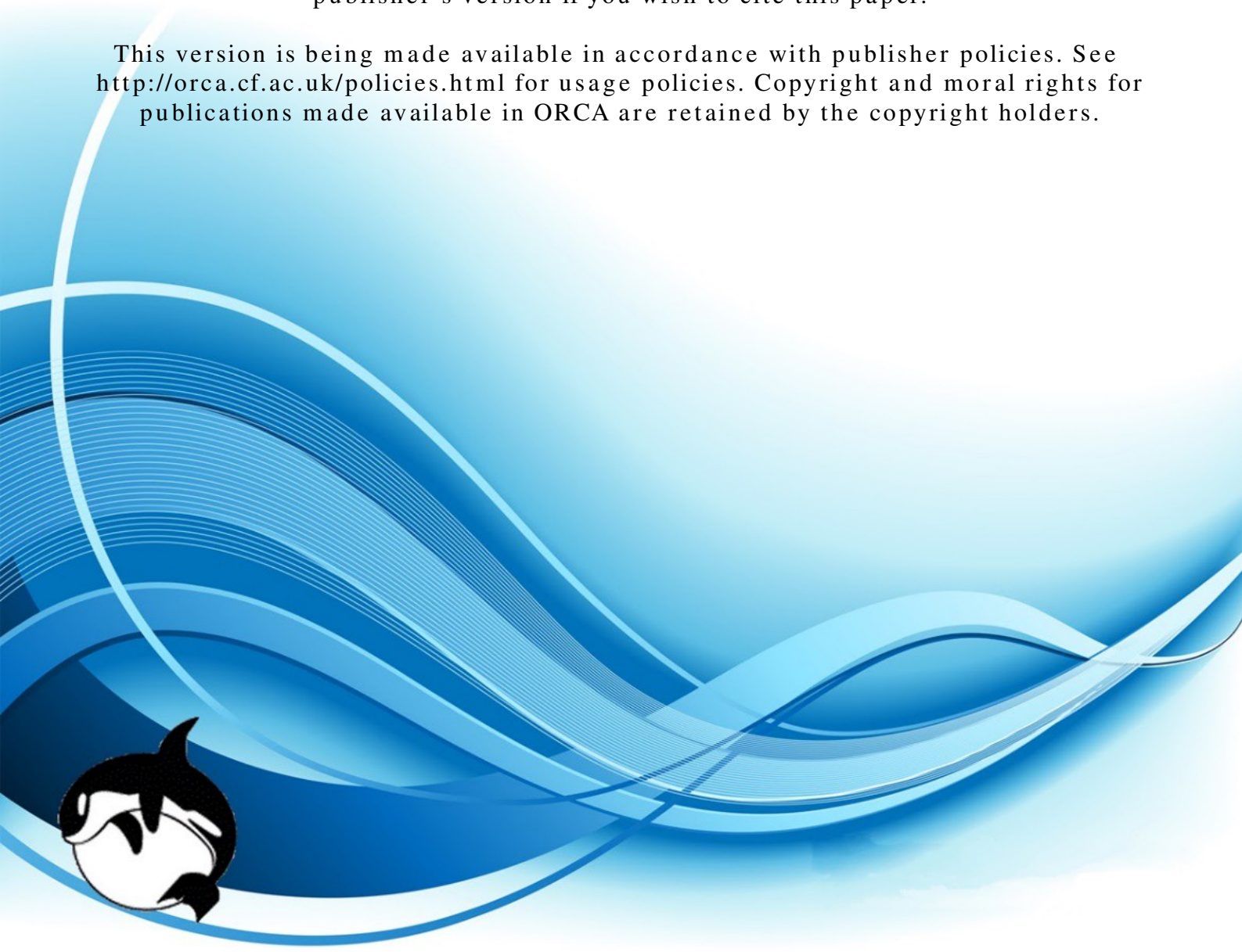

BLUE against OLSE in the location model: energy minimization and asymptotic considerations

Luc Pronzato[1] and Anatoly Zhigljavsky[2]

[1]Université Côte d'Azur-CNRS, Sophia Antipolis, France,
pronzato@i3s.unice.fr
[2]School of Mathematics, Cardiff University, UK,
ZhigljavskyAA@cardiff.ac.uk

Abstract

The main purpose of the paper is to uncover the connections between kriging, energy minimization and properties of the ordinary least squares and best linear unbiased estimators in the location model with correlated observations. We emphasize the special role of the constant function and illustrate our results by several examples.

1 Introduction

The paper makes connections between the following domains: (a) simple and ordinary kriging with kernel K for prediction of values of a random field indexed by a set $\mathscr{X}$, (b) energy minimization for K, and (c) parameter estimation with the Ordinary Least Squares Estimator (OLSE) and the Best Linear Unbiased Estimator (BLUE) in the location model with observations whose correlation is defined by K. These three areas are well-studied in modern literature. For the theory and methodology of kriging, along with numerous references, see [18]; the most advanced modern treatment of the theory of energy minimization for K is contained in the excellent book [3], whereas [10] contains a comprehensive survey on classical properties of the OLSE and the BLUE in regression models with correlated observations. For more recent results on properties of the OLSE and the BLUE, see [4] and [26]. Despite both kriging and designing for correlated observations are well-known and well-studied areas in the DOE ("design of experiments") community, the effect of substituting the OLSE for the BLUE as parameter estimator in kriging prediction models has not been adequately addressed in the DOE literature. Similarly, there is a wealth of research on energy minimization, but the implications of this well-studied area on the properties of the OLSE and the BLUE in regression models have not been sufficiently discussed in the DOE literature; see, however, [14], a previous paper by the present

authors, where some aspects of these implications have been mentioned. The present paper aims at closing these gaps by fully concentrating on the interplay between the above three areas.

The paper is organized as follows. In Section 2 we consider different versions of kriging, based on both the OLSE and BLUE. There are no new results in this section but the view-point towards different aspects of kriging is novel. Section 3 is auxiliary and summarizes some known properties of energies, potentials and minimum-energy measures [23, 21, 14], these properties being used in Section 4. The new results of Section 3 are emphasized in Examples 1 and 2, which constitute a substantial part of this section. The most interesting discussions and important contributions are contained in Section 4, where various properties of the OLSE and BLUE are considered. In Subsection 4.1 we summarize and discuss properties of these estimators for fixed-size designs while in the rest of Section 4 we concentrate on their asymptotic properties. An important result, connecting the energy minimization problem with the asymptotic value of the sequence of variances of the BLUE, is contained in Corollary 4.1: $\lim_{n\to\infty} \mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}}) > 0$ if and only if the constant function on $\mathscr{X}$ belongs to the Reproducing Kernel Hilbert Space (RKHS) generated by the kernel K, and, moreover, this limit is equal to $\mathscr{E}^*_K = \inf_\mu \int_{\mathscr{X}^2} K(\mathbf{x},\mathbf{x}')\mu(\mathrm{d}\mathbf{x})\mu(\mathrm{d}\mathbf{x}')$, where the infimum is taken over the set of all signed measures on $\mathscr{X}$ of total mass 1. Another important theoretical result of the paper is contained in subsection 4.5, where we extend some well-known results of Schoenberg [20] to the general class of so-called reduced kernels.

2 Kriging with the BLUE and OLSE

Consider the location model

$$y(\mathbf{x}) = \theta + \varepsilon(\mathbf{x}),\ \ \mathsf{E}\{\varepsilon(\mathbf{x})\} = 0,\ \ \mathsf{E}\{\varepsilon(\mathbf{x})\varepsilon(\mathbf{x}')\} = K(\mathbf{x},\mathbf{x}'),\ \ \mathbf{x},\mathbf{x}' \in \mathscr{X}, \qquad (1)$$

where $\mathscr{X}$ is some set and observations $y(\mathbf{x}_j)$ of $y(\mathbf{x})$ are performed at distinct points $\mathbf{x}_j$ of an n-point design $\mathbf{X}_n = \{\mathbf{x}_1,\ldots,\mathbf{x}_n\} \subset \mathscr{X}$. The kernel K is assumed to be strictly positive definite (SPD), which means that for any $n \in \mathbb{N}$, the kernel matrix $\mathbf{K}_n = (K(\mathbf{x}_i,\mathbf{x}_j))_{i,j=1}^n$ is SPD (as we assume that $\mathbf{X}_n$ contains no repetitions). We denote $\mathbf{k}_n(\cdot) = [K(\mathbf{x}_1,\cdot),\ldots,K(\mathbf{x}_n,\cdot)]^\top$.

The construction of best linear predictors of the value of $y(\mathbf{x}_0)$ at a given point $\mathbf{x}_0 \in \mathscr{X}$ is called "kriging". Linear predictors of $y(\mathbf{x}_0)$ have the form $\eta_n(\mathbf{x}_0) = \sum_{i=1}^n w_i y(\mathbf{x}_i) = \mathbf{w}_n^\top \mathbf{y}_n$, where $\mathbf{w}_n = (w_1,\ldots,w_n)^\top$ is the vector of weights and $\mathbf{y}_n = [y(\mathbf{x}_1),\ldots,y(\mathbf{x}_n)]^\top$ is the vector of observations. The simpler case when θ is known (and hence can be assumed to be 0), corresponds to "simple kriging"; the general case of an unknown θ is called "ordinary kriging". In this paper we shall not consider the case of "universal kriging", where the mean of the random field $y(\mathbf{x})$ is linearly parameterized, which corresponds to the situation when $y(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{h}(\mathbf{x}) + \varepsilon(\mathbf{x})$ with $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}),\ldots,h_p(\mathbf{x})]^\top$ a vector of p known functions on $\mathscr{X}$ and $\boldsymbol{\beta}$ a vector of unknown parameters in $\mathbb{R}^p$.

2.1 Simple kriging

In this model, the Mean Squared Prediction Error (MSPE) of a general linear predictor $\eta_n(\mathbf{x}_0) = \mathbf{w}_n^\top \mathbf{y}_n$ is

$$\rho_n^2(\mathbf{x}_0, \mathbf{w}_n) = \mathsf{E}\{[\eta_n(\mathbf{x}_0) - y(\mathbf{x}_0)]^2\} = K(\mathbf{x}_0, \mathbf{x}_0) - 2\,\mathbf{w}_n^\top \mathbf{k}_n(\mathbf{x}_0) + \mathbf{w}_n^\top \mathbf{K}_n \mathbf{w}_n\,.$$

The Best Linear Predictor (BLP) at $\mathbf{x}_0$ minimizes $\rho_n^2(\mathbf{x}_0, \mathbf{w}_n)$ in the class of all linear predictors $\eta_n(x_0) = \mathbf{w}_n^\top \mathbf{y}_n$ and is given by $\eta_n^*(\mathbf{x}_0) = \mathbf{w}_n^{*\top} \mathbf{y}_n$ with

$$\mathbf{w}_n^* = \mathbf{w}_n^*(\mathbf{x}_0) = \mathbf{K}_n^{-1} \mathbf{k}_n(\mathbf{x}_0)\,.$$

Therefore, the BLP is

$$\eta_n^*(\mathbf{x}_0) = \mathbf{k}_n^\top(\mathbf{x}_0) \mathbf{K}_n^{-1} \mathbf{y}_n\,. \tag{2}$$

Its MSPE equals

$$\rho_n^{*\,2}(\mathbf{x}_0) = \rho_n^2(\mathbf{x}_0, \mathbf{w}_n^*) = K(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{k}_n^\top(\mathbf{x}_0) \mathbf{K}_n^{-1} \mathbf{k}_n(\mathbf{x}_0)\,. \tag{3}$$

2.2 Ordinary kriging

BLUP. Consider the general model (1). The unbiasedness condition for the predictor $\eta_n(\mathbf{x}_0) = \mathbf{w}_n^\top \mathbf{y}_n$ is $\mathsf{E}\{\eta_n(\mathbf{x}_0)\} = \mathsf{E}\{y(\mathbf{x}_0)\} = \theta$ for all θ, which implies $\mathbf{w}_n^\top \mathbf{1}_n = 1$, where $\mathbf{1}_n = (1, \ldots, 1)^\top \in \mathbb{R}^n$. The Best Linear Unbiased Predictor (BLUP) at $\mathbf{x}_0$ minimizes $\rho_n^2(\mathbf{x}_0, \mathbf{w}_n)$ with respect $\mathbf{w}_n \in \mathbb{R}^n$ satisfying $\mathbf{w}_n^\top \mathbf{1}_n = 1$. The Lagrangian of this constrained minimization problem is

$$L(\mathbf{w}_n, \lambda_n) = \mathbf{w}_n^\top \mathbf{K}_n \mathbf{w}_n - 2\,\mathbf{w}_n^\top \mathbf{k}_n(\mathbf{x}_0) + 2\,\lambda_n(\mathbf{w}_n^\top \mathbf{1}_n - 1),$$

with λ_n the Lagrange coefficient for the constraint, which gives the stationarity conditions

$$\begin{bmatrix} \mathbf{K}_n & \mathbf{1}_n \\ \mathbf{1}_n^\top & 0 \end{bmatrix} \begin{pmatrix} \mathbf{w}_n \\ \lambda_n \end{pmatrix} = \begin{pmatrix} \mathbf{k}_n(\mathbf{x}_0) \\ 1 \end{pmatrix}. \tag{4}$$

