BAYESIAN QUADRATURE, ENERGY MINIMIZATION AND SPACE-FILLING DESIGN

LUC PRONZATO[†] AND ANATOLY ZHIGLJAVSKY[*]

Abstract. A standard objective in computer experiments is to approximate the behavior of an unknown function on a compact domain from a few evaluations inside the domain. When little is known about the function, space-filling design is advisable: typically, points of evaluation spread out across the available space are obtained by minimizing a geometrical (for instance, covering radius) or a discrepancy criterion measuring distance to uniformity. The paper investigates connections between design for integration (quadrature design), construction of the (continuous) BLUE for the location model, space-filling design, and minimization of energy (kernel discrepancy) for signed measures. Integrally strictly positive definite kernels define strictly convex energy functionals, with an equivalence between the notions of potential and directional derivative, showing the strong relation between discrepancy minimization and more traditional design of optimal experiments. In particular, kernel herding algorithms, which are special instances of vertex-direction methods used in optimal design, can be applied to the construction of point sequences with suitable space-filling properties.

Keywords: Bayesian quadrature, BLUE, energy minimization, potential, discrepancy, space-filling design
AMS subject classifications: 62K99, 65D30, 65D99.

1. Introduction. The design of computer experiments, where observations of a real physical phenomenon are replaced by simulations of a complex mathematical model (e.g., based on PDEs), has emerged as a full discipline, central to uncertainty quantification. The final objective of the simulations is often goal-oriented and precisely defined. It may correspond for example to the optimization of the response of a system with respect to its input factors, or to the estimation of the probability that the response will exceed a given threshold when input factors have a given probability distribution. Achieving this objective generally requires sequential learning of the behavior of the response in a particular domain of interest for input factors: the region where the response is close to its optimum, or is close to the given threshold; see, e.g., the references in [38]. When simulations are computationally expensive, sequential inference based on the direct use of the mathematical model is unfeasible due to the large number of simulations required and simplified prediction models, approximating the simulated response, have to be used. A most popular approach relies on Gaussian process modelling, where the response (unknown prior to simulation) is considered as the realization of a Gaussian Random Field (RF), with parameterized mean and covariance, and Bayesian inference gives access to the posterior distribution of the RF (after simulation). Typically, in a goal-oriented approach based on stepwise uncertainty reduction [7, 8], the prediction model is used to select the input factors to be used for the next simulation, the selection being optimal in terms of predicted uncertainty on the target. The construction of a first, possibly crude, prediction model is necessary to initialize the procedure. This amounts to approximating the behavior of an unknown function f (the model response) on a compact domain $\mathscr{X} \subset \mathbb{R}^d$ (the feasible set for d input factors) from a few evaluations inside the domain. That is the basic design objective we shall keep in mind throughout the paper, although we may use diverted paths where approximation/prediction will be shadowed by other objectives, integration in particular.

[*]Cardiff University, UK
[†]CNRS, Université Côte d'Azur, I3S, France
[‡]Luc.Pronzato@cnrs.fr (corresponding author)

In general, little is known about the function *a priori*, and it seems intuitively reasonable to spread out points of evaluation across the available space; see [10]. Such space-filling designs can be obtained by optimizing a geometrical measure of dispersion or a discrepancy criterion measuring distance to uniformity. When using a Gaussian RF model, minimizing the Integrated Mean-Squared Prediction Error (IMSPE) is also a popular approach, although not very much used due to its apparent complexity, see, e.g., [36, 41]. The paper promotes the use of designs optimized for integration with respect to the uniform measure for their good space-filling properties. It gives a survey of recent results on energy functionals that measure distance to uniformity and places recent approaches proposed for space-filling design, such as [64], in a general framework and perspective encompassing design for integration, construction of the (continuous) Best Linear Unbiased Estimator (BLUE) in a location model with correlated errors, and minimization of energy (kernel discrepancy) for signed measures. Our objective is to foster the use of designs obtained by minimizing a quadratic measure of discrepancy, which can be easily computed, for function approximation at the initial exploratory stage of computer experiments. In particular, we believe that such constructions are especially useful when the number of function evaluations is not fixed in advance, and one wishes to have an ordered sequence of points such that any first n points have suitable space-filling properties.

We start by a quick introduction to Bayesian function approximation and integration (Section 2), where the function is considered as the realization of a Gaussian RF with covariance structure defined by some kernel K; see in particular [57, 58, 72, 16] for Bayesian integration. Section 3 summarizes recent results on the minimization of energy functionals [21, 88, 89] and extends some to kernels with singularities, which we believe have great potential for the construction of space-filling designs. Integrally strictly positive definite kernels define strictly convex energy functionals (Lemmas 3.1 and 3.2), which yields an equivalence between the notions of potential and directional derivative that reveals the strong relation between discrepancy minimization and more traditional design of optimal experiments. Further connections are discussed: Bayesian integration is equivalent to the construction of the BLUE in a model with modified correlation structure (Section 3.5.2), so that the two associated design problems coincide; the posterior variance in Bayesian integration corresponds to the minimum of a squared kernel discrepancy for signed measures with total mass one (Theorem 4.3) and to the minimum of an energy functional for a reduced kernel (Theorem 4.4). Since the posterior variance criterion in Bayesian integration takes a very simple form, its minimization constitutes an attractive alternative to the minimization of the IMSPE. This is considered in Section 4, which starts by exploring relations between discrepancy and covering radius. In particular, kernel herding algorithms from machine learning, which are special instances of vertex-direction methods used in optimal design and can be used for the construction of point sequences with suitable space-filling properties, are considered in Section 4.4. Section 5 provides a few numerical examples. The main results are stated as theorems or lemmas; links to related work, or comments on specific aspects, are isolated in a few remarks; several illustrative examples are given to help keeping track of technical developments. Several auxiliary results are given in appendices. Appendix A describes convergence properties of the algorithms used in Section 4; it adapts some known results in the community of optimal design theory to the particular case of a quadratic criterion. Extension to design for the simultaneous estimation of several integrals is considered in Appendix B. Appendix C contains technical details for computing energy and

potential for a particular kernel.

2. Random-field models for function approximation and integration.

2.1. Space-filling design and kernel choice for function approximation.

Let $K(\cdot,\cdot)$ denote a symmetric positive definite kernel on $\mathscr{X} \times \mathscr{X}$, with associated Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_K$. Denote $K_{\mathbf{x}}(\cdot) = K(\mathbf{x},\cdot)$ and $\langle\cdot,\cdot\rangle_K$ the scalar product in $\mathcal{H}_K$, so that the reproducing property gives $\langle f, K_{\mathbf{x}}\rangle_K = f(\mathbf{x})$ for any $f \in \mathcal{H}_K$.

Consider first the common framework where the function f to be approximated is supposed to belong to $\mathcal{H}_K$. Let $\eta_n(\mathbf{x}) = \sum_{i=1}^n w_i f(\mathbf{x}_i) = \mathbf{w}_n^\top \mathbf{y}_n$ be a linear predictor of $f(\mathbf{x})$ based on evaluations of f at the n-point design $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, with $\mathbf{x}_i \in \mathscr{X}$ for all i. Throughout the paper we denote $\mathbf{w}_n = (w_1, \ldots, w_n)^\top$, $\mathbf{y}_n = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\top$, $\mathbf{k}_n(\cdot) = [K_{\mathbf{x}_1}(\cdot), \ldots, K_{\mathbf{x}_n}(\cdot)]^\top$ and $\{\mathbf{K}_n\}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$, $i,j = 1,\ldots,n$. The Cauchy-Schwarz inequality gives the classical result

$$
\begin{aligned}
|f(\mathbf{x}) - \eta_n(\mathbf{x})| = \left| f(\mathbf{x}) - \sum_{i=1}^n w_i f(\mathbf{x}_i) \right| &= \left| \langle f, K_{\mathbf{x}} - \sum_{i=1}^n w_i K_{\mathbf{x}_i} \rangle_K \right| \\
&\leq \|f\|_{\mathcal{H}_K} \left\| K_{\mathbf{x}} - \sum_{i=1}^n w_i K_{\mathbf{x}_i} \right\|_{\mathcal{H}_K},
\end{aligned}
$$

where $\|f\|_{\mathcal{H}_K}$ depends on f but not on $\mathbf{X}_n$, and $\rho_n(\mathbf{x},\mathbf{w}) = \|K_{\mathbf{x}} - \sum_{i=1}^n w_i K_{\mathbf{x}_i}\|_{\mathcal{H}_K}$ depends on $\mathbf{X}_n$ (and $\mathbf{w}_n$) but not on f. Suppose that $\mathbf{K}_n$ has full rank. For a given $\mathbf{X}_n$, the Best Linear Predictor (BLP) minimizes $\rho_n(\mathbf{x},\mathbf{w})$ and corresponds to $\eta_n^*(\mathbf{x}) = (\mathbf{w}_n^*)^\top \mathbf{y}_n$, with $\mathbf{w}_n^* = \mathbf{w}_n^*(\mathbf{x}) = \mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x})$, which gives ${\rho_n^*}^2(\mathbf{x}) = \rho_n^2(\mathbf{x},\mathbf{w}_n^*) = K(\mathbf{x},\mathbf{x}) - \mathbf{k}_n^\top(\mathbf{x})\mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x})$.

A less restrictive assumption on f is to suppose that it corresponds to a realization of a RF Z_x, with zero mean ($\mathbb{E}\{Z_x\} = 0$) and covariance $\mathbb{E}\{Z_x Z_{x'}\} = \sigma^2\, K(\mathbf{x},\mathbf{x}')$ for all $\mathbf{x}$, $\mathbf{x}'$ in $\mathscr{X}$, $\sigma^2 > 0$. Then, straightforward calculation shows that $\eta_n^*(\mathbf{x})$ is still the BLP (the posterior mean if Z_x is Gaussian), and $\sigma^2\, {\rho_n^*}^2(\mathbf{x})$ is the Mean-Squared Prediction Error (MSPE) at $\mathbf{x}$. This construction corresponds to simple kriging; see, e.g., [4, 94]. IMSPE-optimal designs minimize the integrated squared error $\mathsf{IMSPE}(\mathbf{X}_n) = \sigma^2 \int_{\mathscr{X}} {\rho_n^*}^2(\mathbf{x}) \mathrm{d}\mu(\mathbf{x})$, with μ generally taken as the uniform probability measure on $\mathscr{X}$, see, e.g., [36, 41, 83].

IMSPE-optimal designs $\mathbf{X}_n^*$ depend on the chosen K. It is well known that the asymptotic rate of decrease of $\mathsf{IMSPE}(\mathbf{X}_n^*)$ as n increases depends on the smoothness of K (the same is true for the integration problem); see for instance [82]. It is rather usual to take K stationary (translation invariant), i.e., satisfying $K(\mathbf{x},\mathbf{x}') = \Psi(\mathbf{x} - \mathbf{x}')$ for all $\mathbf{x}$ and $\mathbf{x}'$, with Ψ in some parametric class selected according to prior knowledge on the smoothness properties of f. A typical example is the Matérn class of covariances, see [90, Chap. 2]. On the other hand, for reasons explained in Section 1, computer experiments often use small values of n, and the asymptotic behavior of the approximation error is hardly observed. Its behavior on a short horizon is much more important and strongly depends on the correlation lengths in K, which are difficult to choose *a priori*. Robustness with respect to the choice of K favours space-filling designs, where the $\mathbf{x}_i$ are suitably spread over $\mathscr{X}$. Noticeably, it is shown in [85] that for translation invariant and isotropic kernels (i.e., such that $K(\mathbf{x},\mathbf{x}') = \Psi(\|\mathbf{x} - \mathbf{x}'\|)$, with $\|\cdot\|$ the Euclidean distance in $\mathbb{R}^d$), one has $\rho_n^2(\mathbf{x}) \leq S_K[h_r(\mathbf{x})]$ for some increasing function $S_K(\cdot)$. Here $h_r(\mathbf{x}) = \max_{\|\mathbf{x}-\mathbf{x}'\| \leq r} \min_{1 \leq i \leq n} \|\mathbf{x}' - \mathbf{x}_i\|$

measures the density of design points $\mathbf{x}_i$ around $\mathbf{x}$, with r a fixed positive constant. It satisfies, $\max_{\mathbf{x}\in\mathscr{X}} h_r(\mathbf{x}) \geq \max_{\mathbf{x}\in\mathscr{X}} h_0(\mathbf{x}) = \mathsf{CR}(\mathbf{X}_n)$, with

$$\mathsf{CR}(\mathbf{X}_n) = \max_{\mathbf{x}\in\mathscr{X}} \min_{1\leq i\leq n} \|\mathbf{x}-\mathbf{x}_i\|\,,$$

the *covering radius* of $\mathbf{X}_n$: $\mathsf{CR}(\mathbf{X}_n)$ defines the smallest r such that the n closed balls of radius r centred at the $\mathbf{x}_i$ cover $\mathscr{X}$. $\mathsf{CR}(\mathbf{X}_n)$ is also called the dispersion of $\mathbf{X}_n$ [69, Chap. 6] and corresponds to the minimax-distance criterion [49] used in space-filling design. Loosely speaking, the property $\rho_n^2(\mathbf{x}) \leq S_K[h_r(\mathbf{x})]$ quantifies the intuition that designs with a small value of CR provide precise predictions over $\mathscr{X}$ since for any $\mathbf{x}$ in $\mathscr{X}$ there always exists a design point $\mathbf{x}_i$ at proximity where $f(\mathbf{x}_i)$ has been evaluated. Another standard geometrical criterion of spreadness is the *packing* (or *separating*) *radius*

$$\mathsf{PR}(\mathbf{X}_n) = \frac{1}{2}\min_{i\neq j} \|\mathbf{x}_i-\mathbf{x}_j\|\,.$$

It corresponds to the largest r such that the n open balls of radius r centred at the $\mathbf{x}_i$ do not intersect; $2\,\mathsf{PR}(\cdot)$ corresponds to the maximin-distance criterion [49] often used in computer experiments. The packing radius $\mathsf{PR}(\mathbf{X}_n)$ is a simpler characteristic than the covering radius $\mathsf{CR}(\mathbf{X}_n)$, in terms of evaluation and optimization, see, e.g., [74]. Regularized versions of $\mathsf{PR}(\mathbf{X}_n)$ are well-known, see Example 3.5; regularization of $\mathsf{CR}(\mathbf{X}_n)$ is considered in [78].

In this paper, we shall adopt the following point of view. We do not intend to construct designs adapted to a particular K chosen from *a priori* knowledge on f. Neither shall we estimate the parameters in K (such as correlation lengths) when K is taken from a parametric class. We shall rather consider the kernel K as a tool for constructing a space-filling design, the quality of which will be measured in particular through the value of CR. The motivation is twofold: (*i*) the construction will be much easier than the direct minimization of CR, (*ii*) it will facilitate the construction of *sequences of points* suitably spread over $\mathscr{X}$.

