a factors for a weighted kernel that adapts the RKHS to the multivariable structure of the target function.

We introduce total sensitivity kernels (TSKs), a family of weighted ANOVA kernels whose component weights are determined by d parameters, called TSK factors, that represent the importance of each input.

$$\kappa_{\Sigma}(\mathbf{x}, \mathbf{y}) = \prod_{k=1}^d \left(1 - \Sigma_k + \Sigma_k \kappa_k(x_k, y_k)\right) = \sum_{\mathbf{u} \subseteq [d]} \gamma_{\mathbf{u}} \kappa_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}, \mathbf{y}_{\mathbf{u}}), \quad \gamma_{\mathbf{u}} = \prod_{k \in \mathbf{u}} \Sigma_k \prod_{l \notin \mathbf{u}} (1 - \Sigma_l).$$

Like product weights in QMC, TSK factors parameterize the 2^d ANOVA component weights. TSK factors define a parameterized family of RKHSs from which to choose the approximation space.. We choose TSK factors that determine the RKHS where the target function has minimal norm and so has geometry that aligns with the function’s multivariable structure. Under suitable conditions, we show that determining these weights is a strongly convex optimization problem and therefore admits a unique solution. The TSK method provides a principle for learning these parameters from available function evaluations. This population problem motivates a practical finite-data procedure where the norm of the unknown target is replaced by the norm of its minimum-norm interpolant. The finite-data objective is related to maximum likelihood estimation objectives that appear in Gaussian process methods such as automatic relevance determination (ARD) [51]. We establish theoretical consistency of the resulting finite-data objective. The method considered here is developed for deterministic approximation, so that function evaluations are assumed to be noise-free, and the input variables are assumed to be independent.

Structure of this work. We start with Section 2, where we introduce known results about kernel spaces and the ANOVA decomposition. In Section 3 we define our proposed kernel, a weighted kernel, which adapts to the multivariate structure of the target function. We state properties of this kernel, especially in Theorem 3.4 where we show strong convexity of our main objective which determines TSK factor optimality. Section 4 describes our main optimization algorithm in more detail and the numerical challenges that come with the algorithm. Finally, in Section 5 we give numerical results, which underpin the benefits of our proposed algorithm.

2. BACKGROUND

We start with the introduction of some notation. We use boldface for vectors and matrices. Denote the domains $\mathbb{D}, \Omega \subseteq \mathbb{R}$. Throughout, let $\mathbf{x} \in \mathbb{D}^d$ and $\boldsymbol{\omega} \in \Omega^d$ be d -dimensional vectors. We denote $[d] := \{1, \dots, d\}$ and the index $\mathbf{u} \subseteq [d]$ indicates a subset of the inputs. For a vector $\mathbf{x}$, the truncated vector $\mathbf{x}_{\mathbf{u}} := (x_k)_{k \in \mathbf{u}}$ features only those inputs associated to $\mathbf{u}$. If $\mathbf{x}^i$ is the i th point in a data set, we denote the k th entry as x_k^i .

2.1. Reproducing Kernel Hilbert Spaces. Consider a feature map $\Phi : \mathbb{D}^d \times \Omega^d \rightarrow \mathbb{C}$ along with a probability measure τ on Ω^d . We take the feature map to be the Fourier map $\Phi(\mathbf{x}, \boldsymbol{\omega}) = e^{i\langle \boldsymbol{\omega}, \mathbf{x} \rangle}$ which, due to Bochner's theorem [7, 22], induces a reproducing kernel $\kappa : \mathbb{D}^d \times \mathbb{D}^d \rightarrow \mathbb{C}$ via

$$\kappa(\mathbf{x}, \mathbf{y}) = \int_{\Omega^d} e^{i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle} d\tau(\boldsymbol{\omega}). \quad (2.1)$$

For comprehensive theory and properties of reproducing kernels, see [3, 11, 5]. We recall some basic properties and theories. There is a unique Hilbert space $\mathcal{H}$ associated with the kernel $\kappa : \mathbb{D}^d \times \mathbb{D}^d \rightarrow \mathbb{C}$, known as a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of functions $f : \mathbb{D}^d \rightarrow \mathbb{C}$, with the property that $\kappa(\mathbf{x}, \cdot) \in \mathcal{H}$, for all $\mathbf{x} \in \mathbb{D}^d$, and

$$f(\mathbf{x}) = \langle \kappa(\mathbf{x}, \cdot), f \rangle_{\mathcal{H}},$$

for all $\mathbf{x} \in \mathbb{D}^d$. The reproducing kernel κ has the key property that, for distinct points $\mathbf{x}^1, \dots, \mathbf{x}^M \in \mathbb{D}^d$, the $M \times M$ matrix $\mathbf{K}$, where $\mathbf{K}_{i,j} = \kappa(\mathbf{x}^i, \mathbf{x}^j)$, is symmetric positive definite. Consider the operator $T : L_2(\Omega^d, \tau) \rightarrow L_2(\mathbb{D}^d)$,

$$(T\beta)(\mathbf{x}) = \int_{\Omega^d} \beta(\boldsymbol{\omega}) e^{i\langle \boldsymbol{\omega}, \mathbf{x} \rangle} d\tau(\boldsymbol{\omega}), \quad \beta \in L_2(\Omega^d, \tau).$$

The image of T is the RKHS $\mathcal{H}$. The norm of f in the RKHS $\mathcal{H}$ is then

$$\|f\|_{\mathcal{H}}^2 = \min_{T\beta=f} \|\beta\|_{L_2(\Omega^d, \tau)}^2.$$

The function β is called the spectral factor of f . In this work, we consider kernels with a tensor product structure. These are multivariable kernels that are the Fourier transform of a probability measure which corresponds to independently distributed random variables. The probability measure takes the form $\tau = \tau_1 \otimes \dots \otimes \tau_d$, where each τ_k is a one-dimensional probability measure. The kernel can be written

$$\kappa(\mathbf{x}, \mathbf{y}) = \int_{\Omega^d} e^{i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle} d\tau(\boldsymbol{\omega}) = \prod_{k=1}^d \int_{\Omega} e^{i\omega_k(x_k - y_k)} d\tau_k(\omega_k). \quad (2.2)$$

We often take τ_k to be continuous distributions.

Kernel learning is a subset of machine learning. It involves searching an RKHS for a function that closely approximates a target function or a set of data. An RKHS is the closure of the span of $\{\kappa(\mathbf{x}, \cdot); \mathbf{x} \in \mathcal{X}\}$. So kernel learning means approximating by an expansion $\sum_{i=1}^M \beta_i \kappa(\mathbf{x}_i, \cdot)$.

2.2. Kernel ridge regression and generalization bounds. Kernel ridge regression (KRR) is a cornerstone algorithm for non-parametric function estimation. Consider a dataset $\{(\mathbf{x}^i, f^i)\}_{i=1}^M$. Here, each $\mathbf{x}_i \sim \mu$ is drawn identically and independently, while $f(\mathbf{x}^i) = f^i$. KRR finds an estimator $\hat{f}_{\lambda}$ by minimizing the regularized empirical risk over the RKHS $\mathcal{H}$ [43, 49],

$$\hat{f}_{\lambda} = \arg \min_{f \in \mathcal{H}} \frac{1}{M} \sum_{i=1}^M (f(\mathbf{x}^i) - f^i)^2 + \lambda \|f\|_{\mathcal{H}}^2,$$

where $\lambda > 0$ is the regularization parameter controlling the trade-off between data fit and generalization.

The theoretical performance of the estimator $\hat{f}_{\lambda}$ is typically evaluated via its expected generalization error in $L_2(\mathbb{D}^d, \mu)$. Assuming the true target function f is contained within the RKHS $\mathcal{H}$, standard results from statistical learning theory and integral operator techniques [9, 46] provide a high-probability upper bound on the squared $L_2(\mathbb{D}^d, \mu)$ error,

$$\|\hat{f}_{\lambda} - f\|_{L_2(\mu)}^2 \leq \mathcal{O} \left(\frac{K^2 \log(1/\delta)}{\lambda M} \right) + \lambda \|f\|_{\mathcal{H}}^2, \quad (2.3)$$

where $K = \sup_{\mathbf{x} \in \mathbb{D}^d} \sqrt{\kappa(\mathbf{x}, \mathbf{x})}$ bounds the kernel and $1 - \delta$ is the confidence level.

The bound in (2.3) decomposes the error into two competing terms. The first term represents the estimation error, which vanishes as the sample size $M \rightarrow \infty$. When restricting to shift-invariant kernels, this term is independent of the choice of kernel since $K = 1$ for all shift-invariant kernels. The second term represents the approximation bias introduced by the Tikhonov regularization. Balancing these two terms optimally by setting $\lambda \sim 1/\sqrt{M}$ yields an overall learning rate of $\mathcal{O}(\|f\|_{\mathcal{H}} M^{-1/2})$.

Equation (2.3) reveals a fundamental principle for methods using shift-invariant kernels: the tightness of the generalization bound is strictly governed by the RKHS norm of the true function, $\|f\|_{\mathcal{H}}$. Selecting a kernel that actively minimize the objective $\|f\|_{\mathcal{H}}$ explicitly minimizes the leading constant of the approximation bias in (2.3).

3. THEORY OF TOTAL SENSITIVITY KERNELS

In this section we propose the procedure to learn a kernel κ that adapts to the multivariable structure of the target function. The kernels take a form similar to ANOVA kernels (1.2). We modify the product kernel (2.2) by introducing tunable parameters that control input importance.

Definition 3.1. For each $k \in [d]$, let κ_k be a univariate kernel. A total sensitivity kernel (TSK) is defined by

$$\kappa_{\Sigma}(\mathbf{x}, \mathbf{y}) = \prod_{k=1}^d (1 - \Sigma_k + \Sigma_k \kappa_k(x_k, y_k)). \quad (3.1)$$

There, we denote the TSK factors $\Sigma = (\Sigma_1, \dots, \Sigma_d)$, with $\Sigma_k \in (0, 1)$. Let $\mathcal{H}_{\Sigma}$ be the RKHS associated to κ_{Σ} , with norm $\|\cdot\|_{\Sigma}$.

The TSK (3.1) takes an ANOVA form when we multiply out,

$$\kappa_{\Sigma}(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{u} \subseteq [d]} \gamma_{\mathbf{u}} \kappa_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}, \mathbf{y}_{\mathbf{u}}), \quad \kappa_{\mathbf{u}} = \prod_{k \in \mathbf{u}} \kappa_k, \quad \gamma_{\mathbf{u}} = \prod_{k \in \mathbf{u}} \Sigma_k \prod_{l \notin \mathbf{u}} (1 - \Sigma_l).$$

Therefore, an input's TSK factor being larger means that interaction terms corresponding to that input are weighted more. In this work, we let the univariate kernels be shift-invariant kernels, specifically those that come from characteristic functions,

$$\kappa_k(x_k, y_k) = \phi_k(x_k - y_k) = \int_{\Omega_k} e^{i\omega_k(x_k - y_k)} d\tau_k(\omega_k).$$

Here, ϕ_k is the characteristic function of a distribution τ_k . Then, $\kappa_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}, \mathbf{y}_{\mathbf{u}}) = \prod_{k \in \mathbf{u}} \phi_k(x_k - y_k)$.

3.1. RKHS properties. We discuss the properties of the TSKs and characterize the corresponding RKHS. As a corollary of [3, Part I.6], we have the following lemma for the norm in a sum of RKHSs.

Lemma 3.2. Let κ_{Σ} be a TSK. For $f \in \mathcal{H}_{\Sigma}$, with corresponding RKHS norm $\|\cdot\|_{\Sigma}$, then

$$\|f\|_{\Sigma}^2 = \min_{f = \sum_{\mathbf{u}} f_{\mathbf{u}}} \sum_{\mathbf{u} \subseteq [d]} \frac{1}{\gamma_{\mathbf{u}}} \|f_{\mathbf{u}}\|_{\mathcal{H}_{\mathbf{u}}}^2, \quad f_{\mathbf{u}} \in \mathcal{H}_{\mathbf{u}}. \quad (3.2)$$

Each RKHS $\mathcal{H}_{\mathbf{u}}$ corresponds to the kernel $\kappa_{\mathbf{u}} = \prod_{k \in \mathbf{u}} \kappa_k$ with attached norm $\|\cdot\|_{\mathcal{H}_{\mathbf{u}}}$.