By direct calculation, the solution of (4) is

$$\widehat{\mathbf{w}}_n = \widehat{\mathbf{w}}_n(\mathbf{x}_0) = \mathbf{K}_n^{-1} \mathbf{k}_n(\mathbf{x}_0) + \frac{[1 - \mathbf{k}_n^\top(\mathbf{x}_0) \mathbf{K}_n^{-1} \mathbf{1}_n] \mathbf{K}_n^{-1} \mathbf{1}_n}{\mathbf{1}_n^\top \mathbf{K}_n^{-1} \mathbf{1}_n}\,. \tag{5}$$

The place of the BLUE of θ in the BLUP. From (5), we get the following explicit expression for the BLUP:

$$\widehat{\eta}_n(\mathbf{x}_0) = \widehat{\theta}^n_{\mathsf{BLUE}} + \mathbf{k}_n^\top(\mathbf{x}_0) \mathbf{K}_n^{-1} (\mathbf{y}_n - \widehat{\theta}^n_{\mathsf{BLUE}} \mathbf{1}_n) \tag{6}$$

where

$$\widehat{\theta}^n_{\mathsf{BLUE}} = \frac{\mathbf{1}_n^\top \mathbf{K}_n^{-1} \mathbf{y}_n}{\mathbf{1}_n^\top \mathbf{K}_n^{-1} \mathbf{1}_n}$$

is the BLUE of θ in model (1). Indeed, the variance $\mathsf{var}(\widehat{\theta}^n)$ of any linear estimator $\widehat{\theta}^n = \mathbf{w}_n^\top \mathbf{y}_n$ of θ equals $\mathbf{w}_n^\top \mathbf{K}_n \mathbf{w}_n$, the unbiasedness constraint $\mathsf{E}\{\widehat{\theta}^n\} = \theta$ for any θ imposes $\mathbf{1}_n^\top \mathbf{w}_n = 1$, so that $\widehat{\theta}^n_{\mathsf{BLUE}} = \mathbf{w}_{n,\mathsf{BLUE}}^\top \mathbf{y}_n$ with

$$\mathbf{w}_{n,\mathsf{BLUE}} = \frac{\mathbf{K}_n^{-1}\mathbf{1}_n}{\mathbf{1}_n^\top \mathbf{K}_n^{-1}\mathbf{1}_n} = \arg \min_{\mathbf{w}_n :\, \mathbf{1}_n^\top \mathbf{w}_n = 1} \mathbf{w}_n^\top \mathbf{K}_n \mathbf{w}_n \,. \tag{7}$$

The variance of $\widehat{\theta}^n_{\mathsf{BLUE}}$ equals

$$\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}}) = \mathbf{w}_{n,\mathsf{BLUE}}^\top \mathbf{K}_n \mathbf{w}_{n,\mathsf{BLUE}} = (\mathbf{1}_n^\top \mathbf{K}_n^{-1} \mathbf{1}_n)^{-1} \,. \tag{8}$$

In Section 4.1 we shall see that $\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})$ is the energy for K of the minimum-energy signed measure μ_n^* of total mass 1, supported on $\mathbf{X}_n$, μ_n^* having weights $\mathbf{w}_{n,\mathsf{BLUE}}$.

Equations (2) and (6) show that we can perceive the BLUP $\widehat{\eta}_n$ as the BLP η_n^* applied to model (1) where the unknown θ is replaced by the BLUE $\widehat{\theta}^n_{\mathsf{BLUE}}$. Also, the BLP $\eta_n^*(\mathbf{x}_0)$ and the BLUP $\widehat{\eta}_n(\mathbf{x}_0)$ are related by

$$\widehat{\eta}_n(\mathbf{x}_0) = \eta_n^*(\mathbf{x}_0) + [1 - \mathbf{1}_n^\top \mathbf{w}_n^*(\mathbf{x}_0)]\, \widehat{\theta}^n_{\mathsf{BLUE}} \,. \tag{9}$$

Equation (9) shows that $\widehat{\eta}_n(\mathbf{x})$ corrects $\eta_n^*(\mathbf{x})$ when $\mathbf{1}_n^\top \mathbf{w}_n^*(\mathbf{x}) \neq 1$. The MSPE of $\widehat{\eta}_n(\mathbf{x}_0)$, called the ordinary kriging variance, equals

$$\begin{aligned} \widehat{\rho}_n^2(\mathbf{x}_0) &= {\rho_n^*}^2(\mathbf{x}_0) + [1 - \mathbf{1}_n^\top \mathbf{w}_n^*(\mathbf{x}_0)]^2\, \mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}}) \\ &= K(\mathbf{x}_0,\mathbf{x}_0) - \mathbf{k}_n^\top(\mathbf{x}_0)\mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x}_0) + \frac{(1 - \mathbf{k}_n^\top(\mathbf{x}_0)\mathbf{K}_n^{-1}\mathbf{1}_n)^2}{\mathbf{1}_n^\top \mathbf{K}_n^{-1}\mathbf{1}_n} \,. \end{aligned} \tag{10}$$

Therefore $\widehat{\rho}_n^2(\mathbf{x}_0) \geq {\rho_n^*}^2(\mathbf{x}_0)$ given by (3); this is due to the presence of the unknown constant mean θ that needs to be estimated.

Use of a reduced kernel. Let us show that expressions of $\widehat{\eta}_n(\mathbf{x}_0)$ and $\widehat{\rho}_n^2(\mathbf{x}_0)$ can be simplified by considering a formal modification of the model. Define the random field $\varepsilon_0(\mathbf{x}) = \varepsilon(\mathbf{x}) - \varepsilon(\mathbf{x}_0)$; it is centered and has covariance function

$$R(\mathbf{x},\mathbf{x}') = \mathsf{E}\{\varepsilon_0(\mathbf{x})\varepsilon_0(\mathbf{x}')\} = K(\mathbf{x},\mathbf{x}') - K(\mathbf{x},\mathbf{x}_0) - K(\mathbf{x}',\mathbf{x}_0) + K(\mathbf{x}_0,\mathbf{x}_0) \,,$$

sometimes called the reduction of K with respect to the delta measure $\delta_{\mathbf{x}_0}$ [19, 7], or the version of K centered at $\delta_{\mathbf{x}_0}$ [21]. It satisfies $R(\mathbf{x},\mathbf{x}_0) = 0$ for all $\mathbf{x}$, and the kernel matrix $\mathbf{R}_n$ with elements $\{\mathbf{R}_n\}_{i,j} = R(\mathbf{x}_i,\mathbf{x}_j)$ satisfies

$$\mathbf{R}_n = [\mathbf{I}_n \;\; -\mathbf{1}_n] \begin{bmatrix} \mathbf{K}_n & \mathbf{k}_n(\mathbf{x}_0) \\ \mathbf{k}_n^\top(\mathbf{x}_0) & K(\mathbf{x}_0,\mathbf{x}_0) \end{bmatrix} \begin{bmatrix} \mathbf{I}_n \\ -\mathbf{1}_n^\top \end{bmatrix} ,$$

with $\mathbf{I}_n$ the $n \times n$ identity matrix. The kernel matrix $\mathbf{R}_n$ is strictly positive definite when the $\mathbf{x}_i$ are all pairwise different and $\mathbf{x}_0 \notin \mathbf{X}_n$. Since we can write $y(\mathbf{x}) = \theta + \varepsilon(\mathbf{x}) = \vartheta + \varepsilon_0(\mathbf{x})$, with $\vartheta = \theta + \varepsilon(\mathbf{x}_0)$ defining a new unknown

mean parameter, previous expressions for the predictor $\widehat{\eta}_n(\mathbf{x}_0)$ and the kriging variance $\widehat{\rho}_n^2(\mathbf{x}_0)$ remain valid when we substitute R for K. Since $R(\mathbf{x}, \mathbf{x}_0) = 0$ for all $\mathbf{x}$, we obtain

$$\widehat{\eta}_n(\mathbf{x}_0) = \frac{\mathbf{1}_n^\top \mathbf{R}_n^{-1} \mathbf{y}_n}{\mathbf{1}_n^\top \mathbf{R}_n^{-1} \mathbf{1}_n} \quad \text{and} \quad \widehat{\rho}_n^2(\mathbf{x}_0) = \frac{1}{\mathbf{1}_n^\top \mathbf{R}_n^{-1} \mathbf{1}_n} \,.$$

Notice that $\widehat{\eta}_n(\mathbf{x}_0)$ is the BLUE of ϑ and $\widehat{\rho}_n^2(\mathbf{x}_0)$ is the variance of this estimator. Reduction of K with respect to a general measure μ is considered in Section 4.5.

Prediction with the OLSE of θ. The OLSE $\widehat{\theta}_{\mathsf{OLSE}}^n$ of θ in the model (1) is simply the empirical mean $\widehat{\theta}_{\mathsf{OLSE}}^n = \bar{\mathbf{y}}_n = \mathbf{w}_{n,\mathsf{OLSE}}^\top \mathbf{y}_n$, where $\mathbf{w}_{n,\mathsf{OLSE}} = \mathbf{1}_n/n$. As $\mathbf{1}_n^\top \mathbf{w}_{n,\mathsf{OLSE}} = 1$, $\widehat{\theta}_{\mathsf{OLSE}}^n$ is unbiased; its variance equals

$$\mathsf{var}(\widehat{\theta}_{\mathsf{OLSE}}^n) = \mathbf{w}_{n,\mathsf{OLSE}}^\top \mathbf{K}_n \mathbf{w}_{n,\mathsf{OLSE}} = \frac{1}{n^2} \mathbf{1}_n^\top \mathbf{K}_n \mathbf{1}_n \,. \tag{11}$$

From the Cauchy-Schwarz inequality,

$$\frac{\mathsf{var}(\widehat{\theta}_{\mathsf{OLSE}}^n)}{\mathsf{var}(\widehat{\theta}_{\mathsf{BLUE}}^n)} = \frac{1}{n^2} (\mathbf{1}_n^\top \mathbf{K}_n \mathbf{1}_n)(\mathbf{1}_n^\top \mathbf{K}_n^{-1} \mathbf{1}_n) \geq \frac{(\mathbf{1}_n^\top \mathbf{1}_n)^2}{n^2} = 1 \,. \tag{12}$$

To predict with the OLSE $\widehat{\theta}_{\mathsf{OLSE}}^n$, we apply the simple kriging predictor (2) to the centered observations $\mathbf{y}'_n = \mathbf{y}_n - \bar{\mathbf{y}}_n$ and then add the empirical mean $\bar{\mathbf{y}}_n$ to the predictor η_n^*. This yields a predictor having the same form as (6), but with the OLSE $\bar{\mathbf{y}}_n = \widehat{\theta}_{\mathsf{OLSE}}^n$ substituted for the BLUE $\widehat{\theta}_{\mathsf{BLUE}}^n$:

$$\widehat{\eta}_{n,\mathsf{OLSE}}(\mathbf{x}_0) = \bar{\mathbf{y}}_n + \mathbf{k}_n^\top(\mathbf{x}_0) \mathbf{K}_n^{-1} (\mathbf{y}_n - \bar{\mathbf{y}}_n \mathbf{1}_n) \,.$$

The MSPE of $\widehat{\eta}_{n,\mathsf{OLSE}}(\mathbf{x}_0)$ is

$$\begin{aligned} \widehat{\rho}_{n,\mathsf{OLSE}}^2(\mathbf{x}_0) &= {\rho_n^*}^2(\mathbf{x}_0) + [1 - \mathbf{1}_n^\top \mathbf{w}_n^*(\mathbf{x}_0)]^2 \, \mathsf{var}(\widehat{\theta}_{\mathsf{OLSE}}^n) \\ &= K(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{k}_n^\top(\mathbf{x}_0) \mathbf{K}_n^{-1} \mathbf{k}_n(\mathbf{x}_0) + (1 - \mathbf{k}_n^\top(\mathbf{x}_0) \mathbf{K}_n^{-1} \mathbf{1}_n)^2 \, \frac{\mathbf{1}_n^\top \mathbf{K}_n \mathbf{1}_n}{n^2} \,, \end{aligned}$$

and (12) implies $\widehat{\rho}_{n,\mathsf{OLSE}}^2(\mathbf{x}_0) \geq \widehat{\rho}_n^2(\mathbf{x})$, where $\widehat{\rho}_n^2(\mathbf{x}) = \widehat{\rho}_{n,\mathsf{BLUE}}^2(\mathbf{x}_0)$ is given by (10). Although $\mathsf{var}(\widehat{\theta}_{\mathsf{OLSE}}^n)$ may be significantly larger than $\mathsf{var}(\widehat{\theta}_{\mathsf{BLUE}}^n)$, in practice $\widehat{\rho}_{n,\mathsf{OLSE}}^2(\mathbf{x}_0)$ is often only marginally larger than $\widehat{\rho}_n^2(\mathbf{x})$ due to the factor $(1 - \mathbf{k}_n^\top(\mathbf{x}) \mathbf{K}_n^{-1} \mathbf{1}_n)^2$, which tends to be small especially for large n. The right panel of Figure 4 provides an illustration.