2.2. Bayesian quadrature. Denote by $\mathscr{M} = \mathscr{M}[\mathscr{X}]$ the set of finite signed Borel measures on a nonempty set $\mathscr{X}$, and by $\mathscr{M}(q)$, $q\in\mathbb{R}$, the set of signed measures with total mass q: $\mathscr{M}(q) = \{\mu\in\mathscr{M} : \mu(\mathscr{X}) = q\}$. The set of Borel probability measures on $\mathscr{X}$ is denoted by $\mathscr{M}^+(1)$, $\mathscr{M}^+$ is the set of finite positive measures on $\mathscr{X}$. Typical applications correspond to $\mathscr{X}$ being a compact subset of $\mathbb{R}^d$ for some d.

Suppose we wish to integrate a real function defined on $\mathscr{X}$ with respect to $\mu\in\mathscr{M}^+(1)$. Assume that $\mathsf{E}_\mu\{|f(X)|\} < +\infty$ and denote

$$I_\mu(f) = \mathsf{E}_\mu\{f(\mathbf{X})\} = \int_{\mathscr{X}} f(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x})\,.$$

We set a prior on f, and assume that f is a realization of a Gaussian RF, with covariance $\sigma^2\,K(\cdot,\cdot)$, $\sigma^2>0$, and unknown mean β_0; that is, we consider the location model with correlated errors

$$f(\mathbf{x}) = \beta_0 + Z_x\,, \tag{2.1}$$

where $\mathbb{E}\{Z_x\}=0$ and $\mathbb{E}\{Z_xZ_{x'}\} = \sigma^2\,K(\mathbf{x},\mathbf{x}')$ for all $\mathbf{x},\mathbf{x}'\in\mathscr{X}$. Regression models more general than (2.1) are considered in Appendix B; one may refer to [16] for an

extensive review of Bayesian quadrature. Here K is a symmetric Positive Definite (PD) kernel; that is, $K(\mathbf{x},\mathbf{x}') = K(\mathbf{x}',\mathbf{x})$, and for all $n \in \mathbb{N}$ and all pairwise different $\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathscr{X}$, the matrix $\mathbf{K}_n$ is non-negative definite; if $\mathbf{K}_n$ is positive definite, then K is called Strictly Positive Definite (SPD). Note that $K^2(\mathbf{x},\mathbf{x}') \leq K(\mathbf{x},\mathbf{x})K(\mathbf{x}',\mathbf{x}') < +\infty$ for all $\mathbf{x},\mathbf{x}' \in \mathscr{X}$ since K corresponds to a covariance. We will call a general kernel K bounded when $K(\mathbf{x},\mathbf{x}) < \infty$ for all $\mathbf{x} \in \mathscr{X}$, and uniformly bounded when there is a constant C such that $K(\mathbf{x},\mathbf{x}) \leq C$ for all $\mathbf{x} \in \mathscr{X}$. Any PD kernel is bounded.

Similarly to Section 2.1, we denote by $\mathcal{H}_K$ the associated RKHS and by $\langle\cdot,\cdot\rangle_K$ the scalar product in $\mathcal{H}_K$. The assumption that K is bounded will be relaxed in Section 3.2 where we shall also consider singular kernels, but throughout the paper we assume that K is symmetric, $K(\mathbf{x},\mathbf{x}') = K(\mathbf{x}',\mathbf{x})$ for all $\mathbf{x},\mathbf{x}' \in \mathscr{X}$. Also, we always assume, as in [34, Sect. 2.1], that either K is non-negative on $\mathscr{X}\times\mathscr{X}$, or $\mathscr{X}$ is compact.

We set a vague prior on β_0 and assume that $\beta_0 \sim \mathcal{N}(\hat{\beta}_0^0, \sigma^2 A)$ with $A \to +\infty$. This amounts to setting $1/A = 0$ in all Bayesian calculations; the choice of $\hat{\beta}_0^0$ is then irrelevant. Suppose that f has been evaluated at an n-point design $\mathbf{X}_n = \{\mathbf{x}_1,\ldots,\mathbf{x}_n\} \in \mathscr{X}^n$ with pairwise different points. We assume that $\mathbf{K}_n$ has full rank. For any $\mathbf{x} \in \mathscr{X}$, the posterior distribution of $f(\mathbf{x})$ (conditional on σ^2 and K) is normal, with mean

$$\hat{\eta}_n(\mathbf{x}) = \hat{\beta}_0^n + \mathbf{k}_n^\top(\mathbf{x})\mathbf{K}_n^{-1}(\mathbf{y}_n - \hat{\beta}_0^n\mathbf{1}_n)$$

and variance (mean-squared error)

$$\sigma^2\rho_n^2(\mathbf{x}) = \sigma^2\left[K(\mathbf{x},\mathbf{x}) - \mathbf{k}_n^\top(\mathbf{x})\mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x}) + \frac{(1-\mathbf{k}_n^\top(\mathbf{x})\mathbf{K}_n^{-1}\mathbf{1}_n)^2}{\mathbf{1}_n^\top\mathbf{K}_n^{-1}\mathbf{1}_n}\right], \tag{2.2}$$

where

$$\hat{\beta}_0^n = \frac{\mathbf{1}_n^\top\mathbf{K}_n^{-1}\mathbf{y}_n}{\mathbf{1}_n^\top\mathbf{K}_n^{-1}\mathbf{1}_n} \tag{2.3}$$

and $\mathbf{1}_n$ is the n-dimensional vector $(1,\ldots,1)^\top$, see for instance [84, Chap. 4]. The posterior mean of $I_\mu(f)$ is thus

$$\widehat{I}_n = \int_{\mathscr{X}} \hat{\eta}_n(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}) = \mathsf{E}_\mu\{\hat{\eta}_n(\mathbf{X})\} = \hat{\beta}_0^n + \mathbf{p}_n(\mu)^\top\mathbf{K}_n^{-1}(\mathbf{y}_n - \hat{\beta}_0^n\mathbf{1}_n)\,, \tag{2.4}$$

with

$$\mathbf{p}_n(\mu) = (P_\mu(\mathbf{x}_1),\ldots,P_\mu(\mathbf{x}_n))^\top\,, \tag{2.5}$$

where, for any $\nu \in \mathscr{M}$ and $\mathbf{x} \in \mathscr{X}$, we denote

$$P_\nu(\mathbf{x}) = \int_{\mathscr{X}} K(\mathbf{x},\mathbf{x}')\,\mathrm{d}\nu(\mathbf{x}')\,. \tag{2.6}$$

$P_\nu(\cdot)$ is called the kernel imbedding of ν into $\mathcal{H}_K$, see [88, Def. 9]; $P_\nu(\mathbf{x})$ is well-defined and finite for any $\nu \in \mathscr{M}$ and $\mathbf{x} \in \mathscr{X}$ when K is uniformly bounded. On the other hand, there always exists $\nu \in \mathscr{M}$ such that $P_\nu(\mathbf{x})$ is infinite for all $\mathbf{x} \in \mathscr{X}$ when K is not uniformly bounded on $\mathscr{X}$. The function $P_\nu(\cdot)$ is called a potential in potential theory, see Section 3.2.

Similarly to (2.2), we obtain that the posterior variance of $I_\mu(f)$ is

$$\sigma^2 s_n^2 = \sigma^2 \left[\mathscr{E}_K(\mu) - \mathbf{p}_n^\top(\mu) \mathbf{K}_n^{-1} \mathbf{p}_n(\mu) + \frac{(1 - \mathbf{p}_n^\top(\mu) \mathbf{K}_n^{-1} \mathbf{1}_n)^2}{\mathbf{1}_n^\top \mathbf{K}_n^{-1} \mathbf{1}_n} \right] , \tag{2.7}$$

where, for any $\nu \in \mathscr{M}$, we denote

$$\mathscr{E}_K(\nu) = \int_{\mathscr{X}^2} K(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\nu(\mathbf{x}) \mathrm{d}\nu(\mathbf{x}') \, . \tag{2.8}$$

This is one of the key notions in potential theory, called the energy of ν; see Section 3.2. For μ in $\mathscr{M}^+(1)$, we have $\mathscr{E}_K(\mu) = \mathsf{E}_\mu\{K(\mathbf{X}, \mathbf{X}')\}$ where $\mathbf{X}$ and $\mathbf{X}'$ are independently identically distributed (i.i.d.) with μ. The quantity $-\mathscr{E}_K(\mu)$ corresponds to the quadratic entropy introduced by C.R. Rao [80]; see also Remark 3.1. Define

$$\mathscr{M}_K^\alpha = \left\{ \nu \in \mathscr{M} : \int_{\mathscr{X}} K^\alpha(\mathbf{x}, \mathbf{x}) \, \mathrm{d}|\nu|(\mathbf{x}) < +\infty \right\} , \ \alpha > 0 \, . \tag{2.9}$$

When $\mu \in \mathscr{M}_K^{1/2}$, the reproducing property and the Cauchy-Schwarz inequality imply that

$$\begin{aligned} \mathscr{E}_K(\mu) &= \int_{\mathscr{X}^2} \langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{x}') \rangle_K \, \mathrm{d}\mu(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}') \\ &\leq \left[\int_{\mathscr{X}} K^{1/2}(\mathbf{x}, \mathbf{x}) \, \mathrm{d}|\mu|(\mathbf{x}) \right]^2 < +\infty \, . \end{aligned} \tag{2.10}$$

When β_0 is assumed to be known (equal to zero for instance), we simply substitute β_0 for $\hat{\beta}_0^n$ in (2.4) and the posterior variance is

$$\sigma^2 s_{n,0}^2 = \sigma^2 \left[\mathscr{E}_K(\mu) - \mathbf{p}_n^\top(\mu) \mathbf{K}_n^{-1} \mathbf{p}_n(\mu) \right] . \tag{2.11}$$

Bayesian quadrature relies on the estimation of $I_\mu(f)$ by $\widehat{I}_n$. An optimal design for estimating $I_\mu(f)$ should minimize s_n^2 given by (2.7). One may refer to [24] for a historical perspective and to [45] for a recent exposition on Bayesian numerical computation. The framework presented above is similar to that considered in [72] (where an improper prior density $p(\beta_0, \sigma^2) \propto \sigma^{-2}$ is set on β_0 and σ^2), restricted to the case (recommended in that paper) where the known trend function is simply the constant 1 (which corresponds to the presence of an unknown mean β_0 in the model (2.1)). In Section 4, we shall see that $s_{n,0}^2$ is equal to the minimum value of a (squared) kernel discrepancy between the measure μ and a signed measure supported on $\mathbf{X}_n$, and that s_n^2 corresponds to the minimum of a squared discrepancy for signed measures that are constrained to have total mass one, and also corresponds to the minimum of an energy functional for a modified kernel K_μ. Note that $\sigma^2 s_n^2 \leq \mathsf{IMSPE}(\mathbf{X}_n) = \sigma^2 \int_{\mathscr{X}} \rho_n^2(\mathbf{x}) \, \mathrm{d}\mu(\mathbf{x})$ (which requires $\mu \in \mathscr{M}_K^1 \subset \mathscr{M}_K^{1/2}$ to be well-defined); we show in Theorem 4.5 that $\mathsf{IMSPE}(\mathbf{X}_n) \leq \sigma^2 s_n^2 + \sigma^2 \left[\int_{\mathscr{X}} K(\mathbf{x}, \mathbf{x}) \, \mathrm{d}\mu(\mathbf{x}) - \mathscr{E}_K(\mu) \right]$. One of the key ideas of the paper is that space-filling design may be based on the minimization of s_n^2 rather than the minimization of the IMSPE.

3. Kernel discrepancy, energy and potentials.

3.1. Maximum mean discrepancy: a metric on probability measures related to integration. Suppose that K is bounded and f belongs to the RKHS $\mathcal{H}_K$. Let μ and ν be two probability measures in $\mathscr{M}^+(1) \cap \mathscr{M}_K^{1/2}$. Since $f \in \mathcal{H}_K$, using the reproducing property, we obtain $I_\mu(f) = \int_{\mathscr{X}} \langle f, K_{\mathbf{x}} \rangle_K \, \mathrm{d}\mu(\mathbf{x})$, $I_\nu(f) = \int_{\mathscr{X}} \langle f, K_{\mathbf{x}} \rangle_K \, \mathrm{d}\nu(\mathbf{x})$ and

$$|I_\mu(f) - I_\nu(f)| = \left| \int_{\mathscr{X}} \langle f, K_{\mathbf{x}} \rangle_K \, \mathrm{d}(\mu - \nu)(\mathbf{x}) \right| = |\langle f, P_\mu - P_\nu \rangle_K| \, ,$$

with $P_\mu(\cdot)$ and $P_\nu(\cdot)$ the kernel imbeddings (2.6). Define

$$\gamma_K(\mu, \nu) = \|P_\mu - P_\nu\|_{\mathcal{H}_K} \, . \tag{3.1}$$

The Cauchy-Schwarz inequality yields the Koksma-Hlawka type inequality, see, e.g., [47], [69, Chap. 2], $|I_\mu(f) - I_\nu(f)| \leq \|f\|_{\mathcal{H}_K} \gamma_K(\mu, \nu)$, and

$$\gamma_K(\mu, \nu) = \sup_{\|f\|_{\mathcal{H}_K} = 1} |I_\mu(f) - I_\nu(f)| \, , \tag{3.2}$$

see, e.g., [89, Th. 1]. Also, the expansion of $\|P_\mu - P_\nu\|_{\mathcal{H}_K}^2$ gives

$$\begin{aligned} \gamma_K(\mu, \nu) &= \left(\|P_\mu\|_{\mathcal{H}_K}^2 + \|P_\nu\|_{\mathcal{H}_K}^2 - 2\langle P_\mu, P_\nu \rangle_K \right)^{1/2} \\ &= \left(\int_{\mathscr{X}^2} K(\mathbf{x}, \mathbf{x}') \, \mathrm{d}(\nu - \mu)(\mathbf{x}) \, \mathrm{d}(\nu - \mu)(\mathbf{x}') \right)^{1/2} . \end{aligned} \tag{3.3}$$

Therefore, $\gamma_K(\cdot, \cdot)$ is at the same time a pseudometric between kernel imbeddings (3.1) and an integral pseudometric on probability distributions (3.2). It defines a kernel discrepancy between distributions (3.3), $\gamma_K(\cdot, \cdot)$ is also called the *Maximum Mean Discrepancy* (MMD) between μ and ν in $\mathscr{M}^+(1) \cap \mathscr{M}_K^{1/2}$, see [88, Def. 10].

To define a metric on the whole $\mathscr{M}^+(1)$, we need P_μ to be well-defined and so that $P_\mu = P_\nu$ for μ and ν in $\mathscr{M}^+(1)$ implies $\mu = \nu$. This corresponds to the notion of *characteristic kernel*, see [89, Def. 6], which is closely connected to the following definitions.