Proof. Each $\gamma_{\mathbf{u}} \kappa_{\mathbf{u}}$ is a kernel with an associated RKHS, which we denote $\mathcal{H}_{\gamma, \mathbf{u}}$ with corresponding norm $\|\cdot\|_{\gamma, \mathbf{u}}$. According to [3, Part I.6] about the sum of reproducing kernels, the norm of $f \in \mathcal{H}_{\Sigma}$ is

$$\|f\|_{\Sigma}^2 = \min_{f = \sum_{\mathbf{u}} f_{\mathbf{u}}} \sum_{\mathbf{u} \subseteq [d]} \|f_{\mathbf{u}}\|_{\mathcal{H}_{\gamma, \mathbf{u}}}^2, \quad f_{\mathbf{u}} \in \mathcal{H}_{\gamma, \mathbf{u}}.$$

Yet, note that the class of functions comprising $\mathcal{H}_{\mathbf{u}}$ is the same as that of $\mathcal{H}_{\gamma, \mathbf{u}}$. If $f \in \mathcal{H}_{\gamma, \mathbf{u}}$, then we can express it as

$$f = \sum_{i=1}^M \alpha_i (\gamma_{\mathbf{u}} \kappa_{\mathbf{u}}(\mathbf{x}^i, \cdot)).$$

It is clear that $f = \sum_{i=1}^M \beta_i \kappa_{\mathbf{u}}(\mathbf{x}^i, \cdot)$, with $\beta_i = \gamma_{\mathbf{u}} \alpha_i$ so that $f \in \mathcal{H}_{\mathbf{u}}$. The norm of f in $\mathcal{H}_{\gamma, \mathbf{u}}$ is

$$\|f\|_{\mathcal{H}_{\gamma, \mathbf{u}}}^2 = \sum_{i, \ell=1}^M \alpha_i \alpha_\ell (\gamma_{\mathbf{u}} \kappa_{\mathbf{u}}(\mathbf{x}^i, \mathbf{x}^\ell)) = \frac{1}{\gamma_{\mathbf{u}}} \sum_{i, \ell=1}^M \beta_i \beta_\ell \kappa_{\mathbf{u}}(\mathbf{x}^i, \mathbf{x}^\ell) = \frac{1}{\gamma_{\mathbf{u}}} \|f\|_{\mathcal{H}_{\mathbf{u}}}^2.$$

Summing all terms gives

$$\min_{f=\sum_{\mathbf{u}} f_{\mathbf{u}}} \sum_{\mathbf{u} \subseteq [d]} \|f_{\mathbf{u}}\|_{\mathcal{H}_{\gamma, \mathbf{u}}}^2 = \min_{f=\sum_{\mathbf{u}} f_{\mathbf{u}}} \sum_{\mathbf{u} \subseteq [d]} \frac{1}{\gamma_{\mathbf{u}}} \|f_{\mathbf{u}}\|_{\mathcal{H}_{\mathbf{u}}}^2.$$

This finishes the proof. $\square$

The lemma shows an immediate connection to the ANOVA decomposition. Due to the orthogonality of the terms in the ANOVA decomposition (1.1), there we have $\|f\|^2 = \sum_{\mathbf{u} \subseteq [d]} \|f_{\mathbf{u}}\|^2$. So, the RKHS norm $\|f\|_{\Sigma}$ corresponding to the TSK κ_{Σ} resembles a weighted sum of ANOVA energies in the RKHS.

In the discussion of kernels of the form (2.1), we see every probability measure induces a kernel. TSKs are induced by a probability measure. This gives an alternative formulation of the kernel,

$$\kappa_{\Sigma}(\mathbf{x}, \mathbf{y}) = \int_{\Omega^d} e^{i\langle \omega, \mathbf{x} - \mathbf{y} \rangle} d\tau_{\Sigma}(\omega), \quad \tau_{\Sigma} = \sum_{\mathbf{u} \subseteq [d]} \left(\bigotimes_{k \in \mathbf{u}} \Sigma_k \tau_k \bigotimes_{k \notin \mathbf{u}} (1 - \Sigma_k) \delta_0 \right),$$

where δ_0 is the Dirac measure. This shows that a TSK is induced by a spike-and-slab distribution. Hence a TSK is a shift-invariant kernel as long as the 1D kernels κ_k are shift-invariant. We can interpret the TSK factor Σ_k as the probability of “activation” for the input x_k .

The TSK factors parameterize a family of probability measures, kernel functions, and RKHSs. Choosing TSK factors means choosing which RKHS to approximate f in. What is the relationship between the RKHSs corresponding to two different sets of TSK factors? In fact, the RKHSs describe the same sets of functions, no matter how the TSK factors Σ are chosen. Changing TSK factors only changes the RKHS inner product. The following proposition states this in more detail.

Proposition 3.3. *Consider a family of TSKs parameterized by Σ . Let $\Sigma \in (0, 1)^d$ and $\tilde{\Sigma} \in [0, 1]^d$. Then any $f \in \mathcal{H}_{\tilde{\Sigma}}$ is also a member of $\mathcal{H}_{\Sigma}$.*

Proof. For each $f \in \mathcal{H}_{\tilde{\Sigma}}$, there exists a spectral factor $\tilde{\beta}$ in $L_2(\tau_{\tilde{\Sigma}})$ such that

$$f(\mathbf{x}) = \int \tilde{\beta}(\omega) e^{i\omega^\top \mathbf{x}} d\tau_{\tilde{\Sigma}}(\omega) = \sum_{\mathbf{u} \subseteq [d]} \tilde{\gamma}_{\mathbf{u}} \int \tilde{\beta}_{\mathbf{u}}^0 e^{i\omega_{\mathbf{u}}^\top \mathbf{x}_{\mathbf{u}}} d\tilde{\tau}_{\mathbf{u}}(\omega_{\mathbf{u}}),$$

where $\tilde{\tau}_{\mathbf{u}} := \bigotimes_{k \in \mathbf{u}} \tau_{\tilde{\Sigma}_k}$ and $\tilde{\gamma}_{\mathbf{u}} = \prod_{k \in \mathbf{u}} \tilde{\Sigma}_k \prod_{l \notin \mathbf{u}} (1 - \tilde{\Sigma}_l)$. Now, define the function

$$\frac{d\tau_{\tilde{\Sigma}}}{d\tau_{\Sigma}}(\omega) = \begin{cases} \frac{\tilde{\gamma}_{\emptyset}}{\gamma_{\emptyset}}, & \omega \in \Omega_{\emptyset} \\ \frac{\tilde{\gamma}_{\{1\}}}{\gamma_{\{1\}}}, & \omega \in \Omega_{\{1\}} \\ \vdots & \\ \frac{\tilde{\gamma}_{\mathbf{u}}}{\gamma_{\mathbf{u}}}, & \omega \in \Omega_{\mathbf{u}} \\ \vdots & \\ \frac{\tilde{\gamma}_{[d]}}{\gamma_{[d]}}, & \omega \in \Omega_{[d]} \end{cases}.$$

This function is piecewise constant on the 2^d different sets $\Omega_{\mathbf{u}} \subset \mathbb{R}^d$, with $\Omega_{\mathbf{u}} = \{\omega \in \mathbb{R}^d : \text{supp } \omega = \mathbf{u}\}$. For example, $\Omega_{\emptyset} = \{0\}$, $\Omega_{\{1\}} = \{\omega \in \mathbb{R}^d : \omega_k = 0, k \neq 1\}$, and $\Omega_{[d]} = \{\omega \in \mathbb{R}^d : \omega_k \neq 0\}$. The function $\frac{d\tau_{\tilde{\Sigma}}}{d\tau_{\Sigma}}(\omega)$ is the Radon-Nikodym derivative. This exists when $\tau_{\tilde{\Sigma}}$ is absolutely continuous with respect to τ_{Σ} , which is the case whenever $\Sigma \in (0, 1)^d$. With these we can write

$$\begin{aligned} f(\mathbf{x}) &= \sum_{\mathbf{u} \subseteq [d]} \tilde{\gamma}_{\mathbf{u}} \int_{\Omega_{|\mathbf{u}|}} \tilde{\beta}_{\mathbf{u}}^0 e^{i\omega_{\mathbf{u}}^\top \mathbf{x}_{\mathbf{u}}} d\tilde{\tau}_{\mathbf{u}}(\omega_{\mathbf{u}}) = \sum_{\mathbf{u} \subseteq [d]} \gamma_{\mathbf{u}} \int_{\Omega_{|\mathbf{u}|}} \frac{\tilde{\gamma}_{\mathbf{u}}}{\gamma_{\mathbf{u}}} \tilde{\beta}_{\mathbf{u}}^0 e^{i\omega_{\mathbf{u}}^\top \mathbf{x}_{\mathbf{u}}} d\tilde{\tau}_{\mathbf{u}}(\omega_{\mathbf{u}}) \\ &= \int \tilde{\beta}(\omega) e^{i\omega^\top \mathbf{x}} \frac{d\tau_{\tilde{\Sigma}}}{d\tau_{\Sigma}}(\omega) d\tau_{\Sigma}(\omega). \end{aligned}$$

This shows that f can be expressed through the new spectral factor $\beta = \tilde{\beta} \frac{d\tau_{\Sigma}}{d\tau_{\Sigma}}$ in $\mathcal{H}_{\Sigma}$. To ensure $f \in \mathcal{H}_{\Sigma}$, we verify that $\beta \in L_2(\Omega, \tau_{\Sigma})$. Let $m := \max_{\mathbf{u} \subseteq [d]} \frac{\tilde{\gamma}_{\mathbf{u}}}{\gamma_{\mathbf{u}}} < \infty$. Then,

$$\|\beta\|_{L_2(\tau_{\Sigma})}^2 = \int |\tilde{\beta}|^2 \left(\frac{d\tau_{\Sigma}}{d\tau_{\Sigma}} \right)^2 d\tau_{\Sigma} \leq m \int |\tilde{\beta}|^2 \frac{d\tau_{\Sigma}}{d\tau_{\Sigma}} d\tau_{\Sigma} = m \|\tilde{\beta}\|_{L_2(\tau_{\Sigma})}^2 < \infty.$$

Since β has a finite RKHS norm under τ_{Σ} , we conclude $f \in \mathcal{H}_{\Sigma}$. $\square$

Choosing which RKHS we approximate f in, across all TSK factors in $(0, 1)^d$, is then a well-posed question. It is impossible for the target to be in one RKHS but not another.

3.2. Optimal TSK factors. We now must answer what makes TSK factors optimal. We want the structure of κ_{Σ} to match the multivariable structure of f . Recall that the corresponding RKHS norm in Lemma 3.2 weighs contributions by different groupings of inputs differently. Since each RKHS includes the same set of functions, different weights align the geometry of the space with a specific multivariable structure. A small value of the weight $\gamma_{\mathbf{u}}$ heavily penalizes a large $\mathbf{u}$ -component of f by inflating the RKHS norm. Conversely, a large value of $\gamma_{\mathbf{u}}$ does not penalize a large $\mathbf{u}$ -component. So, the multivariable structure of f aligns most closely with the RKHS where the norm is smallest. Identifying the RKHS where f has the minimal norm also lowers one of the terms in the upper bound on the KRR error in (2.3).

We therefore optimize the TSK factors $\Sigma = (\Sigma_1, \dots, \Sigma_d)$ in (3.1) by minimizing the RKHS norm,

$$g(\Sigma) := \|f\|_{\Sigma}^2.$$

The following result shows that for a fixed function f the norm $g(\Sigma)$ as a function depending on the parameters $\Sigma = (\Sigma_1, \dots, \Sigma_d)$ is not only convex, but also strongly convex, which means that the spectrum of the Hessian is bounded from below by a constant bigger than zero. Strongly convex functions have the benefit that they have a unique global minimum, if it exists.

Theorem 3.4. *Let $0 \neq f \in \mathcal{H}_{\Sigma}$ for some, and hence every, $\Sigma \in (0, 1)^d$. Assume that each kernel κ_k is a strictly positive definite kernel satisfying $\kappa_k(x_k, x_k) = 1$. Then the function*

$$g(\Sigma) = \|f\|_{\Sigma}^2$$

is strongly convex on $(0, 1)^d$.