The variance $\mathsf{var}(\widehat{\theta}_{\mathsf{OLSE}}^n)$ corresponds in fact to the energy for K of the empirical measure $\mu_n = (1/n) \sum_{i=1}^n \delta_{\mathbf{x}_i}$ on $\mathbf{X}_n$. The notion of energy of a measure for a kernel K is introduced in the next section.

3 Energy, potentials and minimum-energy measures

3.1 Energy and MMD

Let $\mathscr{X}$ be a compact set in a metric measurable space and $K(\mathbf{x}, \mathbf{x}')$ be a kernel; that is, a symmetric function on $\mathscr{X} \times \mathscr{X} \to \mathbb{R}$. Unless otherwise stated, we assume that K is uniformly bounded; that is, there exists a constant C such that $|K(\mathbf{x}, \mathbf{x}')| \leq C < \infty$ for all $\mathbf{x}$ and $\mathbf{x}'$ in $\mathscr{X}$. Let $\mathscr{M}$ and $\mathscr{M}(1)$ be respectively the sets of finite signed measures on $\mathscr{X}$ and of signed measures on $\mathscr{X}$ with total mass 1. Also, let $\mathscr{M}^+$ and $\mathscr{M}^+(1)$ be the sets of finite positive measures on $\mathscr{X}$ and of probability measures on $\mathscr{X}$, respectively. Note that the set $\mathscr{M}^+(1)$ is weakly compact whereas $\mathscr{M}(1)$ is not.

For any $\mu \in \mathscr{M}$, the energy of μ is defined as

$$\mathscr{E}_K(\mu) = \int_{\mathscr{X}^2} K(\mathbf{x}, \mathbf{x}') \mu(\mathrm{d}\mathbf{x}) \mu(\mathrm{d}\mathbf{x}') \,,$$

and $|\mathscr{E}_K(\mu)| < \infty$ as K is uniformly bounded.

Definition 3.1. *A kernel K is* Integrally Positive Definite *(IPD) when $\mathscr{E}_K(\nu) \geq 0$ for any $\nu \in \mathscr{M}$. It is* Integrally Strictly Positive Definite *(ISPD) if, in addition, $\mathscr{E}_K(\nu) = 0$ implies $\nu = 0$.*

Definition 3.2. *A kernel K is* Conditionally Integrally Positive Definite *(CIPD) when $\mathscr{E}_K(\nu) \geq 0$ for all signed measures $\nu \in \mathscr{M}$ such that $\nu(\mathscr{X}) = 0$. It is* Conditionally Strictly Integrally Positive Definite *(CISPD) if, in addition, $\mathscr{E}_K(\nu) = 0$ for a ν such that $\nu(\mathscr{X}) = 0$ implies $\nu = 0$.* ◁

If the kernel K is IPD, then $\mathscr{E}_K(\mu) \geq 0$ for any $\mu \in \mathscr{M}$. For $\mu, \nu \in \mathscr{M}$, we also define the cross-energy by $\mathscr{E}_K(\mu, \nu) = \int_{\mathscr{X}^2} K(\mathbf{x}, \mathbf{x}') \mu(\mathrm{d}\mathbf{x}) \nu(\mathrm{d}\mathbf{x}')$. As the energies for both measures μ and ν are bounded, the Cauchy-Schwarz inequality implies that the cross-energy $\mathscr{E}_K(\mu, \nu)$ is also bounded. If K is IPD or more generally CIPD, then $\mathscr{E}_K(\cdot)$ is a convex functional on $\mathscr{M}(1)$; if K is CISPD, then $\mathscr{E}_K(\cdot)$ is strictly convex on $\mathscr{M}(1)$; see [14]. These properties remain valid without the assumption that K is uniformly bounded (in particular for singular kernels such that $K(\mathbf{x}, \mathbf{x}) = +\infty$), but we must then restrict our attention to measures with finite energy. In what follows, we assume that K is at least CIPD.

The Maximum Mean Discrepancy (MMD) between two measures μ and ν in $\mathscr{M}(1)$ is

$$\mathsf{MMD}(\mu, \nu) = \sqrt{\mathscr{E}_K(\mu - \nu)} \,. \tag{13}$$

As we have assumed that K is at least CIPD, $\mathsf{MMD}(\mu, \nu)$ is properly defined for all $\mu, \nu \in \mathscr{M}(1)$. If K is CISPD, then K is a characteristic kernel, and the MMD defines a metric on the space of probability measures $\mathscr{M}^+(1)$; see [23].

Another important notion is that of potential (kernel imbedding) of a (signed) measure $\mu \in \mathscr{M}$, defined by $P_\mu(\mathbf{x}) = \int_{\mathscr{X}} K(\mathbf{x}, \mathbf{x}')\mu(\mathrm{d}\mathbf{x}')$. The space of potentials is $\mathcal{P}_K = \{P_\mu(\cdot),\ \mu \in \mathscr{M}\}$; this set is dense in the RKHS $\mathcal{H}(K)$ associated with K [12, Prop. 2.1]. We also define $\mathcal{P}_K^+ = \{P_\mu(\cdot),\ \mu \in \mathscr{M}^+\}$, the set of potentials associated with finite positive measures $\mu \in \mathscr{M}^+$. Smoothness of the kernel K is naturally inherited by the potentials $P_\mu(\cdot) \in \mathcal{P}_K$ and subsequently by all functions of the RKHS $\mathcal{H}(K)$; see [17, Sect. 2.1.3] for a comprehensive exposition of the relation between the smoothness of K and that of the elements of $\mathcal{H}(K)$.

3.2 Minimum-energy measures: optimality conditions

We denote by $\mu^* = \arg\min_{\mu \in \mathscr{M}(1)} \mathscr{E}_K(\mu)$ the minimum-energy signed measure of total mass 1 (when it exists), and by $\mu^+ = \arg\min_{\mu \in \mathscr{M}^+(1)} \mathscr{E}_K(\mu)$ the minimum-energy probability measure, for a given kernel K. We also denote $\mathscr{E}_K^* = \inf_{\mu \in \mathscr{M}(1)} \mathscr{E}_K(\mu)$ and $\mathscr{E}_K^+ = \min_{\mu \in \mathscr{M}^+(1)} \mathscr{E}_K(\mu) = \mathscr{E}_K(\mu^+)$.

Theorem 3.1. *Assume that K is CIPD.*

(i) μ^+ is a minimum-energy probability measure if and only if $P_{\mu^+}(\mathbf{x}) \geq \mathscr{E}_K^+$ for all $\mathbf{x} \in \mathscr{X}$. For any such measure μ^+, $P_{\mu^+}(\mathbf{x}) = \mathscr{E}_K^+$ on the support of μ^+.

(ii) μ^* is a minimum-energy signed measure of total mass one if and only if $P_{\mu^*}(\mathbf{x}) = \mathscr{E}_K^*$ for all $\mathbf{x} \in \mathscr{X}$.

(iii) There always exists a minimum-energy probability measure μ^+, there may not exist a minimum-energy signed measure μ^* of total mass one. If K is CISPD, then μ^+ is uniquely defined and μ^* is uniquely defined when it exists.

For the proof of Theorem 3.1 see [14, Theorems 3.1 and 3.2].

A noticeable consequence of Theorem 3.1 is that minimum-energy measures for separable kernels can be deduced from their one-dimensional counterparts. Indeed, consider a separable (tensor product) kernel $K(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^d K_i(x_i, x_i')$ on $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_d$, where the K_i are univariate (C)ISPD kernels. Then, the energy and potential of a product measure $\mu(\mathrm{d}\mathbf{x}) = \prod_{i=1}^d \mu_i(\mathrm{d}x_i)$ satisfy $\mathscr{E}_K(\mu) = \prod_{i=1}^d \mathscr{E}_{K_i}(\mu_i)$ and $P_\mu(\mathbf{x}) = \prod_{i=1}^d P_{\mu_i}(x_i)$. Theorem 3.1 thus implies that the minimum-energy probability measure μ_K^+ for K is the product of the univariate minimum-energy probability measures $\mu_{K_i}^+$, i.e., $\mu_K^+(\mathrm{d}\mathbf{x}) = \prod_{i=1}^d \mu_{K_i}^+(\mathrm{d}x_i)$, and if there exists a minimum-energy signed measure $\mu_{K_i}^*$ for K_i on $\mathscr{X}_i$ for each i, the minimum-energy signed measure μ_K^* for K on $\mathscr{X}$ exists and equals $\mu_K^*(\mathrm{d}\mathbf{x}) = \prod_{i=1}^d \mu_{K_i}^*(\mathrm{d}x_i)$. Moreover, if, for each i, there exists a minimum-energy signed measure $\mu_{K_i}^*$ for K_i on $\mathscr{X}_i$ which coincides with $\mu_{K_i}^+$, the minimum-energy probability measure for K_i on $\mathscr{X}_i$, then μ_K^* for K on $\mathscr{X}$ exists and coincides with μ_K^+, the minimum-energy probability measure for K on $\mathscr{X}$.

Let $1_{\mathscr{X}}$ denote constant function 1 on $\mathscr{X}$ (i.e., $1_{\mathscr{X}}(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathscr{X}$). Theorem 3.1 has the following corollary.

Corollary 3.1.

(i) μ^* *exists if and only if* $1_{\mathscr{X}} \in \mathcal{P}_K$,

(ii) $\mu^* = \mu^+$ *if and only if* $1_{\mathscr{X}} \in \mathcal{P}_K^+$.

Corollary 3.1 raises the importance of the establishing conditions on K such that (a) $1_{\mathscr{X}} \in \mathcal{P}_K$ and (b) $1_{\mathscr{X}} \in \mathcal{P}_K^+$. The second condition will be considered in the next subsection and we will return to condition (a) and its consequences in Section 4.4; see Corollary 4.1.

3.3 Conditions guaranteeing that μ^* exists and $\mu^* = \mu^+$

In view of Corollary 3.1-(ii), $1_{\mathscr{X}} \in \mathcal{P}_K^+$ is equivalent to the existence of a minimum-energy signed measure μ^* of mass 1, with μ^* being a probability measure. An easy condition guaranteeing this is given in the following property, which is a simple consequence of Theorem 3.1.

Corollary 3.2. *Let μ^+ be a minimum-energy probability measure. If μ^+ is supported on the whole set $\mathscr{X}$, then μ^+ is also a minimum-energy signed measure of total mass one.*

The proof of the next theorem, provided in [14], is based on the observation that under the conditions of the theorem, the potential P_{μ^*} is subharmonic outside the support of μ^*. This theorem provides the most general sufficient condition known to us.

Theorem 3.2 (see [14, Th. 3.3])**.** *Let K be an ISPD translation invariant kernel, with $K(\mathbf{x}, \mathbf{x}') = \Psi(\mathbf{x} - \mathbf{x}')$ and Ψ continuous, twice differentiable except at the origin, with Laplacian $\Delta\Psi(\mathbf{x}) = \sum_{i=1}^{d} \partial^2\Psi(\mathbf{x})/\partial x_i^2 \geq 0$, $\forall \mathbf{x} \neq 0$. Then there exists a unique minimum-energy signed measure μ^* of mass 1 and μ^* is a probability measure.*

Despite its generality, the conditions stated in Theorem 3.2 are rather disappointing. First, when $d = 1$ they correspond to $\Psi(x)$ being convex for $x > 0$, a condition already mentioned in [8]. Second, when $d \geq 2$ and $\Psi(\mathbf{x} - \mathbf{x}') = \psi(\|\mathbf{x} - \mathbf{x}'\|)$, ψ must have a singularity at 0 to have $\Delta\Psi(\mathbf{x}) \geq 0$ for all $\mathbf{x} \neq 0$ (this is the case of Riesz kernels with $\psi(t) = t^{-s}$ when $s \in [d-2, d)$). Finally, these conditions are not necessary. As the example below illustrates, it is indeed easy to construct translation invariant kernels for $d = 1$ such that $\Psi(x)$ is not convex for $x > 0$ but the minimum-energy signed measure μ^* of mass 1 is a probability measure.