Definition 3.1. *A kernel K is* Integrally Strictly Positive Definite *(ISPD) on $\mathscr{M}$ when $\mathscr{E}_K(\nu) > 0$ for any nonzero measure $\nu \in \mathscr{M}$.*

Definition 3.2. *A kernel K is* Conditionally Integrally Strictly Positive Definite *(CISPD) on $\mathscr{M}$ when it is ISPD on $\mathscr{M}(0)$; that is, when $\mathscr{E}_K(\nu) > 0$ for all nonzero signed measures $\nu \in \mathscr{M}$ such that $\nu(\mathscr{X}) = 0$.*

An ISPD kernel is CISPD. A bounded ISPD kernel is SPD and defines an RKHS. In [89, Lemma 8], the authors show that a uniformly bounded kernel is characteristic if and only if it is CISPD. The proof is a direct consequence of the expression (3.3) for the MMD $\gamma_K(\mu, \nu)$. They also give (Corollary 4) a spectral interpretation of $\gamma_K(\mu, \nu)$ and show that a translation-invariant kernel such that $K(\mathbf{x}, \mathbf{x}') = \Psi(\mathbf{x} - \mathbf{x}')$, with Ψ a uniformly bounded continuous real-valued positive-definite function, satisfies, for any μ and ν in $\mathscr{M}^+(1)$,

$$\gamma_K(\mu, \nu) = \left[\int_{\mathbb{R}^d} |\phi_\mu(\boldsymbol{\omega}) - \phi_\nu(\boldsymbol{\omega})|^2 \, \mathrm{d}\Lambda(\boldsymbol{\omega}) \right]^{1/2} .$$

Here, ϕ_μ and ϕ_ν denote the characteristic functions of μ and ν respectively and Λ is the spectral Borel measure on $\mathbb{R}^d$, defined by

$$\Psi(\mathbf{x}) = \int_{\mathbb{R}^d} e^{-i\mathbf{x}^\top \boldsymbol{\omega}} \, \mathrm{d}\Lambda(\boldsymbol{\omega}) \,. \tag{3.4}$$

Using this spectral representation, they prove (Theorem 9) that K is characteristic if and only if the support of the associated Λ coincides with $\mathbb{R}^d$. For example, the sinc-squared kernel $K(x,x') = \sin^2[\theta(x-x')]/[\theta(x-x')]^2$, $\theta > 0$, is SPD but is not characteristic (and therefore not CISPD) since the support of Λ equals $[-2\theta, 2\theta]$, and the triangular kernel $K_\theta(x,x') = \Psi_\theta(x-x') = \max\{1-\theta|x-x'|, 0\}$ is SPD and characteristic since the Fourier transform of Ψ_θ is the sinc-squared function. When $\gamma_K(\mu, \delta_\mathbf{x})$ is well-defined for all $\mathbf{x} \in \mathscr{X}$, with $\delta_\mathbf{x}$ the Dirac delta measure at $\mathbf{x}$ (and thus in particular when K is characteristic), we may consider the empirical measure $\xi_{n,e} = \xi_{n,e}(\mathbf{X}_n) = (1/n)\sum_{i=1}^n \delta_{\mathbf{x}_i}$ associated with a given design $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, and $\gamma_K(\mu, \xi_{n,e})$ of (3.2) gives the worst-case integration error for $\xi_{n,e}$ when f has norm one in $\mathcal{H}_K$; see Section 4.3.1.

Typical examples of uniformly bounded ISPD, and therefore characteristic, kernels are the squared exponential kernel $K_t(\mathbf{x},\mathbf{x}') = \exp(-t\,\|\mathbf{x}-\mathbf{x}'\|^2)$, $t > 0$, and the isotropic Matérn kernels, in particular

$$K_{3/2,\theta}(\mathbf{x},\mathbf{x}') = (1+\sqrt{3}\theta\,\|\mathbf{x}-\mathbf{x}'\|)\,\exp(-\sqrt{3}\theta\,\|\mathbf{x}-\mathbf{x}'\|) \quad \text{(Matérn 3/2)}\,, \tag{3.5}$$

and $K_{5/2,\theta}(\mathbf{x},\mathbf{x}') = [1+\sqrt{5}\theta\,\|\mathbf{x}-\mathbf{x}'\| + 5\theta^2\,\|\mathbf{x}-\mathbf{x}'\|^2/3]\,\exp(-\sqrt{5}\theta\,\|\mathbf{x}-\mathbf{x}'\|)$ (Matérn 5/2), see, e.g., [90]. (They are SPD for any d, see [40], and ISPD since the spectral measure Λ in (3.4) is strictly positive on $\mathbb{R}^d$.) Two other important examples are given hereafter.

Example 3.1 (generalized multiquadric kernel)**.** The sum of ISPD kernels is ISPD. Since the squared exponential kernel $K_t(\mathbf{x},\mathbf{x}')$ is ISPD for any $t > 0$, the integrated kernel obtained by setting a probability distribution on t is ISPD too. One may thus consider $K(\mathbf{x},\mathbf{x}') = \int K_t(\mathbf{x},\mathbf{x}')\,\mathrm{d}\pi(t)$ for π bounded and non decreasing on $[0,+\infty)$, which generates the class of continuous isotropic autocovariance functions in arbitrary dimension, see [87] and [90, p. 44]. In particular, for any $\epsilon > 0$ and $s > 0$, we obtain

$$K(\mathbf{x},\mathbf{x}') = \int_0^{+\infty} K_t(\mathbf{x},\mathbf{x}')\,t^{s/2-1}\,\exp(-\epsilon\,t)\,\mathrm{d}t = \frac{\Gamma(s/2)}{(\|\mathbf{x}-\mathbf{x}'\|^2+\epsilon)^{s/2}}\,,$$

showing that the generalized multiquadric kernel

$$K_{s,\epsilon}(\mathbf{x},\mathbf{x}') = (\|\mathbf{x}-\mathbf{x}'\|^2+\epsilon)^{-s/2}\,,\ \epsilon > 0\,,\ s > 0\,, \tag{3.6}$$

is ISPD, see also [89, Sect. 3.2]. ◁

Example 3.2 (distance-induced kernels)**.** Consider the kernels defined by

$$K^{(s)}(\mathbf{x},\mathbf{x}') = -\|\mathbf{x}-\mathbf{x}'\|^s\,,\ s > 0\,, \tag{3.7}$$

which are CISPD for $s \in (0,2)$ [93], and the related distance-induced kernels

$$K'^{(s)}(\mathbf{x},\mathbf{x}') = \|\mathbf{x}\|^s + \|\mathbf{x}'\|^s - \|\mathbf{x}-\mathbf{x}'\|^s\,,\ s > 0\,.$$

Note that $\mathscr{E}_{K'^{(s)}}(\mu) = \mathscr{E}_{K^{(s)}}(\mu)$ when $\mu(\mathscr{X}) = 0$; in [93] $\mathscr{E}_{K'^{(s)}}$ is called energy distance for $s = 1$ and generalized energy distance for general $s \in (0,2]$. For $s > 0$,

the set $\mathscr{M}_{K'^{(s)}}$ contains all signed measures μ such that $\int_{\mathscr{X}} \|\mathbf{x}-\mathbf{x}_0\|^s \,\mathrm{d}|\mu|(\mathbf{x}) < +\infty$ for some $\mathbf{x}_0 \in \mathscr{X}$. This result is a direct consequence of the triangular inequality when $s \in (0,1]$; for $s > 1$ it follows from considerations involving semimetrics generated by kernels, see [88, Remark 21]. $K'^{(s)}$ is CISPD for $s \in (0,2)$ ($K'^{(s)}/2$ corresponds to the covariance function of the fractional Brownian motion), but is not SPD (one has in particular, $K'^{(s)}(\mathbf{0},\mathbf{0}) = 0$); $K'^{(2)}$ is not CISPD since $\mathscr{E}_{K'^{(2)}}(\mu) = [\int_{\mathscr{X}} \mathbf{x}^\top \,\mathrm{d}\mu(\mathbf{x})][\int_{\mathscr{X}} \mathbf{x}\,\mathrm{d}\mu(\mathbf{x})]$, $\mu \in \mathscr{M}$. $K(x,x') = 1 - K^{(1)}(x,x') = 1 - |x-x'|$ is ISPD for $\mathscr{X} = [0,1]$. ◁

3.2. Energy and potentials, MMD for signed measures and singular kernels. In this section we extend the considerations of the previous section to signed measures and kernels which may have singularity on the diagonal. The expression (3.9) shows that the squared MMD between two measures μ and ν is the energy $\mathscr{E}_K$ of the signed measure $\nu - \mu$, hence the importance of considering signed measures besides probability measures. We believe that singular kernels have great potential interest for the construction of space-filling designs, due to their natural repelling property.

Definitions 3.1 and 3.2 extend to singular kernels, with Riesz kernels as typical examples.

Example 3.3 (Riesz kernels)**.** These fundamental kernels of potential theory are defined by

$$K_{(s)}(\mathbf{x},\mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\|^{-s}\,,\ s>0\,,\ \text{and}\ K_{(0)}(\mathbf{x},\mathbf{x}') = -\log\|\mathbf{x}-\mathbf{x}'\|\,, \tag{3.8}$$

with $\mathbf{x},\mathbf{x}' \in \mathscr{X} \subset \mathbb{R}^d$ and $\|\cdot\|$ the Euclidean norm. When $s \geq d$, the energy $\mathscr{E}_{K_{(s)}}(\mu)$ is infinite for any nonzero signed measure, but for $s \in (0,d)$ the kernel $K_{(s)}$ is ISPD. Since the logarithmic kernel $K_{(0)}(\mathbf{x},\mathbf{x}')$ has a singularity at zero and tends to $-\infty$ when $\|\mathbf{x}-\mathbf{x}'\|$ tends to $+\infty$, it will only be considered for $\mathscr{X}$ compact; $K_{(0)}$ is CISPD, see [56, p. 80]. ◁

Consider again $\mathscr{E}_K(\mu)$ given by (2.8), with $\mu \in \mathscr{M}$. In potential theory, this quantity is called the energy of the signed measure μ for the kernel K. Denote

$$\mathscr{M}_K = \{\nu \in \mathscr{M} : |\mathscr{E}_K(\nu)| < +\infty\}\,.$$

In the following, we shall only consider kernels that are at least CISPD. When K is ISPD, $\mathscr{E}_K(\mu)$ is positive for any nonzero $\mu \in \mathscr{M}$, but when K is only CISPD, $\mathscr{E}_K(\mu)$ can be negative; this is the reason for the presence of the absolute value in the definition of $\mathscr{M}_K$. Note that $\mathscr{M}_K$ is the set of measures such that $\mathscr{E}_K(\mu^+)$, $\mathscr{E}_K(\mu^-)$ and $\mathscr{E}_K(\mu^+,\mu^-) = \int_{\mathscr{X}^2} K(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mu^+(\mathbf{x})\mathrm{d}\mu^-(\mathbf{x}')$ are all finite, with μ^+ and μ^- denoting the positive and negative parts of the Hahn-Jordan decomposition $\mu = \mu^+ - \mu^-$ of μ, see [34, Sect. 2.1]. Also note that when K is bounded and defines an RKHS, $\mathscr{M}_K^\alpha \subset \mathscr{M}_K$ for any $\alpha \geq 1/2$, see (2.9) and (2.10); when K is uniformly bounded, $\mathscr{M}_K = \mathscr{M}$.

For any $\mu \in \mathscr{M}_K$, $P_\mu(\mathbf{x})$ given by (2.6) is called the potential at $\mathbf{x}$ associated with $\mathscr{E}_K(\mu)$. It is well-defined, with values in $\mathbb{R} \cup \{-\infty,+\infty\}$, when $P_{\mu^+}(\mathbf{x})$ and $P_{\mu^-}(\mathbf{x})$ are not both infinite. Also, $P_\mu(\mathbf{x})$ is finite for μ-almost any $\mathbf{x}$, even if K is singular, when $\mu \in \mathscr{M}^+(1) \cap \mathscr{M}_K^{1/2}$.

When K is ISPD, we can still define MMD through (3.3),

$$\gamma_K(\mu,\nu) = \mathscr{E}_K^{1/2}(\nu-\mu)\,, \tag{3.9}$$

since $\mathscr{E}_K(\nu-\mu)$ is nonnegative whenever defined. The set $\mathscr{M}_K$ forms a pre-Hilbert space, with scalar product the mutual energy $\mathscr{E}_K(\mu,\nu) = \int_{\mathscr{X}^2} K(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mu(\mathbf{x})\mathrm{d}\nu(\mathbf{x}')$

and norm $\mathscr{E}_K^{1/2}(\mu)$. Denote by $\mathscr{P}_K$ the linear space of potential fields $P_\mu(\cdot)$, $\mu \in \mathscr{M}_K$; when K defines an RKHS $\mathcal{H}_K$, $\|P_\mu\|_{\mathcal{H}_K} = \mathscr{E}_K^{1/2}(\mu)$, so that $\mathscr{P}_K \subset \mathcal{H}_K$, and $\mathscr{P}_K$ is dense in $\mathcal{H}_K$. For $\mathscr{P}_K$ to contain all functions $K_{\mathbf{x}}(\cdot) = K(\cdot, \mathbf{x})$, $\mathbf{x} \in \mathscr{X}$, we need $\delta_{\mathbf{x}} \in \mathscr{M}_K$ for all $\mathbf{x}$, which requires $K(\mathbf{x}, \mathbf{x}) < \infty$ for all $\mathbf{x} \in \mathscr{X}$.

For $\mu, \nu \in \mathscr{M}_K$, $\mathscr{E}_K(\mu, \nu)$ defines a scalar product $\langle P_\mu, P_\nu \rangle_{\mathscr{P}_K}$ on $\mathscr{P}_K$, with $\gamma_K(\mu, \nu) = \|P_\mu - P_\nu\|_{\mathscr{P}_K}$. Similarly to Section 3.1, we obtain

$$\begin{aligned} \gamma_K(\mu, \nu) &= \sup_{\xi \in \mathscr{M}_K,\, \mathscr{E}_K(\xi)=1} \left| \int_{\mathscr{X}^2} K(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\xi(\mathbf{x}) \mathrm{d}(\mu - \nu)(\mathbf{x}') \right| \\ &= \sup_{\|h\|_{\mathscr{P}_K} \le 1} |I_\mu(h) - I_\nu(h)| \; ; \end{aligned} \tag{3.10}$$

that is, a result that extends (3.2) to general ISPD kernels. If K is only CISPD, we can also define $\gamma_K(\mu, \nu)$ in the same way when considering measures $\mu, \nu \in \mathscr{M}(1)$; we then define $\mathscr{P}_K$ as the linear space of potential fields $P_\mu(\cdot)$, $\mu \in \mathscr{M}_K \cap \mathscr{M}(0)$, and in (3.10) we restrict ξ to be in $\mathscr{M}(0)$.