Proof. We first construct a decomposition of each one-dimensional kernel κ_k into a constant part and a part whose RKHS contains no nonzero constant functions. Define $r_k := \|1\|_{\mathcal{H}_k}^2$. For every $k \in [d]$, there are two cases and we define the constants

$$C_k := \begin{cases} 1 & \text{if } 1 \notin \mathcal{H}_k, \\ 1 - \frac{1}{r_k} & \text{if } 1 \in \mathcal{H}_k. \end{cases} \quad (3.3)$$

Then we define the RKHS, which contains no constant by

$$\tilde{\kappa}_k(x_k, y_k) := \begin{cases} \kappa_k(x_k, y_k) & \text{if } 1 \notin \mathcal{H}_k, \\ \kappa_k(x_k, y_k) - \frac{1}{r_k} & \text{if } 1 \in \mathcal{H}_k. \end{cases}$$

To see that $\tilde{\kappa}_k$ is positive definite in the second case, let $\{x_k^1, \dots, x_k^M\} \subset \mathbb{D}$ and $a_1, \dots, a_M \in \mathbb{R}$. By the reproducing property,

$$\begin{aligned} \sum_{i,j=1}^M a_i a_j \tilde{\kappa}_k(x_k^i, x_k^j) &= \left\| \sum_{i=1}^M a_i \kappa_k(x_k^i, \cdot) \right\|_{\mathcal{H}_k}^2 - \frac{1}{r_k} \left(\sum_{i=1}^M a_i \right)^2 \\ &= \left\| \sum_{i=1}^M a_i \kappa_k(x_k^i, \cdot) \right\|_{\mathcal{H}_k}^2 - \frac{\left| \left\langle 1, \sum_{i=1}^M a_i \kappa_k(x_k^i, \cdot) \right\rangle_{\mathcal{H}_k} \right|^2}{\|1\|_{\mathcal{H}_k}^2} \geq 0, \end{aligned}$$

where the final inequality follows from the Cauchy-Schwarz inequality. Now, if $1 \in \mathcal{H}_k$, $\mathcal{H}_k = \mathcal{H}_k^1 \oplus \mathcal{H}_{\tilde{\kappa}_k}$ where $\mathcal{H}_k^1 = \text{span}\{1\}$ with norm inherited from $\mathcal{H}_k$. The reproducing kernel of this one-dimensional subspace is the constant kernel $1/r_k$.

We next show that $0 < C_k \leq 1$. The result is immediate if $1 \notin \mathcal{H}_k$. Suppose instead that $1 \in \mathcal{H}_k$. Since $\kappa_k(x_k, x_k) = 1$, the reproducing property gives

$$|h(x_k)| \leq \|h\|_{\mathcal{H}_k} \|\kappa_k(x_k, \cdot)\|_{\mathcal{H}_k} = \|h\|_{\mathcal{H}_k}$$

for every $h \in \mathcal{H}_k$. Taking $h = 1$ gives $1 \leq \|1\|_{\mathcal{H}_k}$. Equality would imply equality in the Cauchy–Schwarz inequality for 1 and $\kappa_k(x_k, \cdot)$ for every $x_k \in \mathbb{D}$. This would mean $\kappa_k(x_k, \cdot) = 1$ for every x_k , and hence $\kappa_k \equiv 1$, contradicting strict positive definiteness on a domain containing at least two distinct points. Then, $r_k = \|1\|_{\mathcal{H}_k}^2 > 1$ and

$$0 < C_k = 1 - \frac{1}{r_k} < 1.$$

In either case,

$$0 < C_k \leq 1.$$

The k th factor of the TSK can be written as

$$1 - \Sigma_k + \Sigma_k \kappa_k(x_k, y_k) = 1 + \Sigma_k (\kappa_k(x_k, y_k) - 1) = 1 + \Sigma_k (\tilde{\kappa}_k(x_k, y_k) - C_k),$$

since in both cases $\tilde{\kappa}_k(x_k, y_k) - C_k = \kappa_k(x_k, y_k) - 1$. Therefore,

$$\kappa_{\Sigma}(\mathbf{x}, \mathbf{y}) = \prod_{k=1}^d (1 - C_k \Sigma_k + \Sigma_k \tilde{\kappa}_k(x_k, y_k)). \quad (3.4)$$

Decomposition of the RKHS norm:

Expanding (3.4) gives

$$\kappa_{\Sigma}(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{u} \subseteq [d]} \theta_{\mathbf{u}}(\Sigma) \tilde{\kappa}_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}, \mathbf{y}_{\mathbf{u}}), \quad \theta_{\mathbf{u}}(\Sigma) := \prod_{k \in \mathbf{u}} \Sigma_k \prod_{\ell \notin \mathbf{u}} (1 - C_{\ell} \Sigma_{\ell}),$$

where

$$\tilde{\kappa}_{\mathbf{u}} := \prod_{k \in \mathbf{u}} \tilde{\kappa}_k, \quad \tilde{\kappa}_{\emptyset} = 1.$$

Since $0 < C_k \leq 1$ and $0 < \Sigma_k < 1$, we have $1 - C_k \Sigma_k > 0$. So, $\theta_{\mathbf{u}}(\Sigma) > 0$ for every $\mathbf{u} \subseteq [d]$.

By construction, the constant space and $\mathcal{H}_{\tilde{\kappa}_k}$ have trivial intersection for every $k \in [d]$. Then, the tensor-product spaces corresponding to distinct subsets $\mathbf{u}$ decompose $\mathcal{H}_{\Sigma}$ by direct sum,

$$\mathcal{H}_{\Sigma} = \bigoplus_{\mathbf{u} \subseteq [d]} \left(\bigotimes_{k \in \mathbf{u}} \Sigma_k \mathcal{H}_{\tilde{\kappa}_k} \otimes \bigotimes_{\ell \notin \mathbf{u}} (1 - C_{\ell} \Sigma_{\ell}) \mathcal{H}_{\ell}^1 \right). \quad (3.5)$$

Every $f \in \mathcal{H}_{\Sigma}$ therefore has a unique decomposition

$$f = \sum_{\mathbf{u} \subseteq [d]} f_{\mathbf{u}}, \quad f_{\mathbf{u}} \in \mathcal{H}_{\tilde{\kappa}_{\mathbf{u}}},$$

where the constant factors are the inactive coordinates $[d] \setminus \mathbf{u}$. Note that these component spaces, and therefore the functions $f_{\mathbf{u}}$, do not depend on Σ .

By the scaling property of RKHS norms,

$$g(\Sigma) = \|f\|_{\Sigma}^2 = \sum_{\mathbf{u} \subseteq [d]} \frac{\|f_{\mathbf{u}}\|_{\mathcal{H}_{\tilde{\kappa}_{\mathbf{u}}}}^2}{\theta_{\mathbf{u}}(\Sigma)} = \sum_{\mathbf{u} \subseteq [d]} \frac{\beta_{\mathbf{u}}}{\theta_{\mathbf{u}}(\Sigma)}, \quad \beta_{\mathbf{u}} := \|f_{\mathbf{u}}\|_{\mathcal{H}_{\tilde{\kappa}_{\mathbf{u}}}}^2.$$

The coefficients $\beta_{\mathbf{u}} \geq 0$ depend only on f and the chosen kernels and not on Σ . Since $f \neq 0$, there exists at least one $\mathbf{u} \subseteq [d]$ such that $\beta_{\mathbf{u}} > 0$.

The Hessian matrix:

For $\mathbf{u} \subseteq [d]$, define

$$\xi_k^{\mathbf{u}} := \begin{cases} -\frac{1}{\Sigma_k}, & k \in \mathbf{u}, \\ \frac{C_k}{1 - C_k \Sigma_k}, & k \notin \mathbf{u}. \end{cases}$$

Then $\xi_k^{\mathbf{u}} > 0$ for $k \in \mathbf{u}$ and $\xi_k^{\mathbf{u}} < 0$ for $k \notin \mathbf{u}$. Direct differentiation gives

$$\mathbf{H} = \nabla^2 g(\boldsymbol{\Sigma}) = \sum_{\mathbf{u} \subseteq [d]} \frac{\beta_{\mathbf{u}}}{\theta_{\mathbf{u}}} \mathbf{H}^{\mathbf{u}},$$

with diagonal entries $H_{k,k}^{\mathbf{u}} = 2(\xi_k^{\mathbf{u}})^2$ and, non-diagonal entries for $k \neq j$, $H_{k,j}^{\mathbf{u}} = \xi_k^{\mathbf{u}} \xi_j^{\mathbf{u}}$. Consequently, for any $\mathbf{v} \in \mathbb{R}^d$,

$$\mathbf{v}^\top \mathbf{H}^{\mathbf{u}} \mathbf{v} = \sum_{k=1}^d v_k^2 (\xi_k^{\mathbf{u}})^2 + \left(\sum_{k=1}^d v_k \xi_k^{\mathbf{u}} \right)^2.$$

In particular, $\mathbf{H}^{\mathbf{u}}$ is positive definite because $\xi_k^{\mathbf{u}} \neq 0$ for every k .

It remains to show that the smallest eigenvalue of $\mathbf{H}$ is uniformly bounded below on $(0, 1)^d$. Choose $\mathbf{u}$ such that $\beta_{\mathbf{u}} > 0$. Define

$$\widehat{C} := \min_{k \in [d]} C_k > 0$$

If $k \in \mathbf{u}$, then $|\xi_k^{\mathbf{u}}| = \frac{1}{\Sigma_k} > 1 > \widehat{C}$. If $k \notin \mathbf{u}$, then

$$|\xi_k^{\mathbf{u}}| = \frac{C_k}{1 - C_k \Sigma_k} > C_k > \widehat{C}.$$

Moreover, $0 < \theta_{\mathbf{u}}(\boldsymbol{\Sigma}) \leq 1$ because $0 < \Sigma_k < 1$ and $0 < 1 - C_k \Sigma_k \leq 1$. Hence, for any $\mathbf{v}$ with $\|\mathbf{v}\| = 1$,

$$\mathbf{v}^\top \mathbf{H} \mathbf{v} \geq \frac{\beta_{\mathbf{u}}}{\theta_{\mathbf{u}}} \mathbf{v}^\top \mathbf{H}^{\mathbf{u}} \mathbf{v} \geq \frac{\beta_{\mathbf{u}}}{\theta_{\mathbf{u}}} \sum_{k=1}^d v_k^2 (\xi_k^{\mathbf{u}})^2 \geq \beta_{\mathbf{u}} \widehat{C}^2 \sum_{k=1}^d v_k^2 = \beta_{\mathbf{u}} \widehat{C}^2 > 0.$$

The lower bound is independent of $\mathbf{v}$ and $\boldsymbol{\Sigma}$. Therefore,

$$\inf_{\boldsymbol{\Sigma} \in (0, 1)^d} \lambda_{\min}(\nabla^2 g(\boldsymbol{\Sigma})) \geq \beta_{\mathbf{u}} \widehat{C}^2 > 0,$$

which proves that g is strongly convex on $(0, 1)^d$. $\square$

As a consequence, minimizing TSK factors, if they exists, are unique to a function f .

3.3. Interpretation of the TSK factors $\boldsymbol{\Sigma}$. While the TSK factors $\boldsymbol{\Sigma}$ naturally act as a measure of variable importance, it is crucial to carefully distinguish them from classical total Sobol' indices [48]. In standard global sensitivity analysis, the total Sobol' index quantifies the exact variance contribution of a variable (including all its interactions) relative to the total variance of the model [47]. This variance decomposition relies fundamentally on the ANOVA decomposition, which requires the functions $f_{\mathbf{u}}$ in (1.1) to be strictly orthogonal in L_2 with respect to a given probability measure. We do not force orthogonality with our kernel construction in (3.1). As a consequence, the strict zero-mean property is relaxed, and the L_2 -orthogonality of the ANOVA components is no longer guaranteed. The components may overlap and exhibit non-zero covariance, meaning the total variance of f does not split into an independent sum. Thus, Σ_k cannot be interpreted as an exact variance-based Sobol' index.

Nevertheless, the TSK factors $\boldsymbol{\Sigma}$ maintain a rigorous and highly useful interpretation as a *model-intrinsic, RKHS-based sensitivity weight*. Even without L_2 -orthogonality, the tensor product structure of $\kappa_{\boldsymbol{\Sigma}}$ ensures that the RKHS $\mathcal{H}_{\boldsymbol{\Sigma}}$ splits into a direct sum of sub-Hilbert spaces, see (3.5). Having a closer look at the norm $\|f\|_{\boldsymbol{\Sigma}}^2$, we see that if we optimize for a single index $k \in [d]$ and fix all other parameters Σ_{ℓ} , then the RKHS norm has the form

$$\|f\|_{\boldsymbol{\Sigma}}^2 = \frac{A_k}{\Sigma_k} + \frac{B_k}{1 - \Sigma_k},$$

where A_k and B_k do not depend on Σ_k . Minimizing this expression by setting the derivative to zero yields

$$\Sigma_k = \frac{\sqrt{A_k}}{\sqrt{A_k} + \sqrt{B_k}}.$$

Therefore, the learned parameters $\boldsymbol{\Sigma} = (\Sigma_1, \dots, \Sigma_d)$ serve as a continuous feature selection mechanism and a proxy for total sensitivity. A value of $\Sigma_k \rightarrow 1$ indicates that the k -th variable carries substantial explanatory

energy necessary to fit the data, whereas $\Sigma_k \rightarrow 0$ acts as a shrinkage operator, effectively pruning the irrelevant dimension from the model.