Example 1. Consider the following linear combination of exponential and Gaussian kernels: $\Psi(x) = \exp(-3\,|x|/2) + \exp(-x^2)$ with $\mathscr{X} = [0, 1]$. Then, $\Delta\Psi(x) < 0$ for $x \in (a, b)$ with $a \simeq 0.0978$, $b \simeq 0.3534$, but μ^* is a probability

measure: $\mu^* = \alpha\,\delta_0 + \alpha\,\delta_1 + (1-2\,\alpha)\xi$, with $\alpha \simeq 0.4331$ and ξ having a density φ with respect to the Lebesgue measure on $\mathscr{X}$; see the left panel of Figure 1 for a plot of $\varphi(x)$ (determined numerically).

By modifying the relative weights of the exponential and Gaussian kernels, we can easily create situations where μ^* exists but is not a probability measure. The right panel of Figure 1 corresponds to $\Psi(x) = \exp(-3\,|x|/2) + (3/2)\,\exp(-x^2)$; there $\mu^* = \alpha\,\delta_0 + \alpha\,\delta_1 + (1-2\,\alpha)\xi$ with $\alpha \simeq 0.4893$ and $\varphi(x) < 0$ for $x \in (c, d)$, $c \simeq 0.2659$, $d \simeq 0.7341$. Corollary 3.2 implies that the support of the minimum-energy probability measure μ^+ is strictly included in $\mathscr{X}$ (and indeed numerical computation shows that the density of the continuous component of μ^+ is zero in the central part of the interval $[0, 1]$).

For the exponential kernel alone, $\Psi(x) = \exp(-\beta|x|)$, $\beta \in (0, \infty)$, the continuous component is the Lebesgue measure $\mu_L = \mathscr{U}_{[0,1]}$ on $\mathscr{X}$, and $\mu^* = \mu^+ = (\delta_0 + \delta_1 + \beta\,\mu_L)/(\beta + 2)$; see [14]. The continuous BLUE is thus

$$\widehat{\theta}^{\infty}_{\mathsf{BLUE}} = \frac{1}{\beta + 2}\left[y(0) + y(1) + \beta \int_0^1 y(x)\mathrm{d}x\right] .$$

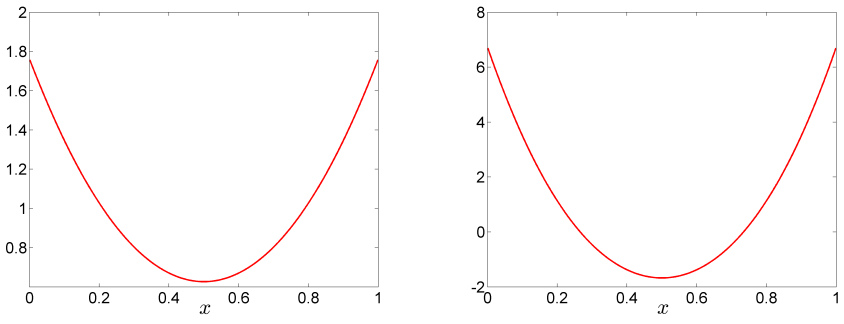

Figure 1: Density $\varphi(x)$ of the continuous component of the minimum-energy signed measure μ^* on $\mathscr{X} = [0, 1]$. Left: $K(x, x') = \exp(-3\,|x - x'|/2) + \exp(-|x - x'|^2)$ (μ^* is a probability measure). Right: $K(x, x') = \exp(-3\,|x - x'|/2) + (3/2)\,\exp(-|x - x'|^2)$ (μ^* is not a probability measure).

The weights $\mathbf{w}_{n,\mathsf{BLUE}}$ defined by (7) are very different for the kernel $\Psi(x) = \exp(-3\,|x|/2) + \exp(-x^2)$ and for the Gaussian kernel $\Psi(x) = \exp(-x^2)$; Figure 2 gives an illustration (log-scale) when $\mathbf{X}_n$ corresponds to $n = 50$ points equally spaced in $[0, 1]$. Whereas the measure that allocates $\mathbf{w}_{n,\mathsf{BLUE}}$ to $\mathbf{X}_n$ forms a discrete approximation of μ^* when $\Psi(x) = \exp(-3\,|x|/2) + \exp(-x^2)$, there is no minimum-energy signed measure of total mass one when $\Psi(x) = \exp(-x^2)$ and the (signed) components w_i of $\mathbf{w}_{n,\mathsf{BLUE}}$ have very large magnitudes and strongly oscillate. ◁

In Example 1, when $K(x, x') = \Psi(x - x') = \exp(-\beta|x - x'|)$ on $\mathscr{X} = [0, 1]$, the *continuous component of* μ^* is uniform on $\mathscr{X}$, but we cannot have $\mu^* =$

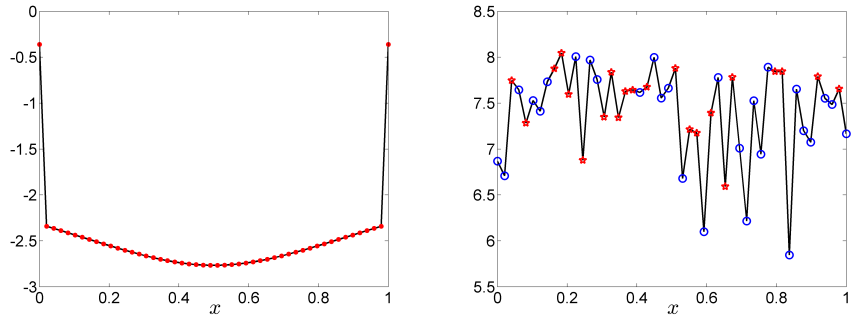

Figure 2: Components w_i of the BLUE weights (7) for $\mathbf{X}_n = \{(i-1)/(n-1),\, i = 1,\dots,n\}$ with $n = 50$. Left: $\log_{10} w_i$ when $\Psi(x) = \exp(-3\,|x|/2) + \exp(-x^2)$; Right: $\log_{10} |w_i|$ when $\Psi(x) = \exp(-x^2)$ ($\star$ for $w_i \geq 0$, $\circ$ for $w_i < 0$).

$\mathscr{U}_{[0,1]}$ when $\Psi(x)$ is bounded, nonnegative, non-constant, and non-increasing for $x > 0$: indeed, simple calculation shows that $P_{\mu_L}(0) < P_{\mu_L}(1/2)$ so that P_{μ_L} is not constant on $[0,1]$; see Theorem 3.1-(ii).

We formulate the following two conjectures for the case where $\mathscr{X}$ is a compact subset of $\mathbb{R}^d$ with nonempty interior and K is translation invariant, i.e., $K(\mathbf{x},\mathbf{x}') = \Psi(\mathbf{x}-\mathbf{x}')$:

C1: For $d > 1$, μ^* does not exist unless Ψ has a singularity at zero (see [15] for properties and regularization of singular kernels).

C2: If Ψ is differentiable at the origin, then μ^* does not exist.

According to Corollary 3.1, to prove C2 it would be enough to show that $1_{\mathscr{X}} \notin \mathcal{P}_K$ when Ψ is differentiable at the origin. Below we provide a simple example that supports this conjecture.

Example 2. Consider the Matérn 3/2 kernel on $\mathscr{X} = [0,1]$ with $\Psi(x) = (1+\beta|x|)\exp(-\beta|x|)$. It defines an RKHS on $\mathbb{R}$, which we denote by $\mathcal{H}_{\mathbb{R}}$; we denote by $\|f\|_{\mathcal{H}_{\mathbb{R}}}$ the norm of a function $f \in \mathcal{H}_{\mathbb{R}}$. The restriction of elements of $\mathcal{H}_{\mathbb{R}}$ to $\mathscr{X}$ defines another RKHS $\mathcal{H}_{\mathscr{X}}$, for which the norm is defined by

$$\forall f \in \mathcal{H}_{\mathscr{X}},\ \|f\|_{\mathcal{H}_{\mathscr{X}}} = \min_{f_0 \in \mathcal{H}_{\mathbb{R}} : f_0(x) = f(x),\, \forall x \in \mathscr{X}} \|f_0\|_{\mathcal{H}_{\mathbb{R}}}\,,$$

see [1, Th. 6]. The inclusion $1_{\mathscr{X}} \in \mathcal{P}_K$ means that there exists a signed measure $\mu \in \mathscr{M}$ such that $1 = 1_{\mathscr{X}}(x) = P_\mu(x)$ for all $x \in \mathscr{X}$. Consider the function $\mathbb{1}_{\mathscr{X}}$ defined on $\mathbb{R}$ by

$$\mathbb{1}_{\mathscr{X}}(x) = \begin{cases} (1-\beta x)\exp(\beta x) & \text{for } x \in (-\infty, 0]\,, \\ 1 & \text{for } x \in [0,1]\,, \\ (1+\beta[x-1])\exp[-\beta(x-1)] & \text{for } x \in [1,\infty)\,. \end{cases}$$

It belongs to $\mathcal{H}_{\mathbb{R}}$ and coincides with $1_{\mathscr{X}}$ on $\mathscr{X}$ (so that $1_{\mathscr{X}} \in \mathcal{H}_{\mathscr{X}}$). We then investigate the possibility that $\mathbb{1}_{\mathscr{X}}(x) = P_\mu(x) = \int \Psi(x-x')\mu(\mathrm{d}x')$ for μ a

measure supported on $\mathscr{X}$. In the Fourier space, this gives $\widehat{\mathbb{1}_{\mathscr{X}}}(\omega) = \widehat{\Psi}(\omega)\widehat{M}(\omega)$, with $\widehat{\Psi}(\omega) = 2\sqrt{2}\beta^3/[\sqrt{\pi}(\beta^2+\omega^2)^2]$ the Fourier transform of Ψ and $\widehat{M}$ the Fourier transform of μ. This yields after some calculation

$$\widehat{M}(\omega) = \frac{i\,(\beta^2-\omega^2)\,[\exp(-i\omega)-1] + 2\,\beta\omega\,[\exp(-i\omega)+1]}{4\,\beta\omega}\,,$$

and then

$$\mu = \frac{\sqrt{2\,\pi}}{4\,\beta}\,\left[\beta^2\mu_L + 2\,\beta(\delta_0+\delta_1) + \delta_0^{(1)} - \delta_1^{(1)}\right]\,,$$

where $\int_{\mathbb{R}} f(x)\delta_x^{(1)}\mathrm{d}x = f'(x)$ for any test function f, with f' the derivative of f. Therefore, μ is not a measure, and the situation would be similar for any other extension $\mathbb{1}_{\mathscr{X}}$ of $1_{\mathscr{X}}$. The (generalized) continuous BLUE of Section 4.4 is

$$\widehat{\theta}^{\infty}_{\mathsf{BLUE}} = \frac{2\,\beta\,[y(0)+y(1)] + \beta^2\int_0^1 y(x)\mathrm{d}x + y'(0) - y'(1)}{\beta^2+2\,\beta}\,,$$

which involves the first-order derivatives of y. We shall see in Section 4.4 (Theorem 4.2) that $\mathsf{var}(\widehat{\theta}^{\infty}_{\mathsf{BLUE}}) = 1/\|1_{\mathscr{X}}\|^2_{\mathcal{H}_{\mathscr{X}}}$, with here $\|1_{\mathscr{X}}\|^2_{\mathcal{H}_{\mathscr{X}}} = \|\mathbb{1}_{\mathscr{X}}\|^2_{\mathcal{H}_{\mathbb{R}}} = 1+\beta/4$; see Figure 3 for a plot of $1/\|1_{\mathscr{X}}\|^2_{\mathcal{H}_{\mathscr{X}}}$ as a function of β.