When K is singular, there always exists ν in $\mathscr{M}_K$ such that $P_\nu(\mathbf{x}_0) = +\infty$ for some $\mathbf{x}_0$. Consider for example the Riesz kernel $K_{(s)}(\mathbf{x}, \mathbf{x}')$ with $s \in (0, d)$; $\mathscr{M}_K$ contains in particular all signed measures with compact support $\mathbb{S}(\mu)$ whose potential $P_\mu(\mathbf{x})$ is bounded on $\mathbb{S}(\mu)$, see [56, p. 81]. Take ν as the measure with density $c/\|\mathbf{x} - \mathbf{x}_0\|^{s-d}$ on $\mathscr{X}$, with $\mathbf{x}_0 \in \mathscr{X}$; we have $\mathscr{E}_{K_{(s)}}(\nu) < \infty$ for $\mathscr{X}$ compact, but $P_\nu(\mathbf{x}_0) = +\infty$. As a consequence, as noted in [21], singular kernels have little interest for integration. Indeed, take $\mu, \nu \in \mathscr{M}_K$ and $h = P_\nu \in \mathscr{P}_K$, then $|I_\mu(h)| \le \|h\|_{\mathscr{P}_K} \mathscr{E}_K^{1/2}(\mu) = \mathscr{E}_K^{1/2}(\nu)\mathscr{E}_K^{1/2}(\mu) < \infty$, whereas $|I_{\xi_n}(h)|$ may be infinite for some discrete approximation ξ_n of μ as h can be infinite at some points. Singular kernels may nevertheless be used for the construction of space-filling designs, see for instance the example in Section 5.3, and this is our motivation for considering them in the following.

The key difficulty with singular kernels is the fact that delta measures do not belong to $\mathscr{M}_K$. An expedient solution to circumvent the problem is to replace a singular kernel with a bounded surrogate. For instance, in space-filling design we may replace the Riesz kernel $K_{(s)}$, $s > 0$, by a generalized inverse multiquadric kernel $K_{s,\epsilon}$ given by (3.6), and consider the limiting behavior of the designs obtained when $\epsilon \to 0$, see Section 4.3.1; see also [79] for other constructions.

3.3. Minimum energy and equilibrium measures. In this section, we show that there exist strong connections between results in potential theory and optimal design theory, where one minimizes a convex functional of $\mu \in \mathscr{M}^+(1)$, with the particularity that here the functional is quadratic. This will be exploited in particular in Section 4.4 for the construction of nested design sequences.

3.3.1. ISPD kernels and convexity of $\mathscr{E}_K(\cdot)$.

Lemma 3.1. *K is ISPD if and only if $\mathscr{M}_K$ is convex and $\mathscr{E}_K(\cdot)$ is strictly convex on $\mathscr{M}_K$.*

Proof. For any K, any μ and ν in $\mathscr{M}_K$ and any $\alpha \in [0, 1]$, direct calculation gives

$$(1-\alpha)\,\mathscr{E}_K(\mu) + \alpha\,\mathscr{E}_K(\nu) - \mathscr{E}_K[(1-\alpha)\mu + \alpha\nu] = \alpha(1-\alpha)\,\mathscr{E}_K(\nu - \mu)\,. \tag{3.11}$$

Assume that K is ISPD. For any μ and ν in $\mathscr{M}_K$, the mutual energy $\mathscr{E}_K(\mu, \nu)$ satisfies $|\mathscr{E}_K(\mu, \nu)| \le \sqrt{\mathscr{E}_K(\mu)\mathscr{E}_K(\nu)} < +\infty$. Therefore, $\mathscr{E}_K(\mu - \nu) = \mathscr{E}_K(\mu) + \mathscr{E}_K(\nu) - 2\,\mathscr{E}_K(\mu, \nu)$ is finite and (3.11) implies that $\mathscr{E}_K[(1-\alpha)\mu + \alpha\nu]$ is finite, showing that

$\mathscr{M}_K$ is convex. Since K is ISPD, $\mathscr{E}_K(\nu-\mu) > 0$ for $\mu, \nu \in \mathscr{M}$, $\nu \neq \mu$, and (3.11) implies that $\mathscr{E}_K(\cdot)$ is strictly convex on $\mathscr{M}_K$.

Conversely, assume that $\mathscr{M}_K$ is convex and $\mathscr{E}_K(\cdot)$ is strictly convex on $\mathscr{M}_K$. Any $\xi \in \mathscr{M}_K$ can be written as $\xi = \nu - \mu$ with, for instance, $\nu = 2\,\xi$ and $\mu = \xi$, both in $\mathscr{M}_K$. If $\mathscr{E}_K(\cdot)$ is strictly convex on $\mathscr{M}_K$, (3.11) with $\alpha \in (0,1)$ implies that $\mathscr{E}_K(\xi) > 0$ when $\nu \neq \mu$, that is, when $\xi \neq 0$. Therefore, K is ISPD. ■

Lemma 3.1 also applies to singular kernels. The lemma below concerns CISPD kernels, which are assumed to be uniformly bounded.

Lemma 3.2. *Assume that K is uniformly bounded. Then, K is CISPD if and only if $\mathscr{E}_K(\cdot)$ is strictly convex on $\mathscr{M}(1)$.*

Proof. Since K is uniformly bounded, $\mathscr{M}_K = \mathscr{M}$. Assume that K is CISPD. Then, $\mathscr{E}_K(\nu-\mu) > 0$ for any $\mu \neq \nu \in \mathscr{M}(1)$, and (3.11) implies that $\mathscr{E}_K(\cdot)$ is strictly convex on $\mathscr{M}(1)$.

Assume now that $\mathscr{E}_K(\cdot)$ is strictly convex on $\mathscr{M}(1)$. Take any non-zero signed measure ξ in $\mathscr{M}(0)$ and consider the Hahn-Jordan decomposition $\xi = \xi^+ - \xi^-$, with $\xi^+(\mathscr{X}) = \xi^-(\mathscr{X}) = c > 0$. Denote $\nu = \xi^+/c$, $\mu = \xi^-/c$, with ν and μ in $\mathscr{M}^+(1)$ (ν and μ are in $\mathscr{M}_K$ since K is uniformly bounded). Then, for any $\alpha \in (0,1)$, (3.11) and the strict convexity of $\mathscr{E}_K(\cdot)$ on $\mathscr{M}(1)$ gives $\mathscr{E}_K(\xi) = c^2\,\mathscr{E}_K(\nu-\mu) > 0$. ■

Note that one may replace $\mathscr{M}(1)$ by $\mathscr{M}^+(1)$, or by any $\mathscr{M}(\gamma)$ with $\gamma \neq 0$, in Lemma 3.2.

3.3.2. Minimum-energy probability measures. In the remaining part of Section 3.3, we assume that K is such that $\mathscr{E}_K(\cdot)$ is strictly convex on $\mathscr{M}^+(1) \cap \mathscr{M}_K$ and $\mathscr{M}(1) \cap \mathscr{M}_K$, which is true under the conditions of Lemma 3.1 or Lemma 3.2.

For $\mu, \nu \in \mathscr{M}_K$, denote by $F_K(\mu;\nu)$ the directional derivative of $\mathscr{E}_K(\cdot)$ at μ in the direction ν,

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

Straightforward calculation gives

$$F_K(\mu;\nu) = 2\left[\int_{\mathscr{X}^2} K(\mathbf{x},\mathbf{x}')\,\mathrm{d}\nu(\mathbf{x})\mathrm{d}\mu(\mathbf{x}') - \mathscr{E}_K(\mu)\right]\,. \tag{3.12}$$

In particular, for any $\mathbf{x} \in \mathscr{X}$, the potential $P_\mu(\mathbf{x})$ associated with μ at $\mathbf{x}$ satisfies

$$P_\mu(\mathbf{x}) = \frac{1}{2}\,F_K(\mu;\delta_\mathbf{x}) + \mathscr{E}_K(\mu)\,.$$

Remark 3.1 (Bregman divergence and Jensen difference)**.** The strict convexity of $\mathscr{E}_K(\cdot)$ implies that $\mathscr{E}_K(\nu) \geq \mathscr{E}_K(\mu) + F_K(\mu,\nu)$ for any $\mu, \nu \in \mathscr{M}_K$, with equality if and only if $\nu = \mu$. This can be used to define a Bregman divergence between measures in $\mathscr{M}_K$ (and thus between probability measures in $\mathscr{M}^+(1) \cap \mathscr{M}_K$), as

$$B_K(\mu,\nu) = \mathscr{E}_K(\nu) - [\mathscr{E}_K(\mu) + F_K(\mu,\nu)]\,;$$

see [81]. Direct calculation gives $B_K(\mu,\nu) = \mathscr{E}_K(\nu-\mu)$ (with therefore $B_K(\mu,\nu) = B_K(\nu,\mu)$), providing another interpretation for the MMD $\gamma_K(\mu,\nu)$, see (3.9).

The squared MMD is also proportional to the dissimilarity coefficient, or Jensen difference, $\Delta_J(\mu,\nu) = (1/2)[\mathscr{E}_K(\mu) + \mathscr{E}_K(\nu)] - \mathscr{E}_K[(\mu+\nu)/2]$ of [80]; indeed, direct calculation gives $\gamma_K^2(\mu,\nu) = \mathscr{E}_K(\nu-\mu) = 4\,\Delta_J(\mu,\nu)$. ◁

Assume that $\mathscr{X}$ is compact. Since $\mathscr{E}_K(\cdot)$ is strictly convex on $\mathscr{M}^+(1)$, there exists a unique minimum-energy probability measure. The measure $\mu_K^+ \in \mathscr{M}^+(1)$ is the minimum-energy measure if and only if $F_K(\mu_K^+;\nu) \geq 0$ for all $\nu \in \mathscr{M}^+(1)$, or equivalently, since ν is a probability measure, if and only if $F_K(\mu_K^+;\delta_{\mathbf{x}}) \geq 0$ for all $\mathbf{x} \in \mathscr{X}$. We thus obtain the following property, called equivalence theorem in the optimal design literature.

Theorem 3.1. *When $\mathscr{E}_K(\cdot)$ is strictly convex on $\mathscr{M}^+(1) \cap \mathscr{M}_K$, $\mu_K^+ \in \mathscr{M}^+(1)$ is the minimum-energy probability measure on $\mathscr{X}$ if and only if*

$$\forall\, \mathbf{x} \in \mathscr{X}\,,\ P_{\mu_K^+}(\mathbf{x}) \geq \mathscr{E}_K(\mu_K^+)\,.$$

Note that, by construction, $\int_{\mathscr{X}} P_{\mu_K^+}(\mathbf{x})\,\mathrm{d}\mu_K^+(\mathbf{x}) = \mathscr{E}_K(\mu_K^+)$, implying $P_{\mu_K^+}(\mathbf{x}) = \mathscr{E}_K(\mu_K^+)$ on the support of μ_K^+. The quantity $C_K^+ = [\inf_{\mu\in\mathscr{M}^+(1)} \mathscr{E}_K(\mu)]^{-1}$, with K an ISPD kernel, is called the *capacity* of $\mathscr{X}$ in potential theory; note that $C_K^+ \geq 0$. The minimizing measure $\mu_K^+ \in \mathscr{M}^+(1)$ is called the *equilibrium measure* of $\mathscr{X}$ (μ_K^+ is sometimes renormalized into $C_K^+\,\mu_K^+$, see [56, p. 138]). Theorem 3.1 thus gives a necessary and sufficient condition for a probability measure μ to be the equilibrium measure of $\mathscr{X}$.

Example 3.4 (continuation of Example 3.2)**.** Properties of minimum-energy probability measures $\mu^+ = \mu_{K^{(s)}}^+$ for $K^{(s)}$ given by (3.7) with $\mathscr{X}$ a compact subset of $\mathbb{R}^d$, $d \geq 2$, are investigated in [11] and [76]. The mass of μ^+ is concentrated on the boundary of $\mathscr{X}$, and its support only comprises extreme points of the convex hull of $\mathscr{X}$ when $s > 1$; for $0 < s < 2$, μ^+ is unique; it is supported on no more than $d+1$ points when $s > 2$.

Take $\mathscr{X} = \mathscr{B}_d(\mathbf{0},1)$, the closed unit ball in $\mathbb{R}^d$. For symmetry reasons, μ^+ for $0 < s < 2$ is uniform on the unit sphere $\mathscr{S}_d(\mathbf{0},1)$ and

$$\mathscr{E}_{K^{(s)}}(\mu^+) = -\int_{\mathscr{X}^2} \|\mathbf{x}-\mathbf{x}'\|^s\,\mathrm{d}\mu^+(\mathbf{x})\mathrm{d}\mu^+(\mathbf{x}') = -\int_{\mathscr{X}} \|\mathbf{x}_0-\mathbf{x}'\|^s\,\mathrm{d}\mu^+(\mathbf{x}')\,,$$

where $\mathbf{x}_0 = (1,0,\ldots,0)^\top$. Denote by $\psi_d(\cdot)$ the density of the first component $t = x_1'$ of $\mathbf{x}' = (x_1',\ldots,x_d')^\top$. We obtain $\psi_d(t) = (d-1)\,V_{d-1}\,(1-t^2)^{(d-3)/2}/(d\,V_d)$, with $V_d = \pi^{d/2}/\Gamma(d/2+1)$ the volume of $\mathscr{B}_d(\mathbf{0},1)$, and

$$\mathscr{E}_{K^{(s)}}(\mu^+) = -\int_{-1}^{1} [(1-t)^2+1-t^2]^{s/2}\,\psi_d(t)\mathrm{d}t = -\frac{2^{d-q-2}\Gamma(d/2)\Gamma[(d+s-1)/2]}{\sqrt{\pi}\Gamma(d+s/2-1)}\,.$$

In particular, $\mathscr{E}_{K^{(1)}}(\mu^+) = -4/\pi$ when $d = 2$ and is a decreasing function of d. When $s = 2$, the uniform distribution on the unit sphere is also optimal, and the minimum energy equals -2 for all $d \geq 1$, but μ^+ is not unique and the measure allocating equal weight $1/(d+1)$ at each of the $d+1$ vertices of a d regular simplex with vertices on the unit sphere is optimal too. ◁

Example 3.5 (continuation of Example 3.3)**.** Consider Riesz kernels $K_{(s)}$, see (3.8), for $\mathscr{X} = \mathscr{B}_d(\mathbf{0},1)$. When $s \geq d$, $\mathscr{E}_{K_{(s)}}(\nu)$ is infinite for any non-zero $\nu \in \mathscr{M}$, but for $0 < s < d$ there exists a minimum-energy probability measure $\mu^+ = \mu_{K_{(s)}}^+$. When $d > 2$ and $s \in (0, d-2]$, μ^+ is uniform on the unit sphere $\mathscr{S}_d(\mathbf{0},1)$ (the boundary of $\mathscr{X}$); the potential at all interior points satisfies $P_{\mu^+}(\mathbf{x}) \geq \mathscr{E}_{K_{(s)}}(\mu^+)$ with strict inequality when $s \in (0, d-2)$. When $s \in (d-2, d)$, μ^+ has a density $\varphi_s(\cdot)$ in $\mathscr{B}_d(\mathbf{0},1)$,

$$\varphi_s(\mathbf{x}) = \frac{\pi^{-d/2}\,\Gamma(1+s/2)}{\Gamma[1-(d-s)/2]}\ \frac{1}{(1-\|\mathbf{x}\|^2)^{(d-s)/2}}\,,$$

and the potential $P_{\mu^+}(\cdot)$ is constant in $\mathscr{B}_d(\mathbf{0},1)$, see, e.g., [56, p. 163].