Unfortunately, there is no closed form for the optimal parameters Σ in Theorem 3.4. However, there are coupled equations, which any optimal parameters $\Sigma^* = (\Sigma_1^*, \dots, \Sigma_d^*)$ fulfill.

Let C_k be defined as in (3.3). Under the assumptions of Theorem 3.4, these constants are bounded by $0 < C_k \leq 1$. Recall that $g(\Sigma) = \|f\|_\Sigma^2$ with

$$\|f\|_\Sigma^2 = \sum_{\mathbf{u} \subseteq [d]} \frac{\beta_{\mathbf{u}}}{\theta_{\mathbf{u}}(\Sigma)}, \quad \theta_{\mathbf{u}}(\Sigma) = \prod_{k \in \mathbf{u}} \Sigma_k \prod_{k \notin \mathbf{u}} (1 - C_k \Sigma_k), \quad \beta_{\mathbf{u}} := \|f_{\mathbf{u}}\|_{\mathcal{H}_{\tilde{\kappa}_{\mathbf{u}}}}^2.$$

Taking the k th partial derivative of gives

$$\frac{\partial g}{\partial \Sigma_k} = -\frac{S_k^{\text{in}}}{\Sigma_k^2} + \frac{C_k S_k^{\text{out}}}{(1 - C_k \Sigma_k)^2},$$

where

$$S_k^{\text{in}} = \sum_{\mathbf{u}: k \in \mathbf{u}} \beta_{\mathbf{u}} \prod_{\substack{\ell \in \mathbf{u} \\ \ell \neq k}} \frac{1}{\Sigma_\ell^*} \prod_{m \notin \mathbf{u}} \frac{1}{1 - C_m \Sigma_m^*}, \quad S_k^{\text{out}} = \sum_{\mathbf{u}: k \notin \mathbf{u}} \beta_{\mathbf{u}} \prod_{\ell \in \mathbf{u}} \frac{1}{\Sigma_\ell^*} \prod_{\substack{m \notin \mathbf{u} \\ m \neq k}} \frac{1}{1 - C_m \Sigma_m^*}.$$

Setting the preceding expression equal to zero and solving for Σ_k^* gives

$$\Sigma_k^* = \frac{1}{C_k + \sqrt{\frac{C_k S_k^{\text{out}}}{S_k^{\text{in}}}}} = \frac{\sqrt{S_k^{\text{in}}}}{C_k \sqrt{S_k^{\text{in}}} + \sqrt{C_k S_k^{\text{out}}}}. \quad (3.6)$$

Note that the exact calculation of the optimal parameter Σ_k^* is not possible without knowing the function to calculate the constants $\beta_{\mathbf{u}}$. Additionally, the previous calculations yield coupled systems of equations, which cannot be solved directly. However, the formula (3.6) shows that the optimal TSK factors Σ_k^* can be interpreted as some sort of sensitivity parameters. Observe that S_k^{in} features indices $\mathbf{u}$ that include k . As such, the expression for Σ_k^* has components of the norm of f associated with k divided by terms that represent all of the function's components.

4. WEIGHT LEARNING: ALGORITHMS AND THEORY

Direct calculation of the minimizer of the true norm $g(\Sigma) = \|f\|_\Sigma^2$, described in the previous section, is not possible without knowing the function f exactly. For numerical applications, we propose in this section to calculate the optimal TSK factors by using a discrete analogue instead.

4.1. Finite-data TSK. While Theorem 3.4 establishes the existence of a unique optimal parameter vector Σ^* that minimizes the theoretical RKHS norm $g(\Sigma) = \|f\|_\Sigma^2$, in practical settings the true f is unknown. Instead, we have access to a finite dataset $\{(\mathbf{x}^i, f^i)\}_{i=1}^M$. We therefore optimize an empirical surrogate objective $g_M(\Sigma)$, such as the empirical RKHS norm of the minimum-norm interpolator,

$$g_M(\Sigma) := \mathbf{f}_M^\top \mathbf{K}_{\Sigma, M}^{-1} \mathbf{f}_M, \quad \text{where } \mathbf{K}_{\Sigma, M} = (\kappa_\Sigma(\mathbf{x}^i, \mathbf{x}^j))_{i, j=1}^M. \quad (4.1)$$

Assuming the samples $\mathbf{x}^i$ are distinct (which holds almost surely for a continuous measure μ), the kernel matrix is strictly positive definite and invertible, meaning g_M is well-defined. A fundamental question in kernel learning is whether the data-driven parameters

$$\hat{\Sigma} := \arg \min_{\Sigma \in (0, 1)^d} g_M(\Sigma) \quad (4.2)$$

reliably approximate the theoretical optimum Σ^* , and whether the resulting approximation error remains small. The strong convexity of $g(\Sigma)$ proven in Theorem 3.4 provides the theoretical guarantee for this parameter stability. The following theorem shows uniform convergence of g_M to the true function g .

Theorem 4.1. Let $\{(\mathbf{x}^i, f^i)\}_{i \in \mathbb{N}}$ be a sequence where each $\mathbf{x}^i \in \mathbb{D}^d$ is sampled i.i.d. from a distribution μ , whose support is $\mathbb{D}^d$. Here, $f^i = f(\mathbf{x}^i)$. Let $g_M(\Sigma)$ be defined as in (4.1). Define $\mathbf{f}^M$ to be the vector of length M containing the first M values $f^i = f(\mathbf{x}^i)$. The matrix $\mathbf{K}_{\Sigma, M}$ is the kernel matrix from the first M points $\mathbf{x}^i$ for the kernel κ_Σ . Let g be as defined in Theorem 3.4 as $g(\Sigma) = \|f\|_\Sigma^2$. Then g_M converges uniformly to g on the domain $[\eta, 1 - \eta]^d$ for any $\eta \in (0, 1)$.

Proof. The structure of g_M and g implies they are continuous. For any Σ , the evaluation of $g_M(\Sigma)$ is equal to

$$\min_{\alpha \in \mathbb{R}^M} \alpha^\top \mathbf{K}_{\Sigma, M} \alpha \quad \text{subject to } \mathbf{K}_{\Sigma, M} \alpha = \mathbf{f}^M.$$

This is the smallest norm possible for a function in $\mathcal{H}_\Sigma$ that interpolates $\{(\mathbf{x}^i, f^i)\}_{i=1}^M$. In other words, it is equal to

$$\min_{\hat{f} \in \mathcal{H}_\Sigma} \|\hat{f}\|_\Sigma^2 \quad \text{subject to } \hat{f}(\mathbf{x}^i) = f^i, \quad i = 1, \dots, M.$$

Since f also interpolates the data, $\|\hat{f}_{\Sigma, M}\|_\Sigma \leq \|f\|_\Sigma$, where $\hat{f}_{\Sigma, M}$ is the norm-minimizing interpolating function. Then, $g_M(\Sigma) \leq g(\Sigma)$ for all Σ . In fact, the subset of interpolating functions in $\mathcal{H}_\Sigma$ shrinks as we increase the number of points to interpolate. Then, $g_M(\Sigma) \leq g_{M+1}(\Sigma)$ and the sequence $\{g_M(\Sigma)\}_{M \in \mathbb{N}}$ is monotone increasing.

We show that $g_M(\Sigma) \rightarrow g(\Sigma)$ pointwise. Each pointwise sequence $g_M(\Sigma)$ is monotone increasing and bounded above by $g(\Sigma)$. Then, $g_M(\Sigma) \rightarrow g_\Sigma^*$, with $g_\Sigma^* \leq g(\Sigma)$. Since $\{\hat{f}_{\Sigma, M}\}_{M \in \mathbb{N}}$ is uniformly bounded in $\mathcal{H}_\Sigma$, there is a subsequence $\{\hat{f}_{\Sigma, M_i}\}_{i \in \mathbb{N}}$ that weakly converges to some f_Σ^* in $\mathcal{H}_\Sigma$. As $\mathcal{H}_\Sigma$ is an RKHS, $\langle h, \kappa_\Sigma(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_\Sigma} = h(\mathbf{x})$ for all $h \in \mathcal{H}_\Sigma$ and $\mathbf{x} \in \mathbb{D}^d$. Weak convergence implies pointwise convergence of the subsequence, $\lim_{i \rightarrow \infty} \langle \hat{f}_{\Sigma, M_i}, \kappa_\Sigma(\cdot, \mathbf{x}) \rangle = \langle f_\Sigma^*, \kappa_\Sigma(\cdot, \mathbf{x}) \rangle$. This implies

$$\|f_\Sigma^*\|_\Sigma^2 \leq \liminf_{i \rightarrow \infty} \|\hat{f}_{\Sigma, M_i}\|_\Sigma^2 \leq g_\Sigma^*.$$

Suppose $\{\mathbf{x}^i\}_{i=1}^\infty$ is dense in $\mathbb{D}^d$, which is true almost surely. Because $f_\Sigma^*(\mathbf{x}^i) = f^i$ for all i , f_Σ^* and f are equal to each other on a dense subset of $\mathbb{D}^d$. Since the kernel κ_Σ is continuous, all functions in the RKHS $\mathcal{H}_\Sigma$ are continuous. Continuity of both f and f_Σ^* implies $f_\Sigma^* = f$. So, $g(\Sigma) \leq g_\Sigma^*$, meaning $g(\Sigma) = g_\Sigma^*$.

Because all g_M and g are continuous and the sequence is monotone increasing and converges pointwise, Dini's theorem [41] yields uniform convergence $g_M \rightarrow g$ on any compact subset of $(0, 1)^d$. $\square$

The previous theorem shows that the empirical surrogate g_M uniformly approximates the true objective g . A direct consequence is that the error between the TSK factors from the empirical minimizing problem and the true minimizer of the objective g is bounded, as the following corollary shows.

Corollary 4.2. By Theorem 4.1, for every $M \in \mathbb{N}$ there exists $\epsilon_M > 0$ such that

$$\sup_{\Sigma \in [\eta, 1 - \eta]^d} |g(\Sigma) - g_M(\Sigma)| \leq \epsilon_M,$$

where $\eta \in (0, 1)$. Suppose g is m -strongly convex, as given by Theorem 3.4. Then, the error between the TSK factors $\hat{\Sigma}_M$ derived by the empirical minimization (4.2) and the true minimizer Σ^* of the function g is bounded by

$$\|\hat{\Sigma}_M - \Sigma^*\| \leq 2\sqrt{\frac{\epsilon_M}{m}}.$$

Proof. Using Theorem 3.4 and the definition of strong convexity evaluated at the minimum Σ^* (see [35]), there exists a constant $m > 0$ such that for any $\Sigma \in [\eta, 1 - \eta]^d$,

$$\frac{m}{2} \|\Sigma - \Sigma^*\|^2 \leq g(\Sigma) - g(\Sigma^*).$$

Evaluating this at the empirical minimizer $\hat{\Sigma}_M$ and utilizing the uniform bound yields

$$\begin{aligned} \frac{m}{2} \|\hat{\Sigma}_M - \Sigma^*\|^2 &\leq g(\hat{\Sigma}_M) - g(\Sigma^*) = g(\hat{\Sigma}_M) - g_M(\hat{\Sigma}_M) + g_M(\hat{\Sigma}_M) - g(\Sigma^*) \\ &\leq g(\hat{\Sigma}_M) - g_M(\hat{\Sigma}_M) + g_M(\Sigma^*) - g(\Sigma^*) \leq |g(\hat{\Sigma}_M) - g_M(\hat{\Sigma}_M)| + |g_M(\Sigma^*) - g(\Sigma^*)| \\ &\leq 2 \sup_{\Sigma \in [\eta, 1 - \eta]^d} |g_M(\Sigma) - g(\Sigma)| \leq 2\epsilon_M. \end{aligned}$$

Here, the first inequality relies on the definition of the empirical minimizer, which guarantees $g_M(\hat{\Sigma}) \leq g_M(\Sigma^*)$. Rearranging this inequality directly provides a bound on the parameter estimation error,

$$\|\hat{\Sigma}_M - \Sigma^*\| \leq 2\sqrt{\frac{\epsilon_M}{m}}.$$

This finishes the proof. $\square$

The previous corollary demonstrates the critical role of strong convexity in ensuring parameter identifiability and stability. Because the parameters $\hat{\Sigma}$ converge to Σ^* at a rate governed by the statistical estimation error ϵ_M , and because the kernel mapping $\Sigma \mapsto \kappa_\Sigma$ is smooth, the learned kernel $\kappa_{\hat{\Sigma}}$ will uniformly converge to the optimal kernel κ_{Σ^*} . Consequently, the approximation error of the model trained with the learned parameters $\hat{\Sigma}$ will only deviate from the optimal approximation error by a term proportional to $\mathcal{O}(\sqrt{\epsilon_M})$. This confirms that optimizing the TSK factors Σ from finite data is mathematically well-posed and preserves the optimal generalization properties of the minimum-norm interpolator.