From Corollary 3.2, the non-existence of a minimum-energy signed measure of total mass 1 implies that the support of the minimum-energy probability measure μ^+ is strictly included in $\mathscr{X}$. After some calculation, we obtain that for β large enough μ^+ is the mixture of delta measures with the uniform distribution on an interval $[a_*, 1-a_*]$: we have $\mu^+ = \mu^+(a_*)$, with $\mu^+(a) = m_0\,(\delta_0+\delta_1) + m_a(\delta_a+\delta_{1-a}) + \alpha(1-2\,a)\mathscr{U}_{[a,1-a]}$ $(a > 1/\beta)$, where $m_0 = \exp(\beta a)/D$, $m_a = (\beta a-1)/D$, $\alpha = \beta^2 a/D$, and $D = D(\beta,a) = 2\,\exp(\beta a)+2\,(\beta a-1)+\beta^2 a(1-2\,a)$; the potential $P_{\mu^+(a)}(\cdot)$ is constant on $[a,1-a]$ and a_* is obtained by minimizing

$$\mathscr{E}_{K(\beta)}[\mu^+(a)] = \frac{2[-2\,\beta a + 6\,\beta^2 a^2 + \exp(2\,\beta a) - 1 - 4\,\beta^3 a^3 + 2\,\beta^3 a^2]}{D^2(\beta,a)}$$

with respect to a or, equivalently, by solving the equation $P_{\mu^+(a)}(a) = P_{\mu^+(a)}(0)$. The solution is $a_* = C/\beta$, with $C \simeq 1.572366$, which gives

$$\mathscr{E}_{K(\beta)}(\mu^+) = \mathscr{E}_{K(\beta)}[\mu^+(a_*)] = \frac{2\left[6C^2 + \exp(2\,C) + 2\beta C^2 - 1 - 2C - 4C^3\right]}{[2\,\exp(C) + 2\,(C-1) + C(\beta - 2\,C)]^2}. \quad (14)$$

When $\beta \to +\infty$, a_* tends to 0, m_0 and m_a tend to 0 and α tends to 1; that is, μ^+ converges in distribution to $\mathscr{U}_{[0,1]}$. This construction only makes sense when $a_* < 1/2$, which corresponds to $\beta > \beta_1 \simeq 3.144731$. For smaller values of β, μ^+ has no continuous component. When $\beta_1 \geq \beta > \beta_0 \simeq 2.512862$, the minimum-energy probability measure is $\mu^+(a) = m_0\,(\delta_0+\delta_1) + (1-2\,m_0)\delta_{1/2}$, with

$$m_0 = m_0(\beta) = \frac{(\beta+2)\exp(-\beta/2) - 2}{2\,[2\,(\beta+2)\exp(-\beta/2) - (\beta+1)\exp(-\beta) - 3]} < 1/2$$

which gives

$$\mathscr{E}_{K(\beta)}(\mu^+) = \frac{(\beta^2 + 2\,\beta + 2)\exp(-\beta) - 2}{2\,[2\,(\beta + 2)\exp(-\beta/2) - (\beta + 1)\exp(-\beta) - 3]}\,. \tag{15}$$

When $\beta \le \beta_0$, $\mu^+ = (\delta_0 + \delta_1)/2$, and $\mathscr{E}_{K(\beta)}(\mu^+) = [1 + (\beta + 1)\exp(-\beta)]/2$. Figure 3 shows $\mathscr{E}_{K(\beta)}(\mu^+)$ as a function of β for $\beta \in [1, 10]$.

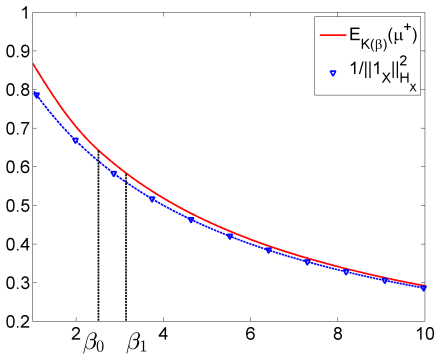

Figure 3: Energy of the minimum-energy probability measure μ^+ for $\Psi(x) = (1 + \beta|x|)\exp(-\beta|x|)$, $x \in [0, 1]$: $\mathscr{E}_{K(\beta)}(\mu^+) = [1 + (\beta + 1)\exp(-\beta)]/2$ for $\beta \le \beta_0$, is given by (15) for $\beta_0 < \beta \le \beta_1$ and by (14) for $\beta_1 < \beta$. The figure also shows $\mathsf{var}(\widehat{\theta}^{\infty}_{\mathsf{BLUE}}) = 1/\|1_{\mathscr{X}}\|^2_{\mathcal{H}_{\mathscr{X}}}$ as a function of β.

In contrast with the Matérn 3/2 kernel considered above, when using the same approach for the exponential (Matérn 1/2) kernel for which $\Psi(x) = \exp(-\beta|x|)$ and $\widehat{\Psi}(\omega) = \sqrt{2}\beta/[\sqrt{\pi}(\beta^2 + \omega^2)]$, by taking $\mathbb{1}_{\mathscr{X}}(x) = \exp(\beta x)$ on $(-\infty, 0]$, 1 on $[0, 1]$, and $\exp[-\beta(x - 1)]$ on $[1, \infty)$, respectively, we get $\widehat{M}(\omega) = \{i\beta\,[\exp(-i\omega) - 1] + \omega[\exp(-i\omega) + 1]\}/(2\,\omega)$. This yields $\mu = \sqrt{2\,\pi}(\delta_0 + \delta_1 + \beta\,\mu_L)/2$, which is indeed a signed measure (here positive); see also Example 1.

For smooth kernels, it is the non-existence of the signed measure μ^* that causes the strongly oscillating behavior of the BLUE weights frequently observed in practice: the (generalized) continuous BLUE involves derivatives of y, the order of which is related to the degree of smoothness of Ψ; see the right panel of Figure 2. There, as $\Psi(x) = \exp(-\beta x^2)$ is infinitely differentiable at the origin, derivatives of all orders are involved. As $1_{\mathscr{X}} \notin \mathcal{H}(K)$, $\mathsf{var}(\widehat{\theta}^{n}_{\mathsf{BLUE}}) \to 0$ when the sequence of design points $x_1, x_2 \ldots$ is dense in $\mathscr{X}$; see Section 4.4. ◁

4 Properties of $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ and $\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})$

4.1 Fixed-size designs

Let us return to the discussion in Section 2 and consider the variances $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ and $\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})$ computed for the design $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where the $\mathbf{x}_i$ are assumed to be distinct and the kernel K is SPD. In view of (11), $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ can be written as the energy

$$\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}}) = \int_{\mathscr{X}^2} K(\mathbf{x}, \mathbf{x}')\mu_n(\mathrm{d}\mathbf{x})\mu_n(\mathrm{d}\mathbf{x}') = \mathscr{E}_K(\mu_n) \tag{16}$$

for the empirical probability measure μ_n assigning weights $1/n$ to the $\mathbf{x}_i \in \mathbf{X}_n$.

From (7), the discrete BLUE $\widehat{\theta}^n_{\mathsf{BLUE}}$ can be written as

$$\widehat{\theta}^n_{\mathsf{BLUE}} = \int_{\mathscr{X}} y(x)\mu^*_n(\mathrm{d}\mathbf{x}) \;\text{ with } \mu^*_n = \arg\min_{\nu_n} \int_{\mathscr{X}^2} K(\mathbf{x}, \mathbf{x}')\nu_n(\mathrm{d}\mathbf{x})\nu_n(\mathrm{d}\mathbf{x}'),$$

where ν_n belongs to the set of signed measures of total mass 1 supported on $\mathbf{X}_n$. The n-point optimal measure μ^*_n concentrated on $\mathbf{X}_n$ has weights $\mathbf{w}^*_n = \mathbf{w}_{n,\mathsf{BLUE}}$, and in view of (8),

$$\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}}) = \int_{\mathscr{X}^2} K(\mathbf{x}, \mathbf{x}')\mu^*_n(\mathrm{d}\mathbf{x})\mu^*_n(\mathrm{d}\mathbf{x}') = \mathscr{E}_K(\mu^*_n)\,.$$

The inequality (12) simply expresses the fact that $\mathscr{E}_K(\mu^*_n) \leq \mathscr{E}_K(\mu_n)$, which obviously follows from the definition of μ^*_n.

Let λ_1 and λ_n respectively denote the minimum and maximum eigenvalues of the matrix $\mathbf{K}_n$ associated with the design $\mathbf{X}_n$. We then have, for any $\mathbf{u} \in \mathbb{R}^n$,

$$(\mathbf{u}^\top\mathbf{u})^2 \leq (\mathbf{u}^\top\mathbf{K}_n\mathbf{u})\,(\mathbf{u}^\top\mathbf{K}_n^{-1}\mathbf{u}) \leq \frac{1}{4}\left(\sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}}\right)^2 (\mathbf{u}^\top\mathbf{u})^2\,, \tag{17}$$

where we use the Cauchy-Schwarz inequality on the left and the Kantorovich inequality on the right.

Since $\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}}) = (\mathbf{1}_n^\top\mathbf{K}_n^{-1}\mathbf{1}_n)^{-1}$ and $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}}) = \mathbf{1}_n^\top\mathbf{K}_n\mathbf{1}_n/n^2$, see (8) and (11), taking $\mathbf{u} = \mathbf{1}_n$ in the left inequality, we obtain (12), where there is equality if and only if $\mathbf{1}_n$ is an eigenvector of $\mathbf{K}_n$; that is, $\mathbf{K}_n\mathbf{1}_n = \lambda\mathbf{1}_n$ for some $\lambda > 0$, which means that the row (and column) sums of $\mathbf{K}_n$ are all identical. For other (equivalent) characterizations of the equality $\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}}) = \mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$, see [16, Sect. 10.2]. One may notice that

$$\begin{aligned}\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}}) - \mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}}) &= (\mathbf{w}_{n,\mathsf{BLUE}} - \mathbf{w}_{n,\mathsf{OLSE}})^\top\mathbf{K}_n(\mathbf{w}_{n,\mathsf{BLUE}} - \mathbf{w}_{n,\mathsf{OLSE}}) \\ &= \mathsf{MMD}^2(\mu_n, \mu^*_n) = \mathscr{E}_K(\mu_n - \mu^*_n) \geq 0\,,\end{aligned}$$

see (13), where μ_n and μ^*_n have respective weights $\mathbf{w}_{n,\mathsf{OLSE}} = \mathbf{1}_n/n$ and $\mathbf{w}_{n,\mathsf{BLUE}}$ given by (7).

The right inequality in (17) coincides with the celebrated upper bound on $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})/\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})$ proved independently in [2] and [9]. However, this upper bound is often rather pessimistic, and a better bound can be obtained as follows. As $\mathscr{E}_K(\cdot)$ defines a convex functional on $\mathscr{M}(1)$, we have

$$\mathscr{E}_K(\mu_n^*) \geq \mathscr{E}_K(\mu_n) + F_{\mathscr{E}_K}(\mu_n, \mu_n^*)\,,$$

where, for any μ and ν in $\mathscr{M}(1)$, $F_{\mathscr{E}_K}(\mu,\nu)$ denotes the directional derivative of $\mathscr{E}_K(\cdot)$ at μ in the direction ν; that is,

$$F_{\mathscr{E}_K}(\mu,\nu) = \lim_{\alpha\to 0^+} \frac{\mathscr{E}_K[(1-\epsilon)\mu + \alpha\nu] - \mathscr{E}_K(\mu)}{\alpha}\,.$$

Direct calculation gives $F_{\mathscr{E}_K}(\mu,\nu) = 2\int_{\mathscr{X}}[P_\mu(\mathbf{x}) - \mathscr{E}_K(\nu)]\nu(\mathrm{d}\mathbf{x})$, and thus

$$\mathscr{E}_K(\mu_n^*) \geq \mathscr{E}_K(\mu_n) + \inf_{\nu\in\mathscr{M}(1)} F_{\mathscr{E}_K}(\mu_n,\nu) = 2\inf_{\mathbf{x}\in\mathscr{X}} P_{\mu_n}(\mathbf{x}) - \mathscr{E}_K(\mu_n)\,.$$

We hence obtain the following bound.

Theorem 4.1. *The variances of the BLUE and OLSE in model* (1) *satisfy*

$$\frac{\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})}{\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})} \geq 2\,\frac{\inf_{x\in\mathscr{X}} P_{\mu_n}(x)}{\mathscr{E}_K(\mu_n)} - 1\,,$$

where μ_n is the uniform probability measure on $\mathbf{X}_n$.

4.2 Designs with increasing n

We start with an illustrative example.