When $d \leq 2$ and $s = 0$, μ^+ has a density in $\mathscr{B}_2(\mathbf{0},1)$ and $P_{\mu^+}(\cdot) = \mathscr{E}_{K_{(0)}}(\mu^+)$ in $\mathscr{B}_2(\mathbf{0},1)$. In particular, for $d = 1$, μ^+ has the arcsine density $1/(\pi\sqrt{1-x^2})$ in $[-1,1]$ with potential $P_{\mu^+}(x) = \log(2) - \log(|x| + \sqrt{x^2-1}|)$, $x \in \mathbb{R}$ (and $P_{\mu^+}(x) = \log(2)$ for $x \in [-1,1]$).

The energy $\mathscr{E}_{K_{(s)}}$ is infinite for empirical measures associated with n-point designs $\mathbf{X}_n$. One may nevertheless consider the "physical" energy

$$\widetilde{\mathscr{E}}_{K_{(s)}}(\mathbf{X}_n) = [2/n(n-1)] \sum_{1 \leq i < j \leq n} \|\mathbf{x}_i - \mathbf{x}_j\|^{-s} \tag{3.13}$$

($\widetilde{\mathscr{E}}_{K_{(s)}}(\mathbf{X}_n) = -[2/n(n-1)] \sum_{1 \leq i < j \leq n} \log \|\mathbf{x}_i - \mathbf{x}_j\|$ when $s = 0$), which is finite provided that all $\mathbf{x}_i$ are distinct, see [21]. An n-point set $\mathbf{X}_n^*$ minimizing $\widetilde{\mathscr{E}}_{K_{(s)}}(\mathbf{X}_n)$ is called a set of Fekete points, and the limit $\lim_{n\to\infty} \widetilde{\mathscr{E}}^{-1}_{K_{(s)}}(\mathbf{X}_n^*)$ exists and is called the transfinite diameter of $\mathscr{X}$. For large s, $\widetilde{\mathscr{E}}^{-1/s}_{K_{(s)}}(\mathbf{X}_n)$ can be considered as a regularized version of the packing radius $\mathsf{PR}(\mathbf{X}_n)$, see [76]. A major result in potential theory, see, e.g., [44], is that the transfinite diameter coincides with the capacity $C^+_{K_{(s)}}$ of $\mathscr{X}$. If $C^+_{K_{(s)}} > 0$, then $\mu^+_{K_{(s)}}$ is the weak limit of a sequence of empirical probability measures associated with Fekete points in $\mathbf{X}_n^*$. In the example considered, $\widetilde{\mathscr{E}}_{K_{(s)}}(\mathbf{X}_n^*)$ tends to infinity when $s \geq d$, but any sequence of Fekete points is asymptotically uniformly distributed in $\mathscr{X}$; $\widetilde{\mathscr{E}}_{K_{(s)}}(\mathbf{X}_n^*)$ grows like $n^{s/d-1}$ for $s > d$ (and like $\log n$ for $s = d$). ◁

Remark 3.2 (Stein variational gradient descent and energy minimization)**.** Variational inference using smooth transform based on kernelized Stein discrepancy provides a gradient descent method for the approximation of a target distribution; see [60] and the references therein; see also [22] for a Newton variational method. The fact that the construction does not require knowledge of the normalizing constant of the target distribution makes the method particularly attractive for approximating a posterior distribution in Bayesian inference. Direct calculation shows that when the kernel is translation invariant and the target distribution is uniform, then Stein variational gradient corresponds to steepest descent for the minimization of the energy $\mathscr{E}_K(\xi_{n,e})$ of the empirical measure $\xi_{n,e} = (1/n) \sum_{i=1}^n \delta_{\mathbf{x}_i}$; that is, at iteration k each design point $\mathbf{x}_i^{(k)}$ is updated into

$$\mathbf{x}_i^{(k+1)} = \mathbf{x}_i^{(k)} + \gamma \sum_{i<j} \frac{\partial K(\mathbf{x}, \mathbf{x}_j^{(k)})}{\partial \mathbf{x}}\bigg|_{\mathbf{x}=\mathbf{x}_i^{(k)}}$$

for some $\gamma > 0$. The construction of space-filling design through energy minimization has already been considered in the literature; see, e.g., [50]. In particular, it is suggested in [3] to construct designs in a compact subset $\mathscr{X}$ of $\mathbb{R}^d$ by minimizing $\widetilde{\mathscr{E}}_{K_{(2)}}(\mathbf{X}_n)$ given by (3.13) (note that for $d \geq 3$ design points constructed in this way are not asymptotically uniformly distributed in $\mathscr{X}$). This approach tends to push points to the border of $\mathscr{X}$, similarly to the maximization of the packing radius $\mathsf{PR}(\mathbf{X}_n)$. ◁

3.3.3. Minimum-energy signed measures. From (3.9), the squared MMD is the energy of a signed measure. Also, even if μ is a probability measure, the measure $\nu \in \mathscr{M}(1)$, with fixed support different from that of μ, that minimizes $\gamma_K(\mu,\nu)$, is not necessarily a probability measure. Hence the importance of considering energy minimization for *signed measures* and not only probability measures.

The situation is slightly different from that in the previous section when we consider measures in $\mathscr{M}(1)$. In that case, μ_K^* is the minimum-energy measure in $\mathscr{M}(1)$ if and only if $F_K(\mu_K^*;\nu) = 0$ for all $\nu \in \mathscr{M}(1)$, this condition being equivalent to $F_K(\mu_K^*;\delta_{\mathbf{x}}) = 0$ for all $\mathbf{x} \in \mathscr{X}$. We thus obtain the following property.

Theorem 3.2. *When $\mathscr{E}_K(\cdot)$ is strictly convex on $\mathscr{M}(1) \cap \mathscr{M}_K$, $\mu_K^* \in \mathscr{M}(1)$ is the minimum-energy signed measure with total mass one on $\mathscr{X}$ if and only if*

$$\forall\, \mathbf{x} \in \mathscr{X}\ ,\ P_{\mu_K^*}(\mathbf{x}) = \mathscr{E}_K(\mu_K^*)\,. \tag{3.14}$$

If we define now a *signed equilibrium measure* on $\mathscr{X}$ as a measure $\mu \in \mathscr{M}(1)$ such that $P_\mu(\mathbf{x})$ is constant on $\mathscr{X}$, from the definition of $P_\mu(\cdot)$, when such a measure exists it necessarily satisfies the condition of Theorem 3.2 and therefore coincides with μ_K^*. Similarly to the case where one considers probability measures in $\mathscr{M}^+(1)$, we can define the (generalized) capacity of $\mathscr{X}$ for measures in $\mathscr{M}(1)$ as $C_K^* = [\inf_{\mu\in\mathscr{M}(1)} \mathscr{E}_K(\mu)]^{-1}$, with $C_K^* = 1/\mathscr{E}_K(\mu_K^*)$ when μ_K^* exists, see [21, p. 824] (note that C_K^* may be negative). However, μ_K^* may not exist even if $\mathscr{X}$ is compact. Notice in particular that $\mathscr{M}(1)$ is not vaguely compact, contrarily to $\mathscr{M}^+(1)$ (and for Riesz kernels (3.8) with $s < d-1$, $\mathscr{M}_{K_{(s)}}$ is not complete contrarily to $\mathscr{M}_{K_{(s)}} \cap \mathscr{M}^+$ [56, Th. 1.19]).

Example 3.6 (continuation of Examples 3.2 and 3.4)**.** Take $K(x,x') = K^{(s)}(x,x') = -|x-x'|^s$ on $\mathscr{X} = [0,1]$, $s \in (0,2)$, see (3.7). K is CISPD, and there exists a unique minimum-energy probability measure $\mu^+ = \mu^+_{K^{(s)}}$ in $\mathscr{M}^+(1)$. On the other hand, below we show that minimum-energy signed measures in $\mathscr{M}(1)$ do not belong to $\mathscr{M}^+(1)$ when $s \in (1,2)$ and that there is no minimum-energy signed measure in $\mathscr{M}(1)$ when $s \geq 2$.

When $s \in (0,1)$, μ^+ has a density $\varphi^{(s)}(\cdot)$ with respect to the Lebesgue measure on $[0,1]$,

$$\varphi^{(s)}(x) = \frac{\Gamma[1-s/2]}{2^s\,\sqrt{\pi}\,\Gamma[(1-s)/2]}\ \frac{1}{[x(1-x)]^{(1+s)/2}}\,,$$

and $P_{\mu^+}(x) = \mathscr{E}(\mu^+) = -\sqrt{\pi}\,\Gamma(1-s/2)/\{2^s\,\Gamma[(1-s)/2]\,\cos(\pi s/2)\}$ for all $x \in \mathscr{X}$ (and $\mathscr{E}(\mu^+) \to -1/2$ as $s \to 1^-$). The fact that $P_{\mu^+}(x) = \mathscr{E}(\mu^+)$ for all $x \in \mathscr{X}$ indicates that μ^+ is the minimum-energy signed measure with total mass one when $s \in (0,1)$.

When $s \in [1,2)$, $\mu^+ = (\delta_0 + \delta_1)/2$; the associated potential is $P_{\mu^+}(x) = -(|x|^s + |1-x|^s)/2 \geq \mathscr{E}(\mu^+) = -1/2$, $x \in \mathscr{X}$ (note that $P_{\mu^+}(x) = -1/2$ for all $x \in \mathscr{X}$ when $s = 1$).

Consider now the signed measure $\mu_w = [(1+w)/2](\delta_0 + \delta_1) - w\delta_{1/2}$, $w > 0$, so that $\mu_w(\mathscr{X}) = 1$ (i.e., $\mu_w \in \mathscr{M}(1)$). Direct calculation gives $\mathscr{E}_{K^{(s)}}(\mu_w) = -(1+w)(1+w-2^{2-s}w)$, which is minimum for $w = w_*(s) = (1-2^{1-s})/(2^{2-s}-1)$ when $s < 2$, with $\mathscr{E}_{K^{(s)}}(\mu_{w_*(s)}) = 2(1-2^{2-s})/(4-2^s)^2$. For $s \in (1,2)$ we get $\mathscr{E}_{K^{(s)}}(\mu_{w_*(s)}) < \mathscr{E}(\mu^+) = -1/2$, and there exist signed measures in $\mathscr{M}(1)$ such that $\mathscr{E}_{K^{(s)}}(\mu_w) < \mathscr{E}(\mu^+)$. Therefore, minimum-energy signed measures with total mass one are not probability measures. For $s \geq 2$, $\lim_{w\to+\infty} \mathscr{E}_{K^{(s)}}(\mu_w) = -\infty$, and there is no minimum-energy signed measure; in particular, $\mathscr{E}_{K^{(s)}}(\mu_w) = -(w+1)/2$ for $s = 2$. ◁

Example 3.7 (continuation of Examples 3.3 and 3.5)**.** Consider Riesz kernels $K_{(s)}$, see (3.8), for $\mathscr{X} = \mathscr{B}_d(\mathbf{0},1)$, $d > 2$ and $s \in (0, d-2)$; the minimum-energy probability measure μ^+ is then uniform on the unit sphere $\mathscr{S}_d(\mathbf{0},1)$ and the potential at all interior points satisfies $P_{\mu^+}(\mathbf{x}) > \mathscr{E}_{K_{(s)}}(\mu^+)$. Consider the signed measure $\mu_w = (1+w)\mu^+ -$

$w\mu^{(r)}$, with $\mu^{(r)}$ uniform on the sphere $\mathscr{S}_d(\mathbf{0}, r)$ with radius $r \in (0,1)$. Calculations similar to those in the proof of [56, Th. 1.32] show that $\mathscr{E}_{K_{(s)}}(\mu_w) < \mathscr{E}_{K_{(s)}}(\mu^+)$ for w small enough, indicating that μ^+ is not the minimum-energy signed measure with total mass one. ◁

3.3.4. When minimum-energy signed measures are probability measures. Unlike minimum-energy probability measures, minimum-energy signed measures do not always exist, but the following property provides a sufficient condition for their existence. Also, we shall see in Section 3.5.1 that the existence is always guaranteed after a suitable modification of the kernel.

Theorem 3.3. *Assume that K is ISPD and translation invariant, with $K(\mathbf{x}, \mathbf{x}') = \Psi(\mathbf{x} - \mathbf{x}')$ and Ψ continuous, twice differentiable except at the origin, with Laplacian $\Delta_\Psi(\mathbf{x}) = \sum_{i=1}^d \partial^2\Psi(\mathbf{x})/\partial x_i^2 \geq 0$, $\mathbf{x} \neq \mathbf{0}$. Then there exists a unique minimum-energy signed measure μ_K^* in $\mathscr{M}(1)$, and μ_K^* is a probability measure.*

Proof. The conditions of Theorem 3.1 are satisfied, and there exists a unique minimum-energy probability measure μ^+ such that $P_{\mu^+}(\mathbf{x}) \geq \mathscr{E}_K(\mu^+)$ for all $\mathbf{x} \in \mathscr{X}$. It also satisfies $P_{\mu^+}(\mathbf{x}) = \mathscr{E}_K(\mu^+)$ on the support of μ^+. On the other hand, the conditions on K imply that for any μ in $\mathscr{M}^+(1)$, $P_\mu(\cdot)$ is subharmonic outside the support of μ, see, e.g., [56, Sect. I.2]. The first maximum principle of potential theory thus holds [56, Th. 1.10]: $P_\mu(\mathbf{x}) \leq c$ on the support of μ implies $P_\mu(\mathbf{x}) \leq c$ everywhere. Applying this to μ^+, we obtain that $P_{\mu^+}(\mathbf{x}) \leq \mathscr{E}_K(\mu^+)$ everywhere; therefore, $P_{\mu^+}(\mathbf{x}) = \mathscr{E}_K(\mu^+)$ for all $\mathbf{x} \in \mathscr{X}$. Theorem 3.2 implies that μ^+ is the minimum-energy signed measure with total mass one. ■

The central argument for the proof of the property above is that $P_\mu(\cdot)$ is subharmonic outside the support of μ for any probability measure μ with finite energy. The weaker condition $\Psi(x - x') = \psi(|x - x'|)$ with $\psi(\cdot)$ convex on $(0, \infty)$ is sufficient when $d = 1$, which corresponds to the result of Hájek (1956). When $d \geq 2$ with $\Psi(\mathbf{x} - \mathbf{x}') = \psi(\|\mathbf{x} - \mathbf{x}'\|)$, $\psi(\cdot)$ must have a singularity at 0 to have $\Delta_\Psi(\mathbf{x}) \geq 0$ for all $\mathbf{x} \neq 0$. For the Riesz kernels $K_{(s)}$ of (3.8), we have $\Delta(\|\mathbf{x}\|^{-s}) = s(s + 2 - d)/\|\mathbf{x}\|^{s+2}$, $\mathbf{x} \neq \mathbf{0}$. When $d > 2$ and $s \in (0, d-2]$, P_μ is superharmonic in $\mathbb{R}^d$, and when $s \in [d-2, d)$, P_μ is subharmonic outside the support of μ, μ^+ being then the minimum-energy signed measure [35, 56]. This is also true for the logarithmic kernel for $d \leq 2$, with $\Delta(-\log\|\mathbf{x}\|) = (2 - d)/\|\mathbf{x}\|^2$, $\mathbf{x} \neq \mathbf{0}$. Examples 3.5 and 3.7 give an illustration.