At present, we cannot guarantee convexity of the finite-data objective, unlike the exact version (3.1). We speculate that the finite-data objective may be convex with high probability under random data sampling. Empirical evidence supports this. In numerical experiments, we have computed the analytical Hessian of g_M at randomly sampled values of Σ throughout $(0, 1)^d$ and consistently observe it to be positive definite.

Connection to automatic relevance determination

The finite-data objective (4.1) is related to kernel hyperparameter learning by maximizing Gaussian process marginal likelihood. Automatic relevance determination (ARD) adapts a kernel through input-specific length scales [51, Section 5.1]. For a Gaussian kernel, ARD takes the form

$$\kappa_\ell(\mathbf{x}, \mathbf{y}) = e^{-\frac{1}{2} \sum_{k=1}^d (x_k - y_k)^2 / \ell_k^2}, \quad \ell_k > 0.$$

A larger value of ℓ_k makes the kernel less sensitive to variation in the k th input. Assuming the data $\mathbf{f}$ are noise-free, maximizing the Gaussian process marginal likelihood is equivalent to

$$\ell^* = \arg \min_{\ell > 0} \mathbf{f}^\top \mathbf{K}_\ell^{-1} \mathbf{f} + \log(\det \mathbf{K}_\ell). \quad (4.3)$$

The first term in (4.3) is the squared RKHS norm of the minimum-norm interpolant and is therefore analogous to the finite-data TSK objective $g_M(\Sigma) = \mathbf{f}^\top \mathbf{K}_\Sigma^{-1} \mathbf{f}$. The two approaches nevertheless adapt the approximation space in different ways. ARD rescales the inputs of the kernel and includes the log-determinant term arising from marginal likelihood. On the other hand, TSK changes the relative weights of the ANOVA component spaces and selects these weights solely through RKHS norm minimization.

4.2. The Algorithm. Our proposed learning algorithm consists of two steps: 1) TSK factor optimization; 2) kernel ridge regression. TSK factor optimization consists of finding a minimizer of the finite-data objective (4.1). The gradient and Hessian of the objective have explicit expressions. We opt for quasi-Newton methods to save on cost. Particularly, we use the L-BFGS algorithm [31, 6] to minimize the finite-data objective. Our method is summarized in Algorithm 1. We must address some numerical concerns that can

Algorithm 1 Total sensitivity kernel learning

Inputs: Data $\{(\mathbf{x}^i, f^i)\}_{i=1}^M$ with data vector $\mathbf{f}^M = (f^i)_{i=1}^M$, characteristic functions ϕ_k , regularization parameter λ , initialization $\Sigma^{(0)}$

Outputs: TSK approximation f^* , TSK factors Σ^*

- 1: Set initial TSK factors to $\Sigma^{(0)}$
- 2: Precompute, for arbitrary TSK factors Σ , the kernel matrix by

$$(\mathbf{K}_{\Sigma, M})_{i,j} = \prod_{k=1}^d (1 - \Sigma_k + \Sigma_k \phi_k(x_k^i - x_k^j))$$

- 3: Compute minimizer $\Sigma^* = \arg \min_{\Sigma} g_M(\Sigma) = (\mathbf{f}^M)^\top \mathbf{K}_{\Sigma, M}^{-1} \mathbf{f}^M$
 - 4: Solve $f^* = \arg \min_{\hat{f} \in \mathcal{H}_\Sigma} \frac{1}{M} \sum_{i=1}^M (\hat{f}(\mathbf{x}^i) - f^i)^2 + \lambda \|\hat{f}\|_\Sigma^2$
-

cause the algorithm to fail.

Reparameterization. The optimization takes place over TSK factors $\Sigma \in (0, 1)^d$. To ensure that the iterates remain within this admissible region, we use a sigmoid reparameterization,

$$\Sigma_k = s(z_k) = \frac{1}{1 + e^{-z_k}}, \quad z_k \in \mathbb{R}.$$

The optimization is then performed over the unconstrained variables $\mathbf{z} \in \mathbb{R}^d$. This prevents optimization iterates from leaving the region where the TSK factors are defined. Centering and scaling the input data also improves the numerical behavior of the kernel matrix. For kernels like the Gaussian kernel, differences in the magnitudes of the inputs can cause excessively small or disparate kernel values. Centering and scaling ensure the data are not spaced too far apart.

Initialization matters. Another issue is that optimization can stall when the input dimension is high. This results from how the input dimension impacts the structure of the kernel matrix in (4.1). Recall the kernel matrix has entries

$$(\mathbf{K}_{\Sigma, M})_{i,j} = \prod_{k=1}^d (1 - \Sigma_k + \Sigma_k \phi_k(x_k^i - x_k^j)).$$

Each entry is a product of convex combinations. Since each ϕ_k is a characteristic function, the diagonal entries are always equal to 1. The off-diagonal entries are a different story, as they correspond to pairs of distinct data points. Consider the standard kernel, when all TSK factors equal 1, whose entries are just products of characteristic functions with $|\phi_k| \leq 1$. For some characteristic functions commonly used in kernel learning, like the Gaussian characteristic function, the value decays to 0 as the distance between the two data points increases. Off-diagonal entries for the standard kernel are then products of d values with magnitude less than 1. Larger input dimensions d accelerate these products towards 0. The result is a kernel matrix that, numerically speaking, approaches the identity matrix. This phenomenon can occur for TSK factor values near 1, too, such as 0.9 or 0.95.

The same phenomenon affects the gradient of the finite-data objective. The gradient is given by entries

$$\frac{\partial g_M}{\partial \Sigma_\ell} = -\boldsymbol{\alpha}^\top \frac{\partial \mathbf{K}_{\Sigma, M}}{\partial \Sigma_\ell} \boldsymbol{\alpha}, \quad \frac{\partial (\mathbf{K}_{\Sigma, M})_{i,j}}{\partial \Sigma_\ell} = (\phi_\ell(x_\ell^i - x_\ell^j) - 1) \prod_{k \neq \ell} (1 - \Sigma_k + \Sigma_k \phi_k(x_k^i - x_k^j)).$$

The diagonal entries of these derivative matrices are 0. When the TSK factors are near 1, the off-diagonal entries can also become numerically close to 0 in high dimensions because they contain the same type of product as the kernel matrix. The gradient of g_M can therefore become numerically small even when the TSK factors are far from the minimizer. The effect is that, numerically speaking, the function $g_M(\Sigma)$ appears nearly constant in the region of near-1 TSK factors, and optimization stalls.

One must be careful about where to initialize the optimizer. We initialize the TSK factors away from 1. Unless otherwise stated, we take

$$\Sigma_k^{(0)} = 0.2, \quad k = 1, \dots, d.$$

This appears to mitigate the vanishing of the off-diagonal entries of the kernel matrix and its derivatives. The corresponding initialization for the unconstrained optimization variables is

$$z_k^{(0)} = \log \left(\frac{\Sigma_k^{(0)}}{1 - \Sigma_k^{(0)}} \right).$$

5. NUMERICAL EXPERIMENTS

The selected numerical experiments display how TSKs exploit multivariable structure. To optimize TSK factors, we initialize each one at a value of $\Sigma_k^{(0)} = 0.2$. Optimization uses the quasi-Newton L-BFGS method. The exact gradient is supplied. This methodology holds for all experiments. All validation sets use 10^4 points and the same distribution as the training data.

We compare our method TSK with the standard tensor product kernel

$$\kappa_{\mathbf{1}}(\mathbf{x}, \mathbf{y}) = \prod_{k=1}^d \phi_k(x_k - y_k), \tag{5.1}$$

and the ANOVA kernel (1.2) with the same one-dimensional kernels as for TSK. Additionally, for the Gaussian kernels we also compare with ARD, described in (4.3).

5.1. Case study: Sobol’ g-function. The Sobol’ g-function, from [42], appears frequently as a sensitivity analysis benchmark. It typically takes 8 inputs

$$f(\mathbf{x}) = \prod_{k=1}^8 \frac{|4x_k - 2| + a_k}{1 + a_k}, \quad (5.2)$$

where $\mathbf{x} \sim \mathcal{U}([0, 1]^8)$. The coefficient a_k determines the importance of the corresponding input. A greater value of a_k decreases the importance of the input x_k : x_k is very important when $a_k \approx 0$, is of lesser importance when $a_k \approx 9$, and makes almost no impact when $a_k \approx 99$ [34]. This control over the input structure makes the g-function an ideal case study. We choose three sets of coefficients $\mathbf{a}_{\text{high}}$, $\mathbf{a}_{\text{med}}$, and $\mathbf{a}_{\text{low}}$ to create high, medium, and low complexity input interaction structures,

$$\begin{aligned} \mathbf{a}_{\text{high}} &= [0 \ 0 \ 1 \ 1 \ 2 \ 2 \ 3 \ 3], & \text{type A,} \\ \mathbf{a}_{\text{med}} &= [0.5 \ 0.5 \ 7 \ 7 \ 7 \ 7 \ 7 \ 7], & \text{type B,} \\ \mathbf{a}_{\text{low}} &= [1 \ 2 \ 5 \ 10 \ 20 \ 50 \ 100 \ 500], & \text{type C.} \end{aligned}$$

The case study demonstrates how TSKs identify and take advantage of structurally low complexity functions. A function is easy or difficult to approximate depending on its ANOVA decomposition. This is formalized with notions of type A, B, and C functions in [25]. A function is complicated (or type A) and challenging to approximate, when no inputs are unimportant and all interactions, low-order to high-order, are important. A function of medium complexity (or type B) has no non-influential inputs, but is dominated by main effects and lower-order interaction terms. The low complexity (or type C) case describes functions where some inputs are unimportant and only lower-order interactions and main effects have influence. Each set of coefficients $\mathbf{a}_{\text{high}}$, $\mathbf{a}_{\text{med}}$, and $\mathbf{a}_{\text{low}}$ falls into either type A, B, or C.

We run Algorithm 1 with $M = 10^3$ samples to create kernel approximations of each case of the g-function (5.2). Because of the nonsmoothness of the g-function, the kernels use the characteristic function $\phi(t) = e^{-|t|}$, which corresponds to the Cauchy distribution. The results for all cases appear in Fig. 1. Scatter plots show the true output values versus what the kernel predicts for each point in the validation set. We present the optimized TSK factors in bar graphs.

Root mean squared error (RMSE) and root relative squared error (RRSE) are evaluated on a validation set of 10^4 points sampled via LHS. The errors are given in Table 5.1.

Complexity	Error Type	TSK	Standard Ker.	ANOVA Ker.
High	RMSE	0.565	0.671	0.507
	RRSE	0.367	0.437	0.330
Medium	RMSE	9.570×10^{-2}	0.233	0.139
	RRSE	8.166×10^{-2}	0.199	0.119
Low	RMSE	2.571×10^{-2}	0.128	6.008×10^{-2}
	RRSE	2.412×10^{-2}	0.120	5.637×10^{-2}

TABLE 1. Root mean squared error (RMSE) and root relative squared error (RRSE) for kernel machines of the TSK, standard kernel, and ANOVA kernel on a validation set. Validation sets consist of 10^4 data points for the high, medium, and low complexity g-function.