Example 3. The left panel of Figure 4 shows the typical behavior of the variances $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ and $\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})$ as n increases. In this example, $\mathscr{X} = [0,1]$, K is the exponential (Matérn 1/2) kernel $K(x,x') = \exp(-5\,|x-x'|)$ and the sequence of designs $\mathbf{X}_n = \{x_1^{(n)},\ldots,x_n^{(n)}\}$ consists of equidistant points $x_i^{(n)} = (i-1)/(n-1)$, $i = 1,\ldots,n$. We can see that $\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})$ is monotonically decreasing with n, whereas the behavior of $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ is non-monotonic. The right panel demonstrates that the significantly larger variance of $\widehat{\theta}^n_{\mathsf{OLSE}}$ compared to $\widehat{\theta}^n_{\mathsf{BLUE}}$ has little consequence on the prediction errors: the figure shows that the difference between the integrated MSPEs for both estimators for the uniform measure, i.e., $\int_{\mathscr{X}} \widehat{\rho}^2_{n,OLSE}(x)\mathrm{d}x - \int_{\mathscr{X}} \widehat{\rho}^2_{n,BLUE}(x)\mathrm{d}x$, is negligible. Note that the situation might be different for $d > 1$ and a design $\mathbf{X}_n$ less dense in $\mathscr{X}$. ◁

Assume that K is a continuous SPD kernel and that the sequence $\{\mu_n\}$ of empirical measures μ_n associated with the designs $\mathbf{X}_n = \{\mathbf{x}_1^{(n)},\ldots,\mathbf{x}_n^{(n)}\}$ (consisting of distinct points $\mathbf{x}_i^{(n)} \in \mathscr{X}$) weakly converges as $n\to\infty$ to some

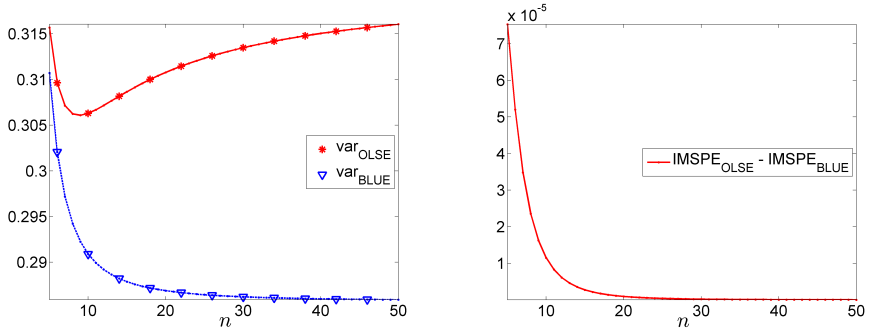

Figure 4: Left: $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ and $\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})$ as functions of n. Right: difference between the integrated MSPEs for $\widehat{\theta}^n_{\mathsf{OLSE}}$ and $\widehat{\theta}^n_{\mathsf{BLUE}}$ as function of n.

limiting probability measure μ_∞ (not necessarily uniform on $\mathscr{X}$). We assume that $\mathscr{X}$ is a compact subset of $\mathbb{R}^d$ and that the support of μ_∞ coincides with $\mathscr{X}$. In the community of optimal design of experiments, the limiting probability measure μ_∞ would be considered as an "approximate design". As $\widehat{\theta}^n_{\mathsf{OLSE}} = \int y(\mathbf{x})\mu_n(\mathrm{d}\mathbf{x})$ for any $\mathbf{X}_n$, the OLSE for the approximate design μ_∞ is $\widehat{\theta}^\infty_{\mathsf{OLSE}} = \int y(\mathbf{x})\mu_\infty(\mathrm{d}\mathbf{x})$. We will call this OLSE the "continuous OLSE".

The inspection of Figure 4 suggests the following general questions.

(i) Does the limit $\lim_{n\to\infty} \mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ exist and what is it?

(ii) Is it possible that $\lim_{n\to\infty} \mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}}) = \lim_{n\to\infty} \mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}}) = 0$?

(iii) Is the non-monotonic behavior of $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ similar to depicted in Figure 4 (left) typical?

(iv) Does the limit $\lim_{n\to\infty} \mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})$ exist and what is it?

(v) Assuming $\lim_{n\to\infty} \mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}}) > 0$, what are the conditions guaranteeing

$$\lim_{n\to\infty} \mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}}) = \lim_{n\to\infty} \mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})\,? \tag{18}$$

We can give (at least partial) answers to all these questions. Let us start with question (i). Since the probability measures μ_n weakly converge to μ_∞, the kernel K is continuous and bounded (as it is PD) and $\mathscr{X}$ is compact, (16) gives $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}}) = \mathscr{E}_K(\mu_n) \to \mathscr{E}_K(\mu_\infty) = \mathsf{var}(\widehat{\theta}^\infty_{\mathsf{OLSE}})$ as $n \to \infty$. As K is PD, $\mathscr{E}_K(\mu_\infty) \geq 0$. Families of kernels K and measures μ_∞ such that $\mathscr{E}_K(\mu_\infty) = 0$ are provided in Section 4.5; since $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}}) \geq \mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})$ for all n, these families provide examples where we have an affirmative answer to question (ii).

The non-monotonic behavior of $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ queried in question (iii) is related to "Smit's paradox" (see [10, p. 50]). This paradox is discussed in Section 4.3

below. Question (iv) will be addressed in Section 4.4. Finally, concerning (v), the conditions for (18) to hold follow from those mentioned in Section 3.3: $\lim_{n\to\infty} \mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}}) = \lim_{n\to\infty} \mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})$ if and only if $\mu_\infty = \mu^+ = \mu^*$: that is, the approximate design μ_∞ coincides with the minimum-energy probability measure μ^+, which at the same time equals μ^*, the minimum-energy signed measure of mass one.

4.3 Behaviour of $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ as n increases

Smit's paradox refers to the non-monotonic behavior of $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ as shown in the left panel of Figure 4. It was firstly observed and investigated in [25] and [22] for a particular class of stationary kernels $K(x,x') = \psi(x-x')$ with $\int_0^1 \psi(x)(1-2x)\,\mathrm{d}x > 0$ in the case of designs formed by equidistant points in $\mathscr{X} = [0,1]$ of the form (d) below for $\mu_\infty = \mathscr{U}_{[0,1]}$. At first glance, such an increase of $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ with n does not look natural, but in fact, when $\mu_\infty \neq \mu^+ = \arg\min_{\mu\in\mathscr{M}^+(1)} \mathscr{E}_K(\mu)$ then any type of convergence of $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ towards $\mathscr{E}_K(\mu_\infty)$ is possible, including the one depicted in Figure 4 (left). Example 4 will give a further illustration.

The following four classes of designs $\mathbf{X}_n$, defined for an arbitrary probability measure $\mu_\infty \in \mathscr{M}^+(1)$, are often used in practical considerations:

(a) $\mathbf{X}_1 = \{\mathbf{x}_1\}$ with arbitrary $\mathbf{x}_1 \in \mathscr{X}$ and $\mathbf{X}_{n+1} = \mathbf{X}_n \cup \{\mathbf{x}_{n+1}\}$, where $\mathbf{x}_{n+1} = \arg\min_{\mathbf{x}} \mathsf{MMD}^2(\mu_{n,\mathbf{x}}, \mu_\infty)$, where $\mu_{n,\mathbf{x}}$ is the discrete measure concentrated on the set of $n+1$ points $\mathbf{X}_n \cup \{\mathbf{x}\}$ and assigning the same weight $1/(n+1)$ to all these points;

(b) $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ are nested designs and the points $\{\mathbf{x}_1, \mathbf{x}_2 \ldots,\}$ are chosen so that the sequence of empirical measures μ_n corresponding to designs $\mathbf{X}_n$ converges to a probability measure μ_∞;

(c) $\mathscr{X} = [0,1]$ and $\mathbf{X}_n = \{x_1^{(n)}, \ldots, x_n^{(n)}\}$, where $x_j^{(n)}$ is the $(j-0.5)/n$-quantile of a given probability measure μ_∞;

(d) $\mathscr{X} = [0,1]$ and $\mathbf{X}_n = \{x_1^{(n)}, \ldots, x_n^{(n)}\}$, where $x_j^{(n)}$ is the $(j-1)/(n-1)$-quantile of a given probability measure μ_∞.

In all four cases, if $\mu_\infty = \mu^+$, then the empirical measures μ_n of designs $\mathbf{X}_n$ converge to μ^+ and therefore $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}}) = \mathscr{E}_K(\mu_n) \to \mathscr{E}_K^+$. In case (a), the sequence $\mathscr{E}_K(\mu_n)$ is monotonously decreasing (with rate at least $\mathcal{O}([\log n]/n)$, see, e.g., [13]); in case (b) one should expect small fluctuations on the route to the limit. In cases (c) and (d) the behavior of $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ is often monotonic; however, for kernels such as $K(x,x') = \cos^2[\beta\pi(x-x')]$ or $K(x,x') = \max\{0, 1-\beta|x-x'|\}$ with $\beta > 1$, a rather irregular behavior of $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ may be observed. Note finally that if θ is estimated by $\widehat{\theta}^n_{\mathsf{OLSE}}$, then the use of any other μ_∞ than μ^+ is not optimal, and hence unreasonable. From this view-point, Smit's paradox has little practical significance.

Example 4. We use again the Matérn 1/2 kernel, as in Example 3 but with a different correlation length. The behavior of $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ shown in the left panel of Figure 5 for the two designs (c) and (d) with $\mu_\infty = \mathscr{U}_{[0,1]}$ $(\neq \mu^+)$ is now much different from that on the left panel of Figure 4. On the right panel the kernel is $K(x,x') = \max\{0, 1 - 5\,|x - x'|\}$, and $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ oscillates during its convergence to $\mathscr{E}_K(\mu_\infty)$ from below or from above. ◁

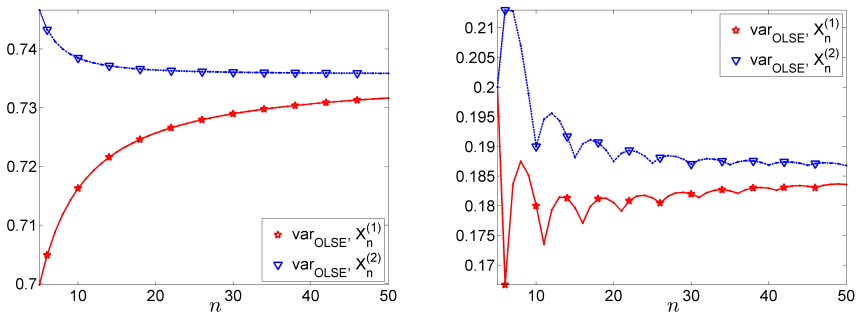

Figure 5: $\mathsf{var}(\widehat{\theta}^n_{\mathsf{OLSE}})$ as a function of n for two different designs in $[0,1]$: $\mathbf{X}_n^{(1)} = \{(i-1)/(n-1),\, i = 1,\dots,n\}$ and $\mathbf{X}_n^{(2)} = \{(i-1/2)/n,\, i = 1,\dots,n\}$. Left: $K(x,x') = \exp(-|x-x'|)$; Right: $K(x,x') = \max\{0, 1 - 5\,|x-x'|\}$.

4.4 Computing $\lim_{n\to\infty} \mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})$, continuous BLUE

Consider the one-parameter model

$$y(\mathbf{x}) = \theta f(\mathbf{x}) + \varepsilon(\mathbf{x}), \quad \mathsf{E}\{\varepsilon(\mathbf{x})\} = 0,\ \mathsf{E}\{\varepsilon(\mathbf{x})\varepsilon(\mathbf{x}')\} = K(\mathbf{x},\mathbf{x}'),$$

which is a generalization of model (1) to an arbitrary measurable function f (we have $f = 1_{\mathscr{X}}$ in model (1)). As usual, assume that K is SPD so that all kernel matrices $\mathbf{K}_n$ below are invertible and K generates an RKHS $\mathcal{H}(K)$.

Assume that $\mathbf{x}_1, \mathbf{x}_2 \dots$ is a dense sequence of distinct points in $\mathscr{X}$ and consider the sequence of nested designs $\mathbf{X}_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, $n = 1, 2 \dots$ From the definition of the discrete BLUE, we immediately conclude that the sequence of $\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})$ is monotonically decreasing with n. Therefore, the limit $\lim_{n\to\infty} \mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})$ exists and it is a non-negative number. The next theorem (which is a corollary of [11, Th. 6C]) gives the value of this limit in case $f \in \mathcal{H}(K)$.