Other examples of kernels satisfying the condition of Theorem 3.3 are given by $K(\mathbf{x}, \mathbf{x}') = h[K_{(s)}(\mathbf{x}, \mathbf{x}')]$ where $K_{(s)}$ is a Riesz kernel with $s \in [d-2, d)$ and h is a twice-differentiable increasing and convex function (in fact, the continuity of h is sufficient, see [1, p. 13]).

In Theorem 3.3 we can also consider CISPD kernels. For example, for the kernels $K^{(s)}$ of (3.7), we have $\Delta(-\|\mathbf{x}\|^s) = s(2 - s - d)/\|\mathbf{x}\|^{2-s}$, $\mathbf{x} \neq \mathbf{0}$. Potentials are superharmonic for $d \geq 2$. When $d = 1$, they are superharmonic for $s \in [1, 2)$; they are subharmonic and satisfy the maximum principle for $s \in (0, 1)$, see Example 3.6.

3.4. Best Linear Unbiased Estimator (BLUE) of β_0. In Section 3.5.2, we shall see that Bayesian integration in the model (2.1) corresponds to the construction of the BLUE of β_0 in a model with a suitably modified covariance. Here we consider the BLUE in the original model, and show that its existence is equivalent to that of a minimum-energy signed measure for K.

3.4.1. Continuous BLUE. Consider again the situation of Section 2.2 where $\sigma^2 K$ corresponds to the covariance of a random field Z_x. Suppose that we may

observe $f(\cdot)$ over $\mathscr{X}$ in order to estimate β_0 in the regression (location) model with correlated errors (2.1). Any linear estimator of β_0 takes the general form

$$\hat{\beta}_0 = \hat{\beta}_0(\xi) = \int_{\mathscr{X}} f(\mathbf{x})\, \mathrm{d}\xi(\mathbf{x}) = I_\xi(f)$$

for some $\xi \in \mathscr{M}$, and $\hat{\beta}_0(\xi)$ is unbiased when $\xi \in \mathscr{M}(1)$. Its variance is

$$V_\xi = \mathbb{E}\{(\hat{\beta}_0(\xi) - \beta_0)^2\} = \sigma^2\, \mathscr{E}_K(\xi)\,;$$

see [68, Sect. 4.2]. The existence of a minimum-energy signed measure μ_K^* is then equivalent to the existence of the continuous BLUE $\hat{\beta}_0^*$ for β_0, with $\hat{\beta}_0^* = \hat{\beta}_0(\mu_K^*)$; the variance of $\hat{\beta}_0^*$ is proportional to the minimum energy $\mathscr{E}_K(\mu_K^*)$, and Theorem 3.2 corresponds to Grenander's theorem [43]. Also, from that theorem, the existence of μ_K^* is equivalent to the existence of an equilibrium measure that yields a constant potential on $\mathscr{X}$. It can be related to a property of the generalized capacity C_K^*, as shown in the following theorem.

Theorem 3.4. *When K is ISPD, the constant function $1_\mathscr{X}$ equal to 1 on $\mathscr{X}$ belongs to the space $\mathscr{P}_K$ of potential fields if and only if there exists a minimum-energy signed measure $\mu_K^* \in \mathscr{M}(1)$, with $\mathscr{E}_K(\mu_K^*) \neq 0$. Moreover, the generalized capacity C_K^* is finite and nonzero, and satisfies $\|1_\mathscr{X}\|_{\mathscr{P}_K}^2 = C_K^*$.*

Proof. Suppose that $1_\mathscr{X} \in \mathscr{P}_K$. There exists $\mu \in \mathscr{M}_K$ such that $P_\mu = 1_\mathscr{X}$; that is, $P_\mu(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathscr{X}$. The definition of P_μ yields $\mathscr{E}_K(\mu) = \mu(\mathscr{X})$, which is finite and strictly positive since K is ISPD and $\mu \neq 0$. Denote $\mu' = \mu/\mu(\mathscr{X}) \in \mathscr{M}(1)$. We obtain $P_{\mu'}(\mathbf{x}) = 1/\mu(\mathscr{X}) = \mathscr{E}_K(\mu') > 0$ for all $\mathbf{x} \in \mathscr{X}$. Theorem 3.2 implies that μ' is the minimum-energy measure μ_K^*. Also, $C_K^* = 1/\mathscr{E}_K(\mu') = \mu(\mathscr{X}) \neq 0$, with $\|1_\mathscr{X}\|_{\mathscr{P}_K}^2 = \mathscr{E}_K(\mu)$, see Section 3.2.

Suppose now that there exists a minimum-energy signed measure $\mu_K^* \in \mathscr{M}(1)$ with $\mathscr{E}_K(\mu_K^*) \neq 0$. Theorem 3.2 implies that $P_{\mu_K^*}(\mathbf{x}) = \mathscr{E}_K(\mu_K^*)$ for all $\mathbf{x} \in \mathscr{X}$. For $\mu = \mu_K^*/\mathscr{E}_K(\mu_K^*)$, we get $P_\mu(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathscr{X}$, and $\|1_\mathscr{X}\|_{\mathscr{P}_K}^2 = \mathscr{E}_K(\mu) = 1/\mathscr{E}_K(\mu_K^*)$. ∎

Under the conditions of Theorem 3.3, the BLUE exists, $\hat{\beta}_0^* = \hat{\beta}_0(\mu_K^+)$, with μ_K^+ the minimum-energy probability measure, and its variance equals $\sigma^2\mathscr{E}_K(\mu_K^+)$. The existence of a minimum-energy signed measure is not guaranteed in general, in particular when $K(\mathbf{x}, \mathbf{x}') = \Psi(\mathbf{x} - \mathbf{x}')$ and Ψ is differentiable at 0; see Example 3.8 below.

3.4.2. Discrete BLUE. Consider the framework of Section 2.2, with the same notation, and suppose that the n design points $\mathbf{x}_i$ in $\mathbf{X}_n$ are fixed. Any linear estimator of β_0 in (2.1) has then the form $\tilde{\beta}_0^n = \mathbf{w}_n^\top \mathbf{y}_n$, with $\mathbf{w}_n = (w_1, \ldots, w_n)^\top \in \mathbb{R}^n$. The unbiasedness constraint imposes $\mathbf{w}_n^\top \mathbf{1}_n = 1$. The variance of $\tilde{\beta}_0^n$ equals $\sigma^2 \mathbf{w}_n^\top \mathbf{K}_n \mathbf{w}_n$, and the BLUE corresponds to the estimator $\hat{\beta}_0^n$ given by (2.3) (we assume that $\mathbf{K}_n$ is nonsingular). The minimum-energy signed measure in $\mathscr{M}(1)$ (here discrete) μ_K^* is defined by the weights $\mathbf{w}_n^* = \mathbf{K}_n^{-1}\mathbf{1}_n/(\mathbf{1}_n^\top \mathbf{K}_n^{-1}\mathbf{1}_n)$ set on the points in $\mathbf{X}_n$; its energy is $\mathscr{E}_K(\mu_K^*) = \mathbf{w}_n^{*\top} \mathbf{K}_n \mathbf{w}_n^* = 1/(\mathbf{1}_n^\top \mathbf{K}_n^{-1}\mathbf{1}_n)$ and the variance of the BLUE equals $\sigma^2\mathscr{E}_K(\mu_K^*)$. Note that some components of $\mathbf{w}_n^*$ may be negative and that the potential associated with the measure $\mu_K^*/\mathscr{E}_K(\mu_K^*)$ on $\mathscr{X} = \mathbf{X}_n$ gives the constant function $1_\mathscr{X} = \mathbf{1}_n$, see Theorem 3.4. The optimal design problem for the discrete BLUE thus corresponds to the determination of the n-point set maximizing $\mathbf{1}_n^\top \mathbf{K}_n^{-1} \mathbf{1}_n$.

Example 3.8. Consider $K(x, x') = \exp(-\theta|x - x'|)$, $\theta > 0$, for $x, x' \in \mathscr{X} = [0, 1]$. K

is ISPD and satisfies

$$1_{\mathscr{X}} = \frac{K(\cdot,0)+K(\cdot,1)}{2} + \frac{\theta}{2}\int_0^1 K(\cdot,x)\mathrm{d}x\,,$$

so that $1_{\mathscr{X}} \in \mathscr{P}_K$, see [2]. The minimum-energy measure in $\mathscr{M}(1)$ is $\mu_K^* = (\delta_0+\delta_1+\theta\mu_L)/(\theta+2)$, with μ_L the Lebesgue measure on $\mathscr{X}$, and $\mu_K^* \in \mathscr{M}^+(1)$. The BLUE of β_0 in (2.1) is $\hat{\beta}_0^* = \int_{\mathscr{X}} f(x)\,\mathrm{d}\mu_K^*(x)$, its variance equals $\sigma^2\mathscr{E}_K(\mu_K^*) = 2\sigma^2/(\theta+2)$, see [68, p. 56]. Note that $K' = K - 2/(\theta+2)$ is still positive definite, but $1_{\mathscr{X}} \notin \mathcal{H}_{K'}$ since $c^2K'-1$ is not positive definite for any $c \neq 0$, see, e.g., [9, p. 30], [73, p. 20].

Consider now the squared exponential kernel $K(x,x') = \exp(-\theta|x-x'|^2)$, $\theta > 0$. The constant $1_{\mathscr{X}}$ does not belong to $\mathcal{H}_K$ [91] and the BLUE of β_0 in (2.1) is not defined for that kernel. On the other hand, the discrete BLUE (2.3) is well-defined for any set of n distinct points x_i, $\hat{\beta}_0^n = \mathbf{w}_n^{*\top}\mathbf{y}_n = \mathbf{1}_n^\top\mathbf{K}_n^{-1}\mathbf{y}_n/(\mathbf{1}_n^\top\mathbf{K}_n^{-1}\mathbf{1}_n)$. Suppose that the n points x_i are equally spaced in $\mathscr{X} = [0,1]$. The process Z_x in (2.1) has mean square derivatives of all orders, and, roughly speaking, for large n the construction of the BLUE mimics the estimation of the derivatives of f and the weights w_i^* strongly oscillate between large positive and negative values. Figure 3.1-Left shows the optimal weights $(w_i^*/|w_i^*|)(\log_{10}(\max\{|w_i^*|,1\})$, truncated to absolute values larger than 1 and in log scale, when $x_i = (i-1)/(n-1)$, $i = 1,\ldots,n = 101$. In Figure 3.1-Right, the kernel is $K(x,x') = (1+\sqrt{5}|x-x'|+5|x-x'|^2/3)\exp(-\sqrt{5}|x-x'|)$ (Matérn 5/2), so that Z_x is twice mean-square differentiable; the construction of the BLUE mimics the estimation of the first and second order derivatives of f at 0 and 1: here, $1_{\mathscr{X}} \notin \mathscr{P}_K$ although $1_{\mathscr{X}} \in \mathcal{H}_K$; see [30, 46] and [23] for more details. ◁

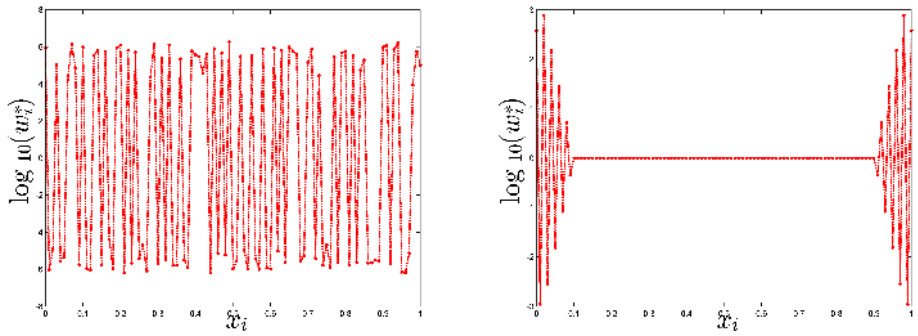

Figure 3.1: BLUE weights $(w_i^*/|w_i^*|)\,\log_{10}(\max\{|w_i^*|,1\})$ for $x_i = (i-1)/(n-1)$, $i = 1,\ldots,n = 101$. Left: $K(x,x') = \exp(-|x-x'|^2)$, Right: $K(x,x') = (1+\sqrt{5}|x-x'|+5|x-x'|^2/3)\exp(-\sqrt{5}|x-x'|)$ (Matérn 5/2).

Although a minimum-energy signed measure may not exist, in the next section we shall see how, for any measure $\mu \in \mathscr{M}(1)$ and any CISPD kernel K, we can modify K in such a way that the minimum-energy signed measure for the modified kernel exists (and coincides with μ).

3.5. Equilibrium measure and kernel reduction: MMD is equivalent to energy minimization for a reduced kernel. Minimum-energy signed measures, when they exist, satisfy the following property.

Lemma 3.3. *If K is CISPD and if a minimum-energy signed measure μ_K^* exists in $\mathscr{M}(1)$, we have $\mathscr{E}_K(\xi) = \mathscr{E}_K[\xi - \xi(\mathscr{X})\mu_K^*] + [\xi(\mathscr{X})]^2\,\mathscr{E}_K(\mu_K^*)\,,\ \forall\,\xi \in \mathscr{M}_K$.*

Proof. For any $\xi \in \mathscr{M}_K$, direct calculation gives

$$\begin{aligned}\mathscr{E}_K[\xi - \xi(\mathscr{X})\mu_K^*] &= \mathscr{E}_K(\xi) + [\xi(\mathscr{X})]^2\,\mathscr{E}_K(\mu_K^*) - 2\xi(\mathscr{X}) \int_{\mathscr{X}^2} K(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mu_K^*(\mathbf{x})\mathrm{d}\xi(\mathbf{x}') \\ &= \mathscr{E}_K(\xi) - [\xi(\mathscr{X})]^2\,\mathscr{E}_K(\mu_K^*)\,,\end{aligned}$$

where the second equality follows from (3.14). ∎

Under the conditions of Lemma 3.3, any $\xi \in \mathscr{M}(1)$ satisfies

$$\mathscr{E}_K(\xi) = \mathscr{E}_K(\xi - \mu_K^*) + \mathscr{E}_K(\mu_K^*)\,,$$

where the first term on the right-hand side equals the squared MMD $\gamma_K^2(\xi, \mu_K^*)$, see (3.9), and the second term does not depend on ξ. Minimizing the energy $\mathscr{E}_K(\xi)$ is thus equivalent to minimizing the MMD $\gamma_K(\xi, \mu_K^*)$. However, (*i*) μ_K^* may not exist, (*ii*) in many situations we wish to select a measure ξ having small MMD $\gamma_K(\xi, \mu)$ for *a given* measure $\mu \in \mathscr{M}_K$. This is the case in particular when one aims at evaluating the integral of a function with respect to some $\mu \in \mathscr{M}^+(1)$ (Section 2.2), or when we want to construct a space-filling design in $\mathscr{X}$, μ being then uniform.