In each complexity case, TSK shows better performance than the standard kernel and performs roughly similar to or better than the ANOVA kernel. Decreasing complexity leads to better performance regardless of the kernel used. The least complex function is easier to approximate. However, TSK appears better able to take advantage of decreasing complexity. In the high complexity case, the performance gap is smaller. We see the optimal TSK factors in Fig. 1 (left column) identify all inputs as important, as all factors take values tending towards 1 rather than 0. The optimized kernel is not so different from the standard kernel. Visually, in Fig. 1 (left column), the TSK predictions do not correlate much better with the true values than those of the standard kernel. The situation changes with the medium complexity case where the TSK performs

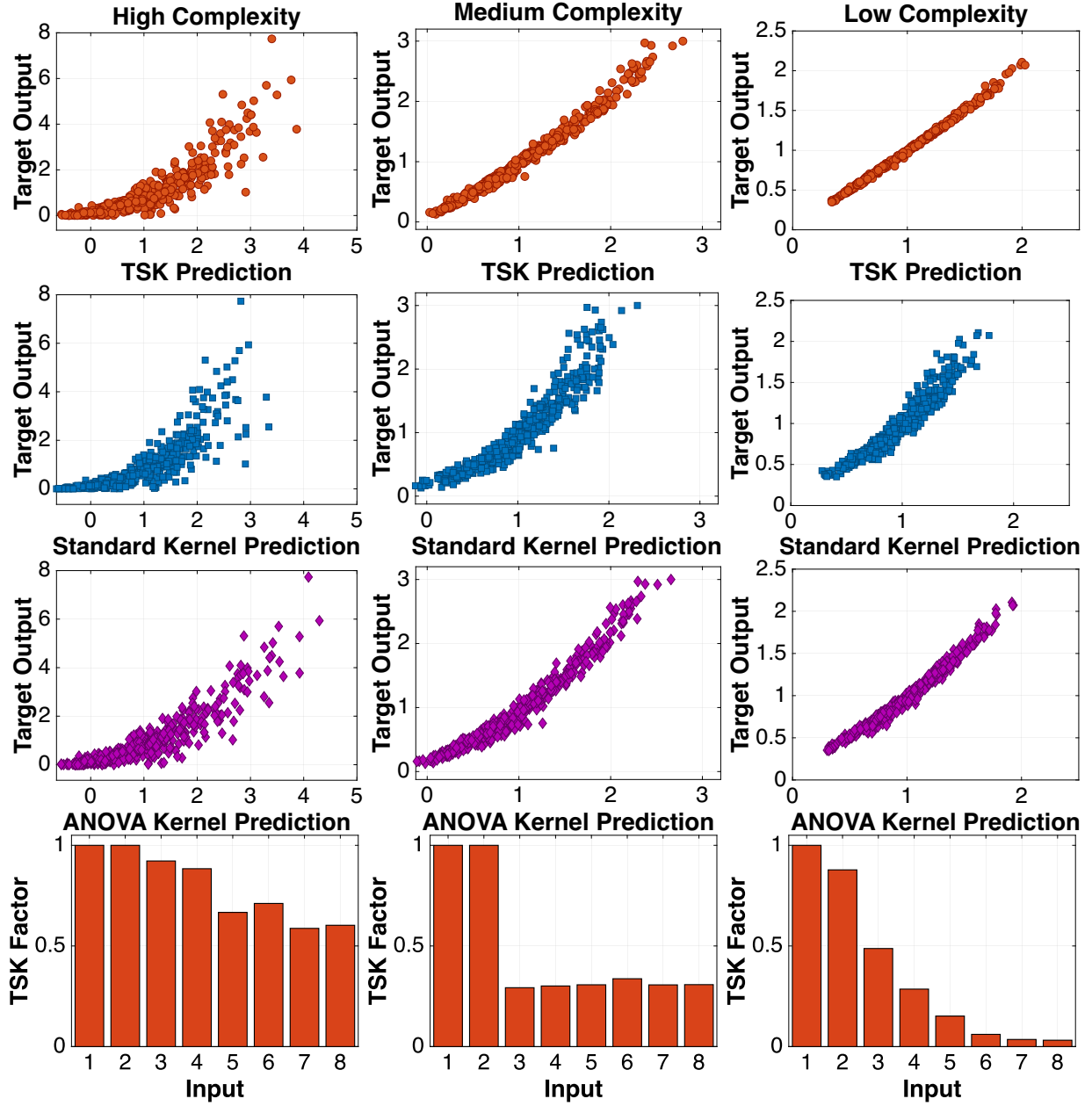


FIGURE 1. Results for approximation of the g-function (5.2) with TSK (top row), standard product kernels (second row), and unweighted ANOVA kernels (third row). The left, middle, and right columns correspond to high, medium, and low complexity structured functions, respectively. We show correlations between the kernel predictions and true outputs for 500 of the 10^4 validation points. Strong correlation, and more accurate prediction, occurs when the scatter plot is closer to a line with slope one. The bottom row presents the TSK factors taken after optimization.

twice as well as the standard kernel. There is a starker difference in correlation of the predictions in Fig. 1 (middle column). There is also more disparity in the TSK factors, which aligns with the choice of $\mathbf{a}_{\text{med}}$. For the low complexity case, the TSK error is less than one fourth of the standard-kernel error and half of the ANOVA-kernel error. In Fig. 1 (right column), the TSK predictions correlate almost perfectly.

The experiment highlights how the optimized kernel performs better as the standard kernel across all cases. Key is how the performance of the optimized kernel κ_{Σ^*} improves over the standard kernel and ANOVA kernel when the function’s interaction structure becomes simpler. This demonstrates how optimizing TSK factors allows the method to identify and take particular advantage of function structure. The ANOVA kernel performs better than the standard kernel by its greater expression of structure. Unlike TSK, the lack of weighting in the ANOVA kernel limits performance.

5.2. Higher-dimensional examples. The following experiments investigate how beneficial it is to adapt the kernel to the multivariable structure as the input dimension increases and the number of available function evaluations remains limited. The kernels are Gaussian, using the characteristic function $\phi(t) = e^{-t^2/2}$, which corresponds to the standard normal distribution. High dimensions present a particular difficulty for the standard product kernel. As discussed in Section 4, high input dimension results in a kernel matrix κ_1 near the identity matrix. When evaluating on the validation set, we construct a prediction kernel matrix built from both the training and validation data. Since the training and validation sets almost surely do not share any data, the prediction kernel matrix can be nearly a zero matrix. The ANOVA and TSK kernels alter this product structure by including lower-dimensional component kernels. Use of Gaussian kernels enables comparison to automatic relevance determination (ARD), which can mitigate the effect by adapting coordinate-wise length scales. The experiments below examine these different mechanisms on functions possessing varying forms of multivariable structure. We study four examples, two of which have smooth analytic expressions.

100D function. The 100D function is a benchmark studied in [33]. The model’s expression is

$$f(\mathbf{x}) = 3 + \frac{1}{100} \sum_{k=1}^{100} k(x_k^3 - 5x_k + \frac{1}{3} \log(x_k^2 + x_k^5)) + x_1x_2^2 + x_2x_4 - x_3x_5 + x_{51} + x_{50}x_{54}^2, \quad (5.3)$$

where $x_{20} \sim \mathcal{U}([1, 3])$ and $x_k \sim \mathcal{U}([1, 2])$ for $k \neq 20$.

Previous results of variance-based global sensitivity analysis show that inputs x_2, x_{20}, x_{51} , and x_{54} are distinctly more important than the remaining inputs. Among the remaining inputs, input importance tends to increase as the input index increases, while x_1, x_3, x_4, x_5 , and x_{50} also have some importance. The 100D function’s expression in (5.3) explains this. Importance scales with the input index because the summand terms are multiplied by the input index k . Inputs $x_1, x_2, x_3, x_4, x_5, x_{50}, x_{51}, x_{54}$ stand out in importance because they appear in extra terms outside the sum. The input x_{20} has special importance because its distribution is unique and wider than that of the other inputs. While all 100 inputs appear in the expression, the 100D function still has a low complexity structure because it features only main effects and second-order interactions.

The scatter plots in Fig. 2 show how well the kernel predictions correlate with true output values. The errors in Table 2 show both TSK and ARD approximate the function accurately. In contrast, the standard kernels and unweighted ANOVA kernels are much less accurate.

The optimal TSK factors in Fig. 2 recover meaningful structure. The TSK factors do not match Sobol’ indices in value. The ranking of the inputs by the TSK factors matches the ranking given by the Sobol’ indices in [33]. This supports that the process of optimizing the TSK factors learns meaningful multivariable structure of the 100D function.

One-dimensional diffusion model. This benchmark is a one-dimensional stochastic diffusion problem. The problem first appeared as a benchmark in [44] and subsequently appeared in [33]. Consider the boundary value problem

$$\begin{aligned} -\frac{d}{dz}(E(z)\frac{du}{dz}(z)) + h(z) &= 0, \quad z \in (0, L), \\ u(0) &= 0, \\ \frac{du}{dz}(L) &= F. \end{aligned} \quad (5.4)$$

The diffusion term is a lognormal random field $E(z) = \exp(\lambda_E + \zeta_E G(z))$, where $G(z)$ is a standard normal stationary Gaussian random field. Suppose we wish to determine how $u(L)$, the solution at the other boundary, depends on the realization of $G(z)$. We represent $G(z)$ through its Karhunen-Loève (KL)

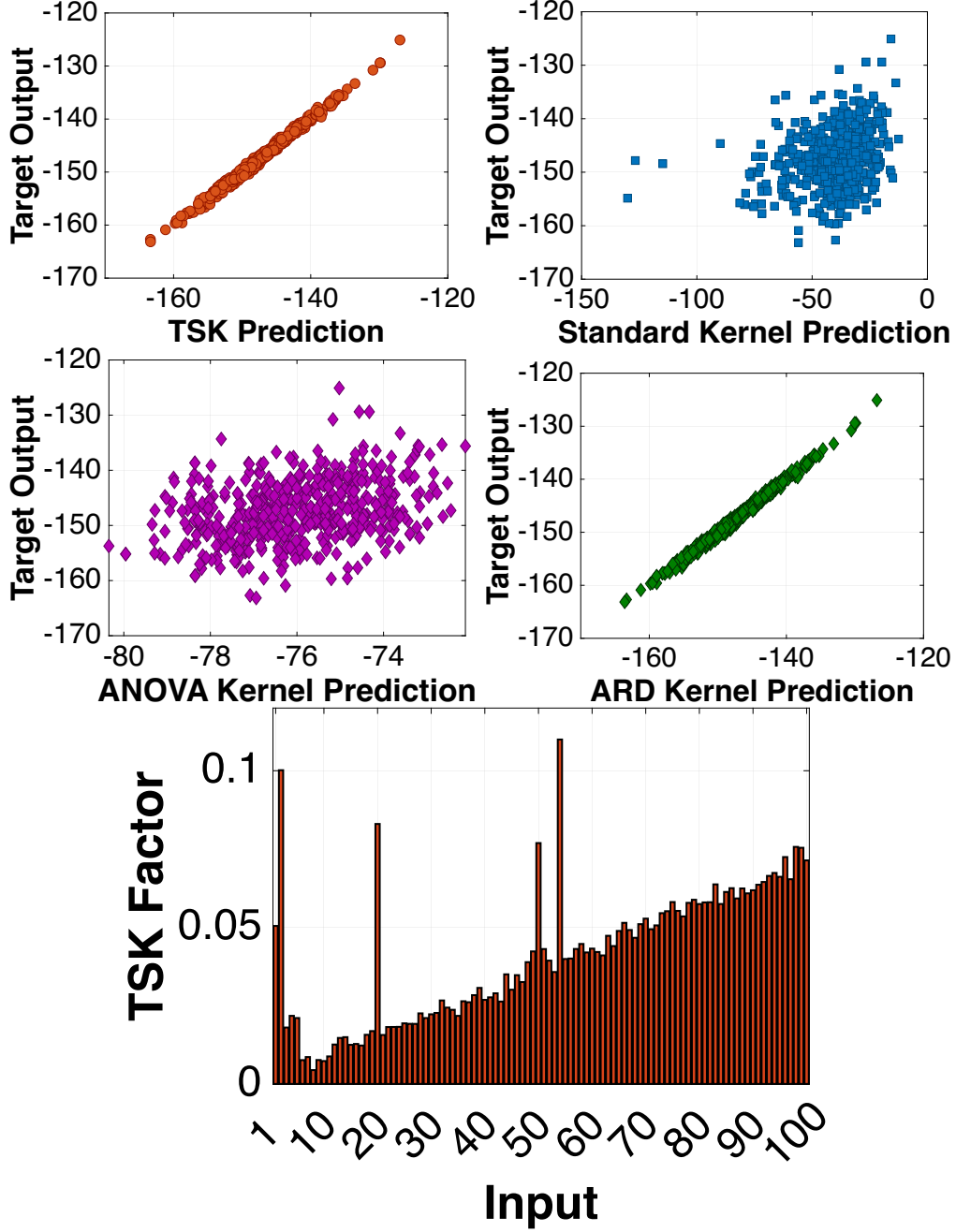


FIGURE 2. Results for approximation of the 100D function with TSK, the standard product kernel, the unweighted ANOVA kernel, and ARD, trained on 1000 function evaluations. Top figures display kernel predictions versus true outputs on 500 out of the 10^4 validation points. Strong correlation, and more accurate prediction, occurs when the scatter plot is closer to a line with slope one. The bottom figure presents the TSK factors taken after optimization.

expansion

$$G(z) = \sum_{k=1}^{\infty} \sqrt{\alpha_k} \varphi_k(z) \xi_k, \quad \xi_k \sim \mathcal{N}(0, 1).$$

The KL expansion is truncated to its first 62 terms to preserve 99% of the random field’s variance. This transforms the problem into a deterministic, finite-dimensional one. Define the inputs

$$x_k = \xi_k, \quad k = 1, \dots, 62.$$

We aim to construct a surrogate for $f(\mathbf{x}) := u(L; \mathbf{x})$, the boundary value we get by solving (5.4) for a realization of $\mathbf{x}$.