Theorem 4.2. *$f \in \mathcal{H}(K)$ if and only if*

$$\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}}) \to 1/\|f\|^2_{\mathcal{H}(K)} \quad \textit{as } n \to \infty. \tag{19}$$

In our main case, when $f = 1_{\mathscr{X}}$, the result (19) can be explained and specialized as follows. In view of (7), the n-point BLUE $\widehat{\theta}^n_{\mathsf{BLUE}}$ can be written as $\widehat{\theta}^n_{\mathsf{BLUE}} = \int y(\mathbf{x})\mu^*_n(\mathrm{d}\mathbf{x})$, where μ^*_n is the minimum-energy signed measure of mass 1 concentrated on $\mathbf{X}_n$ and is defined by the vector of weights $\mathbf{w}_{n,\mathsf{BLUE}}$. Let $\mathcal{H}_n(K)$ be the RKHS induced by the kernel K on the set $\mathbf{X}_n$. The scalar product in $\mathcal{H}_n(K)$ is defined by $\langle \mathbf{a}, \mathbf{b}\rangle_{\mathcal{H}_n(K)} = \mathbf{a}^\top \mathbf{K}_n^{-1}\mathbf{b}$ for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$; see e.g. [12, Sect. 2.3.3]. The restriction of the function $1_{\mathscr{X}}$ to the set $\mathbf{X}_n$ is the vector $\mathbf{1}_n$ and therefore (8) implies that the variance of $\widehat{\theta}^n_{\mathsf{BLUE}}$ equals $\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}}) = 1/\|\mathbf{1}_n\|^2_{\mathcal{H}_n(K)}$.

Assume that the function $1_{\mathscr{X}}$ belongs to $\mathcal{H}(K)$. Then, as this function is continuous and the sequence of points $\mathbf{x}_1, \mathbf{x}_2 \ldots$ is dense in $\mathscr{X}$, $\|\mathbf{1}_n\|_{\mathcal{H}_n(K)} \to \|1_{\mathscr{X}}\|_{\mathcal{H}(K)}$ as $n \to \infty$, implying (19) for $f = 1_{\mathscr{X}}$; see [17, Th. 2.8 & Remark 2.3]. Moreover, if $1_{\mathscr{X}} \in \mathcal{P}_K$ and hence μ^*, the minimum-energy signed measure of mass 1, exists (see Corollary 3.1-(i)), then the sequence of signed measures μ^*_n weakly converges to μ^*. In this case, we can define the continuous BLUE of θ by $\widehat{\theta}^\infty_{\mathsf{BLUE}} = \int y(\mathbf{x})\mu^*(\mathrm{d}\mathbf{x})$.

If $1_{\mathscr{X}} \notin \mathcal{P}_K$, then μ^* does not exist and the sequence of signed measures μ^*_n does not have a weak limit (as the set $\mathscr{M}(1)$ is not weakly compact, the sequence $\{\mu^*_n\} \subset \mathscr{M}(1)$ does not necessarily have a convergent subsequence). In some cases, we can define a generalized continuous BLUE of θ and sometimes write it as $\widehat{\theta}^\infty_{\mathsf{BLUE}} = \int_{\mathscr{X}} y(\mathbf{x})\phi(\mathbf{x})\mathrm{d}\mathbf{x}$, where $\phi(\mathbf{x})$ is a generalized function or distribution (see e.g. [6]). If $\phi(\cdot)$ is a distribution rather than a function, then the expression $\widehat{\theta}^\infty_{\mathsf{BLUE}} = \int y(\mathbf{x})\phi(\mathbf{x})\mathrm{d}\mathbf{x}$ involves mean-square derivatives of the random field $y(\mathbf{x})$, which always exist if the kernel K is differentiable at the diagonal; see [4] for details and references and Example 2 for an illustration.

Theorem 4.2 in the case $f = 1_{\mathscr{X}}$ is also closely related to the minimization problem $\mathscr{E}_K(\mu) \to \min_{\mu\in\mathscr{M}(1)}$ considered in Section 3.2. If K is IPD, then $\mathscr{E}^*_K = \inf_{\mu\in\mathscr{M}^+(1)} \geq 0$ and the above arguments imply that $\mathscr{E}^*_K = \lim_{n\to\infty} 1/\mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}})$ and $1_{\mathscr{X}} \in \mathcal{H}(K)$ if and only if $\mathscr{E}^*_K > 0$.

Summarizing, for the main case of interest here, with $f = 1_{\mathscr{X}}$, we conclude the following (where we define $\|1_{\mathscr{X}}\|_{\mathcal{H}(K)} = \infty$ when $1_{\mathscr{X}} \notin \mathcal{H}(K)$).

Corollary 4.1.

(a) $\lim_{n\to\infty} \mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}}) = \mathscr{E}^*_K = 1/\|1_{\mathscr{X}}\|^2_{\mathcal{H}(K)}$.

(b) $\lim_{n\to\infty} \mathsf{var}(\widehat{\theta}^n_{\mathsf{BLUE}}) > 0$ *if and only if* $1_{\mathscr{X}} \in \mathcal{H}(K)$.

4.5 Reduced kernels

This section shows that for any signed measure $\mu \in \mathscr{M}(1)$, one can construct kernels that admit μ as minimum-energy signed measure of total mass 1. It also establishes links between parameter estimation and integration in the location model.

Consider again model (1), with K an uniformly bounded kernel. Let $\mu \in \mathscr{M}$ and define $\mathbb{P}$ as the orthogonal projection from $L^2(\mathscr{X}, \mu)$ onto the linear space

spanned by $1_{\mathscr{X}}$. We can write

$$y(\mathbf{x}) = \theta + \varepsilon(\mathbf{x}) = \theta + \mathbb{P}\varepsilon(\mathbf{x}) + (\mathrm{Id}_{L^2} - \mathbb{P})\varepsilon(\mathbf{x}) = \vartheta + \varepsilon_\mu(\mathbf{x})\,,$$

where $\vartheta = \theta + \mathbb{P}\varepsilon(\mathbf{x})$ is a new parameter and $\varepsilon_\mu(\mathbf{x}) = (\mathrm{Id}_{L^2} - \mathbb{P})\varepsilon(\mathbf{x})$ is a new random process with zero mean and covariance K_μ given by the reduced kernel (the reduction of K with respect to μ)

$$K_\mu(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}, \mathbf{x}') - P_{K,\mu}(\mathbf{x}) - P_{K,\mu}(\mathbf{x}') + \mathscr{E}_K(\mu)\,. \tag{20}$$

In this model, $\widehat{\vartheta}^n_{\mathsf{BLUE}} = \widetilde{\mathbf{w}}^\top_{n,\mathsf{BLUE}}\mathbf{y}_n$ is the BLUE of the integral $\int_{\mathscr{X}} y(\mathbf{x})\mu(\mathrm{d}\mathbf{x})$ and $(\mathbf{1}_n^\top[\mathbf{K}_\mu]_n^{-1}\mathbf{1}_n)^{-1}$ is its variance, and the continuous BLUE of ϑ always exists: $\widehat{\vartheta}^\infty_{\mathsf{BLUE}} = \int y(\mathbf{x})\mu(\mathrm{d}\mathbf{x})$; see [14].

In the next theorem, we establish fundamental properties of reduced kernels showing their specificity. Property (ii) implies, in particular, that $\mathsf{MMD}^2(\mu,\nu) = \mathscr{E}_K(\mu-\nu) = \mathscr{E}_{K_\mu}(\nu)$ for any $\mu \in \mathscr{M}$ and $\nu \in \mathscr{M}(1)$. A consequence of this is $\mathscr{E}_{K_\mu}(\mu) = 0$ for any kernel K and any signed measure μ. In particular, if K is an IPD kernel and $\mu \in \mathscr{M}(1)$, then μ is the minimum-energy signed measure of mass 1 for K_μ, with $\mathscr{E}^*_{K_\mu} = \mathscr{E}_{K_\mu}(\mu) = 0$ implying that the non-zero constant functions do not belong to the RKHS $\mathcal{H}(K_\mu)$; see Corollary 4.1-(b). Moreover, for any dense sequence of design points $\mathbf{x}_1, \mathbf{x}_2 \ldots$ in $\mathscr{X}$, the measure μ^*_n with weights $\widetilde{\mathbf{w}}_{n,\mathsf{BLUE}} = [\mathbf{K}_\mu]_n^{-1}\mathbf{1}_n/(\mathbf{1}_n^\top[\mathbf{K}_\mu]_n^{-1}\mathbf{1}_n)$ weakly converges to μ. Property (*iii*) is a generalization of Schoenberg's result (see [20, Sect. 3] and [12, Sect. 9.1]) where μ is a delta measure. Properties (*iv*) and (*v*) are further extensions of this result. We recall that a kernel K is Conditionally Positive Definite (CPD) if and only if for any $n \in \mathbb{N}$ and any n-point design $\mathbf{X}_n$, $\mathbf{z}_n\mathbf{K}_n\mathbf{z}_n \geq 0$ for any $\mathbf{z}_n \in \mathbb{R}^n$ such that $\mathbf{1}_n^\top\mathbf{z}_n = 0$.

Theorem 4.3.

(i) For any $\mu \in \mathscr{M}$ and $\nu \in \mathscr{M}(1)$, we have $[K_\mu]_\nu = K_\nu$.

(ii) For any $\mu, \xi \in \mathscr{M}$, we have

$$\mathscr{E}_{K_\mu}(\xi) = \mathscr{E}_K[\xi - \xi(\mathscr{X})\mu] = \mathscr{E}_{K_\mu}[\xi - \xi(\mathscr{X})\mu]\,. \tag{21}$$

(iii) Let $\mu \in \mathscr{M}(1)$ be a discrete measure with finite support. Then

K is CPD if and only if K_μ is PD.

(iv) For any $\mu \in \mathscr{M}(1)$, we have:

K is CIPD if and only if K_μ is IPD.

If, moreover, K is CISPD, then μ is the unique minimum-energy signed measure in $\mathscr{M}(1)$ for K_μ.

(v) Let $\mu \in \mathscr{M}(1)$ be a measure with infinite support and K be CISPD. Then K_μ is SPD.

Proof. (i) We have:

$$\begin{aligned}
&[K_\mu]_\nu(\mathbf{x},\mathbf{x}') = K_\mu(\mathbf{x},\mathbf{x}') - P_{K_\mu,\nu}(\mathbf{x}) - P_{K_\mu,\nu}(\mathbf{x}') + \mathscr{E}_{K_\mu}(\nu) = \\
&[K(\mathbf{x},\mathbf{x}') - P_{K,\mu}(\mathbf{x}) - P_{K,\mu}(\mathbf{x}') + \mathscr{E}_K(\mu)] - \int_{\mathscr{X}} K_\mu(\mathbf{x},\mathbf{z})\nu(\mathrm{d}\mathbf{z}) \\
&- \int_{\mathscr{X}} K_\mu(\mathbf{x}',\mathbf{z})\nu(\mathrm{d}\mathbf{z}) + \int_{\mathscr{X}^2} K_\mu(\mathbf{z},\mathbf{z}')\nu(\mathrm{d}\mathbf{z})\nu(\mathrm{d}\mathbf{z}') \\
= \;& [K(\mathbf{x},\mathbf{x}') - P_{K,\mu}(\mathbf{x}) - P_{K,\mu}(\mathbf{x}') + \mathscr{E}_K(\mu)] \\
- \;& [P_{K,\nu}(\mathbf{x}) - P_{K,\mu}(\mathbf{x}) - \mathscr{E}_K(\mu,\nu) + \mathscr{E}_K(\mu)] \\
- \;& [P_{K,\nu}(\mathbf{x}') - P_{K,\mu}(\mathbf{x}') - \mathscr{E}_K(\mu,\nu) + \mathscr{E}_K(\mu)] \\
+ \;& [\mathscr{E}_K(\nu) - 2\mathscr{E}_K(\mu,\nu) + \mathscr{E}_K(\mu)] = K_\nu(\mathbf{x},\mathbf{x}')\,.
\end{aligned}$$

(ii) Direct calculation using (20) gives

$$\mathscr{E}_{K_\mu}(\xi) = \mathscr{E}_K(\xi) - 2\,\xi(\mathscr{X})\,\mathscr{E}_K(\mu,\xi) + [\xi(\mathscr{X})]^2\,\mathscr{E}_K(\mu) = \mathscr{E}_K[\xi - \xi(\mathscr{X})\mu]\,.$$

Therefore, $\mathscr{E}_{K_\mu}[\xi - \xi(\mathscr{X})\mu] = \mathscr{E}_K[\xi - \xi(\mathscr{X})\mu] = \mathscr{E}_{K_\mu}(\xi)$, which gives (21).