3.5.1. Kernel reduction. Take any $\mu \in \mathscr{M}_K$ such that $\mu(\mathscr{X}) \neq 0$. Without any loss of generality, we assume $\mu \in \mathscr{M}(1)$. Following [21], we show how to modify the kernel K in such a way that minimizing the energy $\mathscr{E}_{K_\mu}(\xi)$, $\xi \in \mathscr{M}(1)$, for the new (reduced) kernel K_μ is equivalent to minimizing $\gamma_{K_\mu}(\xi, \mu)$. Define

$$K_\mu(\mathbf{x},\mathbf{x}') = K(\mathbf{x},\mathbf{x}') - P_\mu(\mathbf{x}) - P_\mu(\mathbf{x}') + \mathscr{E}_K(\mu)\,, \tag{3.15}$$

see [86]. One can readily check that the energy for this new reduced kernel K_μ satisfies $\mathscr{E}_{K_\mu}(\beta\mu) = 0$ for any real β and that the potential for μ associated with K_μ satisfies $\widetilde{P}_\mu(\mathbf{x}) = \int_{\mathscr{X}} K_\mu(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mu(\mathbf{x}') = 0$ for all $\mathbf{x}$.

Next theorem indicates that, for any given μ in $\mathscr{M}(1) \cap \mathscr{M}_K$, when considering signed measures ξ with total mass one, minimizing the energy $\mathscr{E}_{K_\mu}(\xi)$ is equivalent to minimizing the MMD $\gamma_K(\xi, \mu)$, provided that K is CISPD.

Theorem 3.5. *If K is CISPD, then for any $\mu \in \mathscr{M}(1) \cap \mathscr{M}_K$, we have*
(i) the reduced kernel K_μ defined by (3.15) is CISPD;
(ii) μ is the minimum-energy measure in $\mathscr{M}(1)$ for K_μ, and

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

Proof. For any nonzero $\xi \in \mathscr{M}_K$, direct calculation using (3.15) gives

$$\begin{aligned}\mathscr{E}_{K_\mu}(\xi) &= \mathscr{E}_K(\xi) - 2\xi(\mathscr{X}) \int_{\mathscr{X}^2} K(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mu(\mathbf{x})\mathrm{d}\xi(\mathbf{x}') + [\xi(\mathscr{X})]^2\,\mathscr{E}_K(\mu) \\ &= \mathscr{E}_K[\xi - \xi(\mathscr{X})\mu]\,.\end{aligned} \tag{3.16}$$

(*i*) When $\xi(\mathscr{X}) = 0$ we get $\mathscr{E}_{K_\mu}(\xi) = \mathscr{E}_K(\xi)$, which is strictly positive when $\xi \neq 0$, showing that K_μ is CISPD. (*ii*) Since $[\xi - \xi(\mathscr{X})\mu](\mathscr{X}) = 0$ and K is CISPD, $\mathscr{E}_{K_\mu}(\xi) > 0$ for $\xi \neq \xi(\mathscr{X})\mu$, showing that μ is the (unique) minimum-energy signed measure in $\mathscr{M}(1)$ for K_μ. Since $\mathscr{E}_{K_\mu}(\mu) = 0$, Lemma 3.3 with K_μ substituted for K implies that $\mathscr{E}_{K_\mu}(\xi) = \mathscr{E}_{K_\mu}[\xi - \xi(\mathscr{X})\mu]$ for any $\xi \in \mathscr{M}_K$, which, together with (3.16), concludes the proof. ∎

3.5.2. Kernel reduction, BLUE and Bayesian integration. Consider again the situation of Section 3.4, and define $\mathcal{P}_1$ as the orthogonal projection of $L^2(\mathcal{X}, \mu)$ onto the linear space spanned by the constant 1; see [37]. The model (2.1) can then be written as

$$f(\mathbf{x}) = \beta_0 + \mathcal{P}_1 Z_x + (\mathrm{Id}_{L^2} - \mathcal{P}_1) Z_x = \beta_0' + \widetilde{Z}_x \,, \tag{3.17}$$

where $\beta_0' = \beta_0 + \mathcal{P}_1 Z_x$ and $\widetilde{Z}_x = (\mathrm{Id}_{L^2} - \mathcal{P}_1) Z_x$, with $\widetilde{Z}_x$ having zero mean and covariance $\mathbb{E}\{\widetilde{Z}_x \widetilde{Z}_{x'}\} = \sigma^2 K_\mu(\mathbf{x}, \mathbf{x}')$. The extension to a model with a more general linear trend is considered in Appendix B. We have seen in Section 3.4 that the variance of the continuous BLUE of β_0 equals $\sigma^2 \mathscr{E}_K(\mu_K^*)$ provided that the minimum-energy signed measure μ_K^* exists. (Note that the prior on β_0' remains non-informative when the prior on β_0 is non-informative.) On the other hand, we obtain now that the continuous BLUE of β_0' always exists: it coincides with $I_\mu(f)$ and its variance is $\sigma^2 \mathscr{E}_{K_\mu}(\mu) = 0$. Therefore, as mentioned in introduction, Bayesian integration for the model (2.1) with correlated errors is equivalent to parameter estimation in a location model with different correlation structure.

Remark 3.3 (other kernels with zero potential)**.** The approach via kernel reduction, based on a $L^2(\mathcal{X}, \mu)$ orthogonal projection, has the merit of simplicity and pleasant interpretation through the model (3.17), but it is not the only one that can provide a kernel with zero potential P_μ everywhere. Orthogonal projection for the RKHS scalar product is considered in [29] in order to construct the RKHS of zero mean functions; see also [25, Sect. 2.5] and [39] for further developments on ANOVA kernel decomposition. Another possibility is to consider the image of a kernel under a Stein operator, as recently considered in details in [18]; see also [71]. ◁

3.6. Separable kernels. From d kernels K_i respectively defined on $\mathcal{X}_i \times \mathcal{X}_i$, $i = 1, \ldots, d$, we can construct a separable (tensor-product) kernel as

$$K^{\otimes}(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{d} K_i(x_i, x_i') \,, \tag{3.18}$$

where $\mathbf{x} = (x_1, \ldots, x_d)^\top$ and $\mathbf{x}' = (x_1', \ldots, x_d')^\top$ belong to the product space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. The construction is particularly useful when considering product measures on $\mathcal{X}$, since, in some sense, it allows us to decompose an integration problem in a high dimensional space into its one-dimensional counterparts. Suppose that each K_i is uniformly bounded and CISPD on $\mathscr{M}^{(i)} = \mathscr{M}[\mathcal{X}_i]$; that is, K_i is ISPD on $\mathscr{M}^{(i)}(0)$, see Definitions 3.1 and 3.2. One can show that this is equivalent to $K^{\otimes}$ being ISPD on $\otimes_{i=1}^d \mathscr{M}^{(i)}(0)$, see [92, Th. 2]. In the same paper, the authors prove (Theorem 4) that if each K_i is moreover continuous and translation invariant, then $K^{\otimes}$ is ISPD on $\mathscr{M}(0)$; that is, $K^{\otimes}$ is CISPD on $\mathscr{M}$. Their proof relies on the equivalence between the CISPD and characteristic properties for uniformly bounded kernels, and on the characterization of characteristic continuous, uniformly bounded and translation-invariant kernels through a property of the support of the measure Λ defined in (3.4); see Section 3.1. A further attractive feature of separable kernels is that $K^{\otimes}(\mathbf{x}, \mathbf{x}')$ is large when $\mathbf{x}$ and $\mathbf{x}'$ are close in some coordinate, which is a useful feature for the generation of designs having good space-filling performance in projections, see Section 5.3; see also Remark 3.4.

An important property of separable kernels $K^{\otimes}$ is that kernel reductions $K_\mu^{\otimes}$ defined by (3.15) are easily obtained explicitly. Indeed, when $\mu = \otimes_{i=1}^d \mu^{(i)}$ is a

product measure on $\mathscr{X}$, then, for all $\mathbf{x} \in \mathscr{X}$,

$$\mathscr{E}_{K^{\otimes}}(\mu) = \prod_{i=1}^{d} \mathscr{E}_{K_i}(\mu^{(i)}), \tag{3.19}$$

$$P_{\mu}(\mathbf{x}) = \prod_{i=1}^{d} \int_{\mathscr{X}_i} K_i(x_i, x_i') \, \mathrm{d}\mu^{(i)}(x_i') = \prod_{i=1}^{d} P_{\mu^{(i)}}(x_i), \tag{3.20}$$

which facilitates the calculation of $\mathscr{E}_{K_{\mu}^{\otimes}}(\xi)$, in particular when ξ is a discrete measure as considered in Section 4. Table 3.1 gives the expressions of $\mathscr{E}_K(\mu)$ and $P_{\mu}(x)$ obtained for a few kernels, with μ uniform on $\mathscr{X} = [0,1]$; the expressions for the squared exponential and Matérn kernels can be found in [39]. Note that in each case $\mathscr{E}_{K_{\mu}}(\xi) > 0$ for any $\xi \in \mathscr{M}(1)$, $\xi \neq \mu$. Other more general results are provided in Table 1 of [16]. Expressions of $\mathscr{E}_K(\mu)$ and $P_{\mu}(x)$ for the triangular kernel $K_{\theta}(x, x') = \max\{1 - \theta|x - x'|, 0\}$, $\theta > 0$ (with μ uniform on $[0,1]$) are given in Appendix C.

Remark 3.4 (Projections in subspaces with smaller dimension)**.** Let ξ and μ be two measures in $\mathscr{M}(1)$ and consider their squared discrepancy $\gamma^2_{\widehat{K}^{\otimes}}(\xi, \mu)$ for the kernel $\widehat{K}^{\otimes}(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{d}[1 + K_i(x_i, x_i')]$. Direct calculation gives

$$\gamma^2_{\widehat{K}^{\otimes}}(\xi, \mu) = \sum_{m=1}^{d} \sum_{1 \leq i_1 < \cdots < i_m \leq d} \gamma^2_{K_{i_1 \cdots i_m}}(\xi, \mu),$$

where $\gamma^2_{K_{i_1 \cdots i_m}}(\xi, \mu) = \int_{\mathscr{X}^2} \prod_{j=1}^{d} K_{i_j}(x_{i_j}, x_{i_j}') \mathrm{d}(\xi - \mu)(\mathbf{x}) \mathrm{d}(\xi - \mu)(\mathbf{x}')$ corresponds to a squared discrepancy in the m-dimensional space $\mathscr{X} = \mathscr{X}_{i_1} \times \cdots \times \mathscr{X}_{i_m}$. When μ is uniform on $\mathscr{X} = [0,1]^d$, by choosing a discrete measure ξ_n with small $\gamma^2_{\widehat{K}^{\otimes}}(\xi_n, \mu)$ (see Section 4) we may thus construct a design having suitable space-filling properties in all sub-dimensional subspaces. One may refer to [47] for further developments and precisions, including in particular the derivation of quadrature error bounds and the introduction of different weights across dimensions. ◁

Table 3.1: Energy $\mathscr{E}_K(\mu)$ and potential $P_{\mu}(x)$ for different kernels K with μ uniform on $\mathscr{X} = [0,1]$; $P_{\mu}(x) = S_{\mu}(x) + S_{\mu}(1-x) + T_{\mu}(x)$; $S_{\mu}(\cdot)$ is continuously differentiable in $(0,1]$, $T_{\mu} = 0$ when K is translation invariant. $\mathbb{F}$ is the c.d.f. of the standard normal distribution.

$K(x, x')$	$\mathscr{E}_K(\mu)$	$S_{\mu}(x)$ [and $T_{\mu}(x)$]
$\mathrm{e}^{-\theta(x-x')^2}$	$\{\mathrm{e}^{-\theta} - 1 + \sqrt{\pi\theta}[2\mathbb{F}(\sqrt{2\theta}) - 1]\}/\theta$	$\sqrt{\pi}[\mathbb{F}(\sqrt{2\theta}x) - 1/2]/\sqrt{\theta}$
$\mathrm{e}^{-\theta\lvert x-x'\rvert}$	$2(\theta + \mathrm{e}^{-\theta} - 1)/\theta^2$	$x(1 - \mathrm{e}^{-\theta\lvert x\rvert})/(\theta\lvert x\rvert)$
$K_{3/2,\theta/\sqrt{3}}(x, x')$ in (3.5)	$2[\theta(2 + \mathrm{e}^{-\theta}) + 3(\mathrm{e}^{-\theta} - 1)]/\theta^2$	$x[2 - (2 + \theta\lvert x\rvert)\mathrm{e}^{-\theta\lvert x\rvert}]/(\theta\lvert x\rvert)$
$[(x - x')^2 + \epsilon]^{-1}$ $(\epsilon \geq 0)$	$(2/\sqrt{\epsilon})\arctan(1/\sqrt{\epsilon}) - \log(1 + 1/\epsilon)$	$(1/\sqrt{\epsilon})\arctan(x/\sqrt{\epsilon})$
$(\lvert x - x'\rvert + \epsilon)^{-1}$ $(\epsilon > 0)$	$2\,[(1+\epsilon)\log(1 + 1/\epsilon) - 1]$	$\mathrm{sign}(x)\log(1 + \lvert x\rvert/\epsilon)$
$(\lvert x - x'\rvert + \epsilon)^{-1/2}$ $(\epsilon > 0)$	$4\epsilon^{3/2}\,[2(1 + 1/\epsilon)^{3/2} - 2 - 3/\epsilon]/3$	$2\sqrt{\epsilon}\,\mathrm{sign}(x)(\sqrt{1 + \lvert x\rvert/\epsilon} - 1)$
$1 - \theta\lvert x - x'\rvert$ $(0 < \theta \leq 1)$	$1 - \theta/3$	$1/2 - \theta x\lvert x\rvert/2$
$\lvert x - x'\rvert^{-s}$ $(0 < s < 1)$	$2/(s^2 - 3s + 2)$	$x/[(1-s)\lvert x\rvert^s]$
$-\log\lvert x - x'\rvert$	$3/2$	$1/2 - x\log\lvert x\rvert$
$\lvert x\rvert + \lvert x'\rvert - \lvert x - x'\rvert$	$2/3$	$1/4 - x\lvert x\rvert/2$ $[T_{\mu}(x) = \lvert x\rvert]$
$\sqrt{\lvert x\rvert} + \sqrt{\lvert x'\rvert} - \sqrt{\lvert x - x'\rvert}$	$4/5$	$1/3 - 2x\sqrt{\lvert x\rvert}/3$ $[T_{\mu}(x) = \sqrt{\lvert x\rvert}]$

4. Experimental design. From Section 3.3.1, the construction of an optimal measure ξ^* minimizing $\mathscr{E}_K(\xi - \mu)$ forms a particular convex problem (quadratic), and

therefore presents some similarities with optimal experimental design in a parametric framework; see [54, 32] for early contributions. There is a noticeable difference however: optimal experimental design aims at *determining* the probability measure (called design measure) ξ^* that minimizes a convex functional $\phi(\xi)$, usually a function $\Phi[\mathbf{M}(\xi)]$ of the information matrix $\mathbf{M}(\xi)$ in a parametric model. Here, the optimal measure is known ($\xi^* = \mu$), and we wish to construct a discrete measure, with a limited number n of support points, which is close to μ in the sense of having small maximum mean discrepancy $\sqrt{\mathscr{E}_K(\xi - \mu)}$.

Consider an n-point design $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, with $\mathbf{x}_i \in \mathscr{X}$ for all i. For ξ_n a finite signed measure supported on $\mathbf{X}_n$, $\xi_n = \sum_{i=1}^n w_i \, \delta_{\mathbf{x}_i}$, we denote $\mathbf{w}_n = (w_1, \ldots, w_n)^\top$. As in Section 2.2, we assume that $\mu \in \mathscr{M}^+(1)$, with special attention to space-filling design for which μ is uniform on a compact subset $\mathscr{X}$ of $\mathbb{R}^d$. We assume that K is a bounded ISPD kernel (and is thus SPD) and that μ has finite energy $\mathscr{E}_K(\mu)$, see (2.8). For space-filling design, we may restrict our attention to translation-invariant kernels. Direct calculation gives

$$\begin{aligned} \gamma_K^2(\xi_n, \mu) = \mathscr{E}_K(\xi_n - \mu) &= \mathbf{w}_n^\top \mathbf{K}_n \mathbf{w}_n - 2\mathbf{w}_n^\top \mathbf{p}_n(\mu) + \mathscr{E}_K(\mu)\,, \\ &= \sum_{i,j} w_i w_j \, K(\mathbf{x}_i, \mathbf{x}_j) - 2 \sum_{i=1}^n w_i \, P_\mu(\mathbf{x}_i) + \mathscr{E}_K(\mu)\,, \quad (4.1) \end{aligned}$$

where $\{\mathbf{K}_n\}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1, \ldots, n$, and $\mathbf{p}_n(\mu)$ is given by (2.5). Note that $\mathscr{E}_K(\mu)$ and the $P_\mu(\mathbf{x}_i)$ have simple expressions when K is a separable kernel and $\mu = \otimes_{i=1}^d \mu^{(i)}$ is a product measure on $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_d$, see (3.19, 3.20). Monte-Carlo approximation, based on a large i.i.d. sample from μ, or a low-discrepancy sequence, can always be used instead.

4.1. Discrepancies and covering radius. Since our initial motivation is to construct designs having good space-filling properties, in this section we give some arguments supporting the intuition that designs with small MMD have small covering radius. We consider the case where $\mathscr{X}$ is the d-dimensional hypercube $[0,1]^d$.

4.1.1. Star-discrepancy. Low discrepancy sequences and point sets have low dispersion, in the sense that, when $\mathbf{X}_n$ is an n-point design in $\mathscr{X}$,

$$\frac{1}{\sqrt{d}} \, \mathsf{CR}(\mathbf{X}_n) \leq \mathsf{CR}_\infty(\mathbf{X}_n) = \max_{\mathbf{x} \in \mathscr{X}} \min_{1 \leq i \leq n} \|\mathbf{x} - \mathbf{x}_i\|_\infty \leq D^{1/d}(\mathbf{X}_n) \leq 2 D_*^{1/d}(\mathbf{X}_n)\,,$$

with $D(\mathbf{X}_n)$ and $D_*(\mathbf{X}_n)$ respectively the extreme and star discrepancies of $\mathbf{X}_n$; see, e.g., [69, p. 15 and 152]. Hence, low discrepancy sequences or point sets also have low dispersion (small covering radii) — the reverse being wrong, as the example of Ruzsa sequence shows [69, p. 154].

The connection between star discrepancy and covering radius is even stronger when considering design measures and weighted discrepancies. Consider the case $d = 1$, and let ξ_n be a probability measure supported on $\mathbf{X}_n$ with weight w_i on x_i. Assume, without any loss of generality, that $0 \leq x_1 < x_2 < \cdots < x_n \leq 1$. The weighted star discrepancy of ξ_n is defined as $D^*(\xi_n) = \sup_{0 \leq t < 1} \left| \sum_{i:\, x_i \leq t} w_i - t \right|$. The covering radius of $\mathbf{X}_n$ is $\mathsf{CR}(\mathbf{X}_n) = \max\{x_1, (x_2 - x_1)/2, \ldots, (x_n - x_{n-1})/2, 1 - x_n\}$. For simplicity, we restrict our attention to designs with $x_1 = 0$ and $x_n = 1$. We then have the following result.

Theorem 4.1. *For any design ξ_n such that $0 = x_1 < x_2 < \cdots < x_{n-1} < x_n = 1$,*

(i) $\mathsf{CR}(\mathbf{X}_n) = D^*(\xi_n^*)$*, where* ξ_n^* *has the weights* $w_1^* = x_2/2$*,* $w_n^* = (1 - x_{n-1})/2$ *and* $w_i^* = (x_{i+1} - x_{i-1})/2$ *for* $i = 2, \ldots, n-1$*;*

(ii) $D^*(\xi_n) > \mathsf{CR}(\mathbf{X}_n)$ *for any other probability measure* ξ_n *supported on* $\mathbf{X}_n$*.*

Proof. One can check that, for any design ξ_n supported on $\mathbf{X}_n$,

$$D^*(\xi_n) = \max_{1 \le i \le n} \left\{ \frac{w_i}{2} + \left| x_i - \frac{W_i + W_{i-1}}{2} \right| \right\} \tag{4.2}$$

where $W_0 = 0$ and $W_i = w_1 + \cdots + w_i$ for $i = 1, \ldots, n$. This expression is a generalization of that in [69, Theorem 2.6] for the classical star discrepancy. It is then straightforward to check that $D^*(\xi_n^*) = \mathsf{CR}(\mathbf{X}_n)$. Moreover, if we take $\xi_n = \xi_n^*$, then all the terms in the right-hand side of (4.2) are equal to $\mathsf{CR}(\mathbf{X}_n)$:

$$\frac{w_i^*}{2} + \left| x_i - \frac{W_i^* + W_{i-1}^*}{2} \right| = \mathsf{CR}(\mathbf{X}_n), \quad i = 1, \ldots, n\,.$$

This implies that for any other set of weights $w_1, \ldots, w_n$ we have $D^*(\xi_n) > \mathsf{CR}(\mathbf{X}_n)$. ∎

4.1.2. MMD. We have $\mathsf{CR}(\mathbf{X}_n) > n^{-1/d}/V_d$ and $\mathsf{CR}_\infty(\mathbf{X}_n) \ge n^{-1/d}/2$, with V_d the volume of the unit ball $\mathscr{B}_d(\mathbf{0}, 1)$, since the n balls (respectively, hypercubes) centred at the $\mathbf{x}_i$, with common radius $\mathsf{CR}(\mathbf{X}_n)$ (respectively, edge length $\mathsf{CR}_\infty(\mathbf{X}_n)$) must cover $\mathscr{X}$. The more precise bound $\mathsf{CR}_\infty(\mathbf{X}_n) \ge 1/(2\lfloor n^{1/d} \rfloor)$ also holds true [69, Th. 6.8]. The covering radii of optimal designs of size n decrease at the same rate, with $\limsup_{n\to\infty} n^{1/d} \min_{\mathbf{X}_n} \mathsf{CR}(\mathbf{X}_n) \le 2/V_d^{1/d}$, where $V_d = \pi^{d/2}/\Gamma(d/2+1)$; see, e.g., [74, Section 2.2]. The upper bound is slightly worse when considering sequential designs; the existence of an extensible point sequence such that $\lim_{n\to\infty} n^{1/d}\, \mathsf{CR}_\infty(\mathbf{X}_n) = 1/\log(4)$ is proved in [69, Th. 6.9]. The following property indicates that for suitable kernels the MMD discrepancy $\gamma_K(\xi_n, \mu)$ given by (3.2) also yields an upper bound on $\mathsf{CR}_\infty(\mathbf{X}_n)$.

Theorem 4.2. *Let* K *be a bounded ISPD separable kernel on* $\mathscr{X} = [0,1]^d$ *such that each* K_i *in* (3.18) *is translation invariant, with* $K_i(x, x') = \psi(|x - x'|)$ *and* $\psi(0) = 1$*; denote* $P_{0,\mu} = \int_0^1 \psi(r)\,\mathrm{d}r$*. Let* $\mathbf{X}_n$ *be an* n*-point design in* $\mathscr{X}$ *and* ξ_n *denote any probability measure supported on* $\mathbf{X}_n$*, with* $\gamma_K(\xi_n, \mu)$ *its MMD,* μ *being the uniform measure on* $\mathscr{X}$*.*

(i) If ψ *is strictly positive and strictly decreasing on* $\mathbb{R}^+$*, then,*

$$\mathsf{CR}_\infty(\mathbf{X}_n) \le \psi^{-1}\left[P_{0,\mu}^d - \gamma_K(\xi_n, \mu)\right]\,. \tag{4.3}$$

(ii) If $\psi = \psi_\theta$ *has bounded support* $[-1/\theta, 1/\theta]$*, then,*

$$\mathsf{CR}_\infty(\mathbf{X}_n) < 1/\theta \tag{4.4}$$

when $\gamma_K(\xi_n, \mu) < P_{0,\mu}^d$*. In particular, for the triangular kernel defined by* $\psi_\theta(r) = \max\{1 - \theta r, 0\}$*,* $r \ge 0$*, with* $\theta > 1$*,* $\mathsf{CR}_\infty(\mathbf{X}_n) < 1/\theta$ *when* $\gamma_K(\xi_n, \mu) < 1/(2\theta)^d$*.*

Proof. Denote $r_n = \mathsf{CR}_\infty(\mathbf{X}_n)$ and let $\mathbf{x}_0$ be a point in $\mathscr{X}$ such that $\|\mathbf{x}_0 - \mathbf{x}_i\|_\infty \ge r_n$ for all $\mathbf{x}_i \in \mathbf{X}_n$. We take $f = K_{\mathbf{x}_0}$ in (3.2), so that $I_\mu(K_{\mathbf{x}_0}) = P_\mu(\mathbf{x}_0)$, $I_{\xi_n}(K_{\mathbf{x}_0}) = P_{\xi_n}(\mathbf{x}_0)$, $\|f\|_{\mathcal{H}_K} = K^{1/2}(\mathbf{x}_0, \mathbf{x}_0) = 1$, and (3.2) implies $P_\mu(\mathbf{x}_0) - P_{\xi_n}(\mathbf{x}_0) \le \gamma_K(\xi_n, \mu)$.

(i) We have $P_{\xi_n}(\mathbf{x}_0) = \sum_{i=1}^n w_i \prod_{j=1}^d \psi(|x_{0j} - x_{ij}|) \le (\sum_{i=1}^n w_i)\psi(r_n) = \psi(r_n)$, where $w_i = \xi_n(\mathbf{x}_i)$ for all i. Therefore,

$$\min_{\mathbf{x}\in\mathscr{X}} P_\mu(\mathbf{x}) = P_\mu(\mathbf{0}) = P_{0,\mu}^d \le P_\mu(\mathbf{x}_0) \le P_{\xi_n}(\mathbf{x}_0) + \gamma_K(\xi_n, \mu) \le \psi(r_n) + \gamma_K(\xi_n, \mu)\,,$$

which gives (4.3).

(*ii*) Suppose that $r_n \geq 1/\theta$. Then, $P_{\xi_n}(\mathbf{x}_0) = 0$ and (3.2) implies $P_{0,\mu}^d \leq P_\mu(\mathbf{x}_0) \leq \gamma_K(\xi_n, \mu)$. We obtain (4.4) by contradiction. The triangular kernel with $\theta > 1$ satisfies $P_{0,\mu} = 1/(2\theta)$. ■

When K implicitly defines a norm on $\mathscr{X}$, following the same approach as in the proof of Theorem 4.2 we directly get a bound on the covering radius for the corresponding norm. In particular, when K is the product of exponential kernels with $\psi(r) = \exp(-\theta r)$, $\theta > 0$, we obtain for (*i*)

$$\max_{\mathbf{x}\in\mathscr{X}} \min_{1\leq i\leq n} \|\mathbf{x} - \mathbf{x}_i\|_1 \leq -\frac{1}{\theta} \log \left[\left(\frac{1 - \mathrm{e}^{-\theta}}{\theta} \right)^d - \gamma_K(\xi_n, \mu) \right],$$

and when K is the squared exponential kernel $K_\theta(\mathbf{x}, \mathbf{x}') = \exp(-\theta\|\mathbf{x} - \mathbf{x}'\|^2)$, $\theta > 0$, we obtain

$$\mathsf{CR}(\mathbf{X}_n) \leq \overline{r}_{[\theta]}(\mathbf{X}_n) = \left(-\frac{1}{\theta} \log \left\{ \frac{\pi^{d/2}}{\theta^{d/2}} \left[\mathbb{F}(\sqrt{2\theta}) - 1/2 \right]^d - \gamma_K(\xi_n, \mu) \right\} \right)^{1/2}, \quad (4.5)$$

with $\mathbb{F}$ the c.d.f. of the standard normal distribution.

Similarly, for (*ii*), we may use the spherical covariance model in dimension d instead of the product of d triangular kernels: $K(\mathbf{x}, \mathbf{x}')$ is proportional to the volume of the intersection of two balls centered at $\mathbf{x}$ and $\mathbf{x}'$ with radius $\rho/2$ and is therefore zero when $\|\mathbf{x} - \mathbf{x}'\| \geq \rho$; see, e.g., [95, Chap. 8].

Example 4.1. Consider the case $d = 2$ with $K = K_\theta$, the squared exponential kernel. We suppose that $\mathbf{X}_n$ has MMD discrepancy decreasing as $\sqrt{8/n}$ (which is the case when $\mathbf{X}_n$ is constructed by kernel herding; see Appendix A). The left panel of Figure 4.1 presents the upper bound $\overline{r}_{[\theta]}(\mathbf{X}_n)$ given by (4.5) as a function of θ, for three different values of n. The red solid line (bottom) corresponds to the limiting case when n tends to infinity (that is, when $\gamma_K(\xi_n, \mu)$ is set to zero in (4.5)). For each finite n, the upper bound is infinite if θ is larger than the value $\theta_{\max}(n)$ such that $P_{0,\mu}^d = \sqrt{8/n}$, see the right panel of Figure 4.1. Figure 4.2-left shows $\overline{r}_*(\mathbf{X}_n) = \min_\theta \overline{r}_{[\theta]}(\mathbf{X}_n)$ as a function of n; the minimum is obtained at $\theta_*(\mathbf{X}_n)$ shown on the right panel.

The bound $\overline{r}_{[\theta]}(\mathbf{X}_n)$ (or more generally the bound on $\mathsf{CR}_\infty(\mathbf{X}_n)$) is very pessimistic, but Figure 4.1 nevertheless suggests that θ should increase at suitable rate as n increases, in agreement with common intuition. Using a covariance kernel with correlation length $L = \mathcal{