The training and validation sets come from an open-source data set created for the benchmark problem [20]. The training set uses $M = 600$ samples from the standard normal distribution. The 10^4 validation samples are likewise samples from the standard normal distribution. The results in Fig. 3 demonstrate the importance of adapting the kernel to the strongly anisotropic structure of this problem. The standard product kernel gives predictions close to zero and an RRSE of approximately one. The unweighted ANOVA kernel performs similarly, indicating that introducing lower-dimensional ANOVA components without adapting their relative importance is insufficient in this example. Both TSK and ARD produce accurate approximations.

The optimized TSK factors indicate that only a relatively small subset of the 62 KL coefficients contributes substantially to the quantity of interest. Thus, the improvement of TSK over the two fixed kernels is consistent with its ability to adapt to unequal input importance. Variable relevance adaptation through individual kernel length scales provides a particularly effective representation.

Modified Griewank function. The Griewank function is an optimization benchmark [19] structurally composed of polynomial main effect terms and a trigonometric interaction term. We modify the polynomial terms to make the inputs more unequal in importance.

$$f(\mathbf{x}) = \frac{1}{4000} \sum_{k=1}^{40} \frac{x_k^2}{k^2} - \prod_{k=1}^{40} \cos\left(\frac{x_k}{\sqrt{k}}\right) + 1, \quad \mathbf{x} \sim \mathcal{U}([-600, 600]^{40}). \quad (5.5)$$

In terms of multivariable structure, the importance of the polynomial main effects decreases with the input index. The trigonometric product introduces non-additive dependence across all 40 inputs and, when viewed through an ANOVA decomposition, generates interaction components of multiple orders. The resulting function therefore combines strongly anisotropic input importance with richer interaction structure than the preceding high-dimensional examples. In this respect, it is more similar to the g-function.

The training set uses $M = 10^3$ samples from the uniform distribution $\mathcal{U}([-600, 600]^{40})$. The modified Griewank function produces a different comparison between TSK and ARD than the preceding examples. The TSK approximation error is an order of magnitude smaller than the ARD error.

Summary

The three high-dimensional examples highlight the necessity of identifying multivariable structure and incorporating it into the approximation. The errors in Table 2 show that TSK factor optimization can improve performance over the standard kernel by at least two orders of magnitude. The standard product kernel performs poorly in each example, while the unweighted ANOVA kernel provides only modest improvement and performs nearly identically to the standard kernel for the two diffusion problems. Simply introducing an ANOVA decomposition is not sufficient when the contributions of different inputs and interactions are strongly unequal. Both TSK and ARD address this limitation through data-driven kernel adaptation and substantially improve approximation accuracy.

The comparison between TSK and ARD also illustrates that the two methods exploit multivariable structure differently. For the 100D function, which has unequal input importance but only main effects and pairwise interactions, TSK and ARD achieve nearly identical errors. For the diffusion problem, the learned TSK factors indicate that the quantity of interest depends much more strongly on a relatively small subset of the stochastic inputs, and ARD provides the most accurate approximation. In contrast, TSK performs substantially better than ARD for the modified Griewank function, which combines strongly unequal main effects with interaction components of multiple orders. These results suggest that ARD length scale adaptation and adaptation of ANOVA-component weights provide complementary mechanisms for exploiting multivariable structure. No single adaptive mechanism performs best across all four examples, but TSK

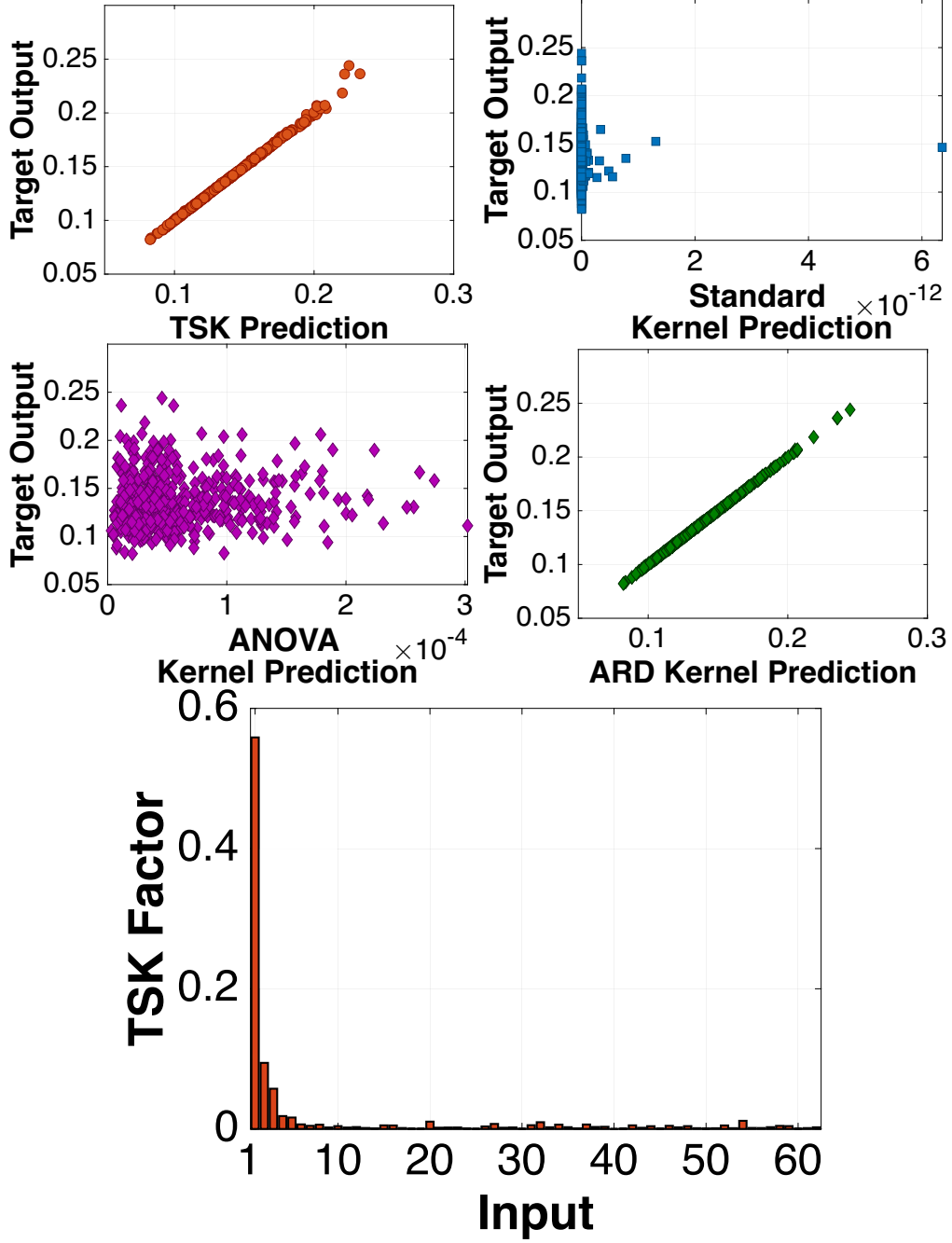


FIGURE 3. Results for approximation of the one-dimensional diffusion model with TSK, the standard product kernel, the unweighted ANOVA kernel, and ARD, trained on 600 function evaluations. Top figures display kernel predictions versus true outputs on 500 out of the 10^4 validation points. Strong correlation, and more accurate prediction, occurs when the scatter plot is closer to a line with slope one. The bottom figure presents the TSK factors Σ^* taken after optimization.

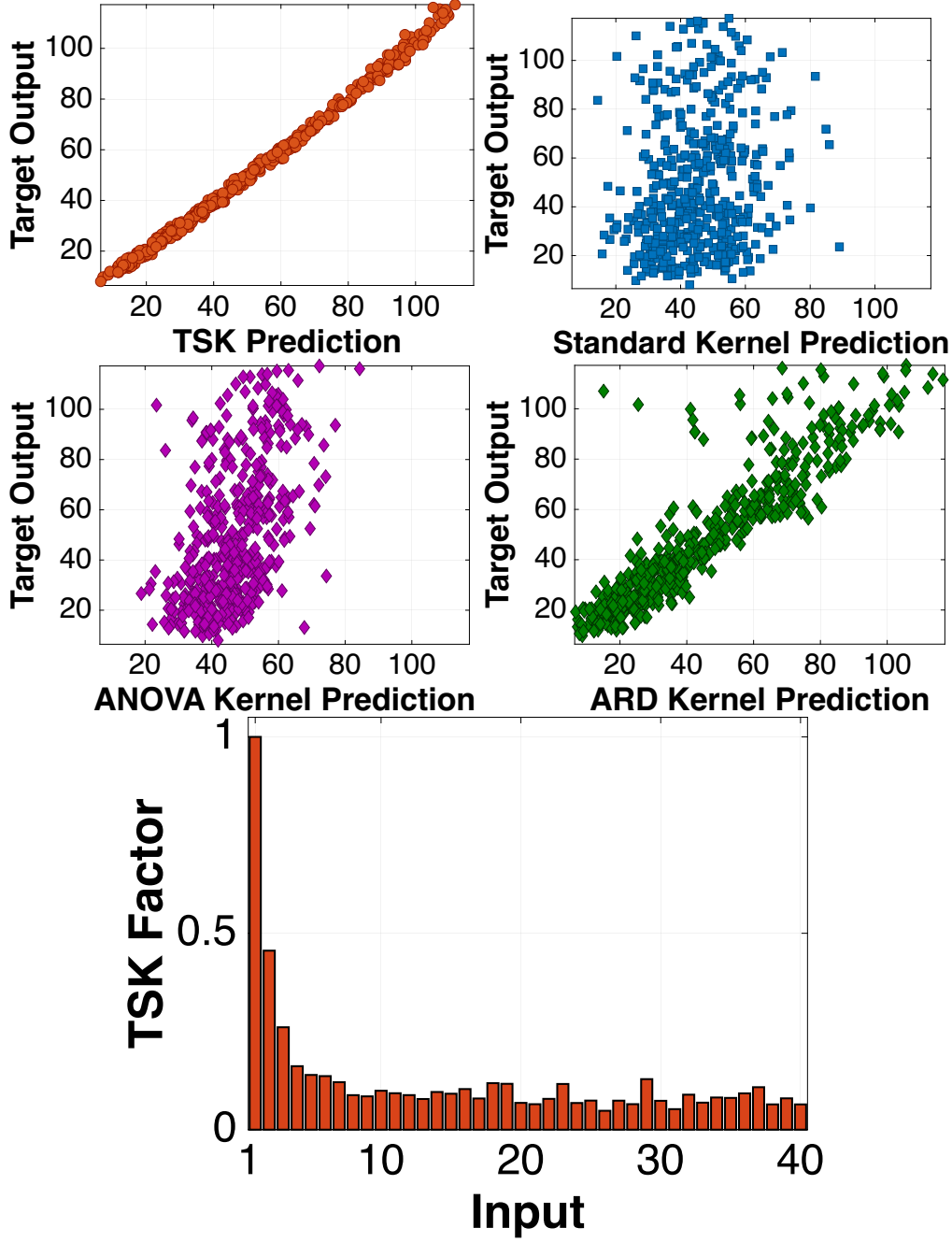


FIGURE 4. Results for approximation of the modified Griewank function with TSK, the standard product kernel, the unweighted ANOVA kernel, and ARD, trained on $M = 10^4$ function evaluations. Top figures display kernel predictions versus true outputs on 500 out of the 10^4 validation points. Strong correlation, and more accurate prediction, occurs when the scatter plot is closer to a line with slope one. The bottom figure presents the TSK factors taken after optimization.