(iii) Let $\mu = \sum_{i=1}^m u_i\delta_{\mathbf{s}_i}$ with $\mathbf{s}_i \in \mathscr{X}$ for all i and $\mathbf{u}_m^\top\mathbf{1}_m = \sum_{i=1}^m u_i = 1$, where $\mathbf{u}_m = (u_1,\ldots,u_m)^\top$. Take any n-point design $\mathbf{X}_n \subset \mathscr{X}$. The corresponding kernel matrix $\mathbf{K}_{\mu_n} = [\mathbf{K}_\mu]_n$ is given by

$$\mathbf{K}_{\mu_n} = \mathbf{K}_n - \mathbf{p}_{K,n}(\mu)\mathbf{1}_n^\top - \mathbf{1}_n\mathbf{p}_{K,n}^\top(\mu) + \mathscr{E}_K(\mu)\mathbf{1}_n\mathbf{1}_n^\top\,,$$

where

$$\mathbf{p}_{K,n}(\mu) = [P_{K,\mu}(\mathbf{x}_1),\ldots,P_{K,\mu}(\mathbf{x}_n)]^\top\,. \tag{22}$$

We have $\mathscr{E}_K(\mu) = \mathbf{u}_m^\top\mathbf{K}_m\mathbf{u}_m$ and $\mathbf{p}_{K,n}(\mu) = \mathbf{K}_{n,m}\mathbf{u}_m$, where $\mathbf{K}_m$ is the kernel matrix for the support points $\mathbf{s}_i$ of μ and $\{\mathbf{K}_{n,m}\}_{i,j} = K(\mathbf{x}_i,\mathbf{s}_j)$; $i = 1,\ldots,n$, $j = 1,\ldots,m$.

For any $\mathbf{z}_n \in \mathbb{R}^n$, we have

$$t_n = \mathbf{z}_n^\top\mathbf{K}_{\mu_n}\mathbf{z}_n = \mathbf{z}_n^\top\mathbf{K}_n\mathbf{z}_n - 2\,(\mathbf{z}_n^\top\mathbf{1}_n)[\mathbf{z}_n^\top\mathbf{p}_{K,n}(\mu)] + (\mathbf{z}_n^\top\mathbf{1}_n)^2\mathscr{E}_K(\mu). \tag{23}$$

Assume that K is CPD. If $\mathbf{z}_n^\top\mathbf{1}_n = 0$, then $t_n = \mathbf{z}_n^\top\mathbf{K}_n\mathbf{z}_n \geq 0$. Otherwise,

$$t_n = (\mathbf{z}_n^\top\mathbf{1}_n)^2\,[\mathbf{w}_n^\top \;\; -\mathbf{u}_m^\top] \begin{bmatrix} \mathbf{K}_n & \mathbf{K}_{n,m} \\ \mathbf{K}_{n,m}^\top & \mathbf{K}_m \end{bmatrix} \begin{bmatrix} \mathbf{w}_n \\ -\mathbf{u}_m \end{bmatrix},$$

where $\mathbf{w}_n = \mathbf{z}_n/(\mathbf{z}_n^\top\mathbf{1}_n)$. As $\mathbf{1}_n^\top\mathbf{w}_n - \mathbf{1}_m^\top\mathbf{u}_m = 0$ and K is CPD, $t_n \geq 0$.

Conversely, assume that K_μ is PD. For any $\mathbf{z}_n \in \mathbb{R}^n$ such that $\mathbf{1}_n^\top\mathbf{z}_n = 0$, (23) implies $\mathbf{z}_n^\top\mathbf{K}_n\mathbf{z}_n = \mathbf{z}_n^\top\mathbf{K}_{\mu_n}\mathbf{z}_n \geq 0$.

(iv) Assume that K is CIPD. As $\mu \in \mathscr{M}(1)$, for any $\xi \in \mathscr{M}$ the measure $\xi - \xi(\mathscr{X})\mu$ has total mass zero, and (21) implies that $\mathscr{E}_{K_\mu}(\xi) \geq 0$. Conversely, assume that K_μ is IPD. For any $\xi \in \mathscr{M}(0)$ we have $\mathscr{E}_K(\xi) = \mathscr{E}_{K_\mu}(\xi) \geq 0$.

Assume now that K is CISPD. Take any $\xi \in \mathscr{M}(1)$. From (21), we have $\mathscr{E}_{K_\mu}(\xi) = \mathscr{E}_K(\xi - \mu) \geq 0$ with equality if and only if $\xi = \mu$, which proves that μ is the unique minimum-energy signed measure in $\mathscr{M}(1)$ for K_μ.

(v) Take any n-point design $\mathbf{X}_n \subset \mathscr{X}$ and any $\mathbf{z}_n = (z_1, \ldots, z_n)^\top \in \mathbb{R}^n$, $\mathbf{z}_n \neq \mathbf{0}_n$. If $s_n = \mathbf{1}_n^\top \mathbf{z}_n \neq 0$, denote $\xi_n = (1/s_n) \sum_{i=1}^n z_i \delta_{\mathbf{x}_i}$; ξ_n belongs to $\mathscr{M}(1)$, it has finite support and therefore cannot coincide with μ. This implies $\mathbf{z}_n^\top \mathbf{K}_{\mu_n} \mathbf{z}_n = s_n^2 \mathscr{E}_{K_\mu}(\xi_n) > 0$. If $s_n = 0$, denote $\nu_n = \sum_{i=1}^n z_i \delta_{\mathbf{x}_i}$. From (21), we have $\mathbf{z}_n^\top \mathbf{K}_{\mu_n} \mathbf{z}_n = \mathscr{E}_{K_\mu}(\nu_n) = \mathscr{E}_K(\nu_n)$, which is strictly positive since K is CISPD. □

Example 5. The kernel $K(\mathbf{x}, \mathbf{x}') = -\|\mathbf{x} - \mathbf{x}'\|$ is CPD; for $\mu = \delta_{\mathbf{0}}$ (the delta measure at the origin), the associated reduced kernel is $K_{\delta_{\mathbf{0}}}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x}\| + \|\mathbf{x}'\| - \|\mathbf{x} - \mathbf{x}'\|$, i.e., the energy-distance kernel of [24], which is PD (but not SPD as $K_{\delta_{\mathbf{0}}}(\mathbf{0}, \mathbf{0}) = 0$). One may refer to [21] for a thorough exposition on distance-induced kernels and their properties. ◁

We finally make the following observation. Let K be a CI(S)PD kernel and assume that there exists a minimum-energy signed measure of total mass one μ^* (μ^* is uniquely defined from Theorem 3.1), with infinite support. From Corollary 3.2, this is the case in particular when the minimum-energy probability measure μ^+ (which always exists) has full support, since then $\mu^* = \mu^+$. The reduced kernel K_{μ^*} is thus (S)PD from Theorem 4.3-(v). Since $P_{K,\mu^*}(\mathbf{x}) = \mathscr{E}_K(\mu^*)$ for all $\mathbf{x} \in \mathscr{X}$ (see Theorem 3.1), we have, $K_{\mu^*}(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}, \mathbf{x}') - \mathscr{E}_K(\mu^*)$ for any $\mathbf{x}, \mathbf{x}' \in \mathscr{X}$, implying that $K - C$ is (S)PD for any constant $C \leq \mathscr{E}_K(\mu^*)$.

5 Conclusions

We have shown the existing connections between the following problems: kriging for prediction of values of a random field (Section 2), energy minimization (Section 3), and parameter estimation in the location model with correlated observations, using the OLSE or the BLUE. As shown in Section 4, the asymptotic variances of those estimators refer to different energy minimization problems. It appears that the constant function $1_{\mathscr{X}}$ plays a very important role in our study. In particular, as outlined in Corollary 4.1, the limiting value of the variances of the BLUE for a dense sequence of designs is positive if and only if $1_{\mathscr{X}}$ belongs to the RKHS generated by the kernel K. In Subsection 3.3, we have formulated two conjectures concerning the existence of μ^*, the minimum energy signed measure of mass 1. In view of Corollary 3.1-*(i)*, connecting the existence of μ^* to the property $1_{\mathscr{X}} \in \mathcal{P}_K$, where $\mathcal{P}_K$ is the space of potentials, these two conjectures can be reformulated in terms of the constant function $1_{\mathscr{X}}$, as written below (C1 and C2). We also add a third conjecture, C3, where the implication "if" has been established in [5].

Assume that $\mathscr{X}$ is a compact subset of $\mathbb{R}^d$ with nonempty interior and that the kernel K is translation invariant, i.e., $K(\mathbf{x}, \mathbf{x}') = \Psi(\mathbf{x} - \mathbf{x}')$.

C1′: If $\Psi(\mathbf{0}) < \infty$ and $d > 1$, then $1_{\mathscr{X}} \notin \mathcal{P}_K$.

C2′: If Ψ is differentiable at the origin, then $1_{\mathscr{X}} \notin \mathcal{P}_K$.

C3: Assume that K is SPD and defines an RKHS $\mathcal{H}(K)$, and that the Fourier transform $\widehat{\Psi}$ of Ψ has no mass at $\mathbf{0}$; then $\widehat{\Psi}$ is moment-determinant if and only if $1_{\mathscr{X}} \notin \mathcal{H}(K)$.

Acknowledgments

The work of the first author was partly supported by project INDEX (INcremental Design of EXperiments) ANR-18-CE91-0007 of the French National Research Agency (ANR). The authors are grateful to Toni Karvonen (University of Helsinki) for useful advice and especially for proposing a family of translation-invariant kernels with non-convex functions Ψ and $1_{\mathscr{X}} \in \mathcal{P}_K$ used in Example 1.

References

[1] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, 2011.

[2] P. Bloomfield and G.S. Watson. The inefficiency of least squares. *Biometrika*, 62(1):121–128, 1975.

[3] S.V. Borodachov, D.P. Hardin, and E.B. Saff. *Discrete energy on rectifiable sets*. Springer, 2019.

[4] H. Dette, A. Pepelyshev, and A. Zhigljavsky. The BLUE in continuous-time regression models with correlated errors. *The Annals of Statistics*, 47(4):1928–1959, 2019.

[5] H. Dette and A. Zhigljavsky. Reproducing kernel Hilbert spaces, polynomials, and the classical moment problem. *SIAM/ASA Journal on Uncertainty Quantification*, 9(4):1589–1614, 2021.

[6] G. Dijk. *Distribution Theory: Convolution, Fourier Transform, and Laplace Transform*. Walter de Gruyter, 2013.

[7] B. Gauthier and L. Pronzato. Convex relaxation for IMSE optimal design in random field models. *Computational Statistics and Data Analysis*, 113:375–394, 2017.

[8] J. Hájek. Linear estimation of the mean value of a stationary random process with convex correlation function. *Czechoslovak Mathematical Journal*, 6(81):94–117, 1956.

[9] M. Knott. On the minimum efficiency of least squares. *Biometrika*, 62(1):129–132, 1975.

[10] W. Näther. *Effective Observation of Random Fields*. Teubner, 1985.

[11] E. Parzen. An approach to time series analysis. *The Annals of Mathematical Statistics*, 32(4):951–989, 1961.

[12] V.I. Paulsen and M. Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, Cambridge, UK, 2016.

[13] L. Pronzato. Performance analysis of greedy algorithms for minimising a maximum mean discrepancy. *Statistics and Computing*, 2023. (to appear, hal-03114891, arXiv:2101.07564).

[14] L. Pronzato and A. Zhigljavsky. Bayesian quadrature, energy minimization, and space-filling design. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):959–1011, 2020.

[15] L. Pronzato and A. Zhigljavsky. Minimum-energy measures for singular kernels. *Journal of Computational and Applied Mathematics*, 382:113089, 2021.

[16] S. Puntanen, G.P.H. Styan, and J. Isotalo. *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*. Springer, 2011.

[17] S. Saitoh and Y. Sawano. *Theory of Reproducing Kernels and Applications*. Springer, 2016.

[18] T.J. Santner, B.J. Williams, and W.I. Notz. *The Design and Analysis of Computer Experiments*. Springer, Heidelberg, 2003.

[19] R. Schaback. Native Hilbert spaces for radial basis functions I. In *New Developments in Approximation Theory*, pages 255–282. Springer, 1999.

[20] I.J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

[21] S. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

[22] J.C. Smit. Estimation of the mean of a stationary stochastic process by equidistant observations. *Trabajos de Estadistica*, 12(1):35–45, 1961.

[23] B.K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G.R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

[24] G.J. Székely and M.L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

[25] S.Y. Vilenkin. On the estimation of the mean in stationary processes. *Theory of Probability & Its Applications*, 4(4):415–416, 1959.

[26] A. Zhigljavsky, H. Dette, and A. Pepelyshev. A new approach to optimal design for linear models with correlated observations. *Journal of the American Statistical Association*, 105:1093–1103, 2010.