Model	Error Type	TSK	Standard Ker.	ANOVA Ker.	ARD
100D	RMSE	0.480	108.492	71.667	0.479
	RRSE	3.255×10^{-3}	0.736	0.486	3.251×10^{-3}
1D Diff.	RMSE	2.027×10^{-3}	0.139	0.139	2.755×10^{-4}
	RRSE	1.461×10^{-2}	1.000	0.999	1.985×10^{-3}
Grienwank	RMSE	1.1814	29.267	24.643	11.900
	RRSE	3.188×10^{-2}	0.515	0.433	0.209

TABLE 2. Root mean squared error (RMSE) and root relative squared error (RRSE) for approximations of the 100D function, 1D diffusion problem, and modified Grienwank function using TSK, the standard product kernel, the unweighted ANOVA kernel, and the ARD kernel. Errors are taken on validation sets of 10^4 points.

provides substantial improvements when its learned weighting of the approximation space aligns with the structure of the target function.

6. CONCLUSION AND OUTLOOK

This work develops a framework for adapting a weighted RKHS to the multivariable structure of an unknown target function. TSKs provide a structured parameterization of ANOVA kernel weights, reducing 2^d possible component weights to d input-specific TSK factors. Rather than prescribing these weights a priori or selecting a sparse collection of components, we learn the TSK factors by seeking the RKHS where the target function has minimum norm. Under suitable conditions, these TSK factors are unique to a target function. The numerical experiments demonstrate that the resulting adaptation improves approximation when the target possesses exploitable multivariable structure. This perspective on adaptive multivariable approximation argues for adapting the geometry of the approximation space to the structure of the target function.

Avenues for future work concern the scope and computational implementation of the method. The product nature of weights induced by TSK factors restricts the method to certain families of weights. While acceptable for many functions, this is unsuitable for functions with complicated interactions, especially functions that feature higher-order interaction terms but no low-order interaction terms. Quasi-Monte Carlo literature offers many examples of weight structures, including product-order-dependent weights [26, 23]. Future work should investigate whether TSK theory can extend to other weight structures. Improved computational scalability for large datasets can be addressed by the use of random features [40] instead of using kernels corresponding to the characteristic functions ϕ_k directly. The structural adaptability of TSKs makes them attractive for direct integration into established kernel-based machine learning methods like automatic relevance determination [53] and multiple kernel learning [18]. While our current theoretical results focus on exact minimum-norm interpolation, extending these uniform error bounds and convergence guarantees to settings with noisy observational data is a crucial next step. In the current work, we establish the method in the approximation of functions with independently distributed inputs. An important future direction is to extend the framework to settings with dependent inputs. Finally, given the recent exploration of sensitivity analysis beyond the classical ANOVA decomposition setting [12, 28, 50, 30], it would be interesting to study how TSK factors relate to alternative notions of sensitivity, especially in kernel-based sensitivity indices.

Acknowledgments. This work is partially supported by the National Science Foundation (NSF) under grant number 2038118 (John E. Darges). This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 569580074 (Laura Weidensager).

REFERENCES

- [1] R. Agrawal and T. Broderick. The SKIM-FA Kernel: High-Dimensional Variable Selection and Nonlinear Interaction Discovery in Linear Time. *J. Mach. Learn. Res.*, 24(27):1–60, 2023.
- [2] A. Antoniadis. Analysis of variance on function spaces. *Statistics*, 15(1):59–71, 1984.
- [3] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68(3):337–404, 1950. doi:10.21236/ada296533.

- [4] F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pages 105–112, Red Hook, NY, USA, 2008. Curran Associates Inc.
- [5] F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *J. Mach. Learn. Res.*, 18(1):714 – 751, 2017.
- [6] A. S. Berahas, J. Nocedal, and M. Takáč. A multi-batch L-BFGS method for machine learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1063–1071, Red Hook, NY, USA, 2016. Curran Associates Inc.
- [7] S. Bochner. *Harmonic Analysis and the Theory of Probability*. University of California Press, Berkeley, CA, USA, 1 edition, 1955. doi:10.1525/9780520345294.
- [8] R. E. Caflisch, W. Morokoff, and A. B. Owen. Valuation of mortgage-backed securities using Brownian bridges to reduce effective dimension. *J. Comput. Finance*, 1:27–46, 1997.
- [9] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007.
- [10] A. Chertkov, G. Ryzhakov, and I. Oseledets. Black Box Approximation in the Tensor Train Format Initialized by ANOVA Decomposition. *SIAM J. Sci. Comput.*, 45(4):A2101–A2118, 2023. doi:10.1137/22m1514088.
- [11] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39(1):1–49, 2002.
- [12] S. da Veiga. Kernel-based ANOVA decomposition and Shapley effects – Application to global sensitivity analysis, 2021. Preprint at <https://arxiv.org/abs/2101.05487>.
- [13] J. E. Darges, A. Alexanderian, and P. A. Gremaud. Extreme learning machines for variance-based global sensitivity analysis. *Int. J. Uncertain. Quantif.*, 14(4):83–103, 2024. doi:10.1615/int.j.uncertaintyquantification.2024049519.
- [14] J. Dick. Random weights, robust lattice rules and the geometry of the cbrc algorithm. *Numer. Math.*, 122(3):443–467, 2012. doi:10.1007/s00211-012-0469-5.
- [15] J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: The quasi-Monte Carlo way. *Acta Numer.*, 22:133–288, 2013. doi:10.1017/s0962492913000044.
- [16] N. Durrande, D. Ginsbourger, O. Roustant, and L. Carraro. ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis. *J. Multivariate Anal.*, 115:57 – 67, 2013. doi:10.1016/j.jmva.2012.08.016.
- [17] Z. Fang, I. Kim, and P. Schaumont. Flexible variable selection for recovering sparsity in nonadditive nonparametric models. *Biometrics*, 72(4):1155–1163, 2016. doi:10.1111/biom.12518.
- [18] M. Gönen and E. Alpaydm. Multiple kernel learning algorithms. *J. Mach. Learn. Res.*, 12:2211–2268, 2011.
- [19] A. Griewank. Generalized descent for global optimization. *J. Optim. Theory Appl.*, 34:11–39, 1981.
- [20] A. Hlobilová, S. Marelli, and B. Sudret. Surrogate modeling benchmark - one-dimensional diffusion model, 2024. Retrieved from <https://doi.org/10.5281/zenodo.12704504>.
- [21] W. Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. *Ann. Math. Statist.*, 19(3):293 – 325, 1948.
- [22] T. Hofmann, B. Schölkopf, and A. Smola. Kernel methods in machine learning. *Ann. Statist.*, 36:1171–1220, 2008. doi:10.1214/009053607000000677.
- [23] V. Kaarnioja, Y. Kazashi, F. Kuo, F. Nobile, and I. Sloan. Fast approximation by periodic kernel-based lattice-point interpolation with application in uncertainty quantification. *Numer. Math.*, 150(1):33–77, 2022. doi:10.1007/s00211-021-01242-3.
- [24] P. Kritzer, F. Pillichshammer, and G. Wasilkowski. On quasi-Monte Carlo methods in weighted ANOVA spaces. *Math. Comp.*, 90:1381–1406, 2021. doi:10.1090/mcom/3598.
- [25] S. Kucherenko, B. Feil, N. Shah, and W. Mauntz. The identification of model effective dimensions using global sensitivity analysis. *Reliab. Eng. Syst. Saf.*, 96(4):440–449, 2011. doi:10.1016/j.ress.2010.11.003.
- [26] F. Y. Kuo, C. Schwab, and I. H. Sloan. Quasi-Monte Carlo Finite Element Methods for a Class of Elliptic Partial Differential Equations with Random Coefficients. *SIAM J. Numer. Anal.*, 50(6):3351–3374, 2012. doi:10.1137/110845537.
- [27] F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski, and H. Woźniakowski. On decompositions of multivariate functions. *Math. Comp.*, 79(270):953–966, 2009. doi:10.1090/s0025-5718-09-02319-9.
- [28] M. Lamboni. Kernel-based Measures of Association Between Inputs and Outputs Using ANOVA. *Sankhya A*, 86:790–826, 2024. doi:10.1007/s13171-024-00354-w.
- [29] G. Larcher, G. Leobacher, and K. Scheicheher. On the tractability of the Brownian bridge algorithm. *J. Complexity*, 19(4):511–528, 2003. doi:10.1016/s0885-064x(03)00045-1.
- [30] T. Larsen and A. Alexanderian. A new kernel-based index for the global sensitivity analysis of models with correlated inputs, 2026. Preprint at <https://arxiv.org/abs/2603.00849>.
- [31] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45:503–528, 1989. doi:10.1007/bf01589116.
- [32] R. Liu and A. B. Owen. Estimating mean dimensionality of analysis of variance decompositions. *J. Amer. Statist. Assoc.*, 101(474):712–721, 2006. doi:10.1198/016214505000001410.
- [33] N. Lüthen, S. Marelli, and B. Sudret. Sparse Polynomial Chaos Expansions: Literature Survey and Benchmark. *SIAM/ASA J. Uncertain. Quantif.*, 9(2):593–649, 2021. doi:10.1137/20m1315774.
- [34] A. Marrel, B. Iooss, F. Van Dorpe, and E. Volkova. An efficient methodology for modeling complex computer codes with Gaussian processes. *Comput. Statist. Data Anal.*, 52(10):4731–4744, 2008. doi:10.1016/j.csda.2008.03.026.
- [35] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Springer, New York, NY, USA, 1st edition, 2004. doi:10.1007/978-1-4419-8853-9.

- [36] A. B. Owen. Effective Dimension of Some Weighted Pre-Sobolev Spaces with Dominating Mixed Partial Derivatives. *SIAM J. Numer. Anal.*, 57(2):547–562, 2019.
- [37] D. Potts and M. Schmischke. Learning multivariate functions with low-dimensional structures using polynomial bases. *J. Comput. Appl. Math.*, 403:113821, 2022. doi:10.1016/j.cam.2021.113821.
- [38] D. Potts and L. Weidensager. ANOVA-boosting for random Fourier features. *Appl. Comput. Harmon. Anal.*, 79:101789, 2025. doi:10.1016/j.acha.2025.101789.
- [39] H. Rabitz, Ömer F. Aliş, J. Shorter, and K. Shim. Efficient input—output model representations. *Comput. Phys. Commun.*, 117(1):11–20, 1999. doi:10.1016/s0010-4655(98)00152-0.
- [40] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, pages 1177—1184, Red Hook, NY, USA, 2007. Curran Associates Inc.
- [41] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, Inc., New York, NY, USA, 3rd edition, 1976.
- [42] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Comput. Phys. Commun.*, 181(2):259–270, 2010. doi:10.1016/j.cpc.2009.09.018.
- [43] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. doi:10.7551/mitpress/4175.001.0001.
- [44] Y. Shin and D. Xiu. Nonadaptive Quasi-Optimal Points Selection for Least Squares Linear Regression. *SIAM J. Sci. Comput.*, 38(1):A385–A411, 2016. doi:10.1137/15m1015868.
- [45] I. H. Sloan and H. Woźniakowski. When Are Quasi-Monte Carlo Algorithms Efficient for High Dimensional Integrals? *J. Complexity*, 14(1):1–33, 1998. doi:10.1006/jcom.1997.0463.
- [46] S. Smale and D.-X. Zhou. Learning Theory Estimates via Integral Operators and Their Approximations. *Constr. Approx.*, 26(2):153–172, 2007. doi:10.1007/s00365-006-0659-y.
- [47] I. Sobol’. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comp. Simulation*, 55(1–3):271–280, 2001. doi:10.1016/s0378-4754(00)00270-6.
- [48] I. M. Sobol’. *Multidimensional quadrature formulas and Haar functions*. Nauka, Moscow (In Russian), 1969.
- [49] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, New York, NY, 1st edition, 2008. doi:10.1007/978-0-387-77242-4.
- [50] L. Weidensager. Sensitivity Analysis on the Sphere and a Spherical ANOVA Decomposition. *J. Fourier Anal. Appl.*, 32(66), 2026. doi:10.1007/s00041-026-10275-x.
- [51] C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT Press, Cambridge, MA, USA, 2006.
- [52] X. Yang, M. Choi, G. Lin, and G. E. Karniadakis. Adaptive ANOVA decomposition of stochastic incompressible and compressible flows. *J. Comput. Phys.*, 231(4):1587–1614, 2012. doi:10.1016/j.jcp.2011.10.028.
- [53] H. Zhang, Z. Ye, X. Wang, X. Guo, Z. Xu, Y. Cheng, Z. Hu, and Y. Qi. Efficient network automatic relevance determination. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 76541–76562, 2025. URL: <https://openreview.net/forum?id=RfQntVTY7Q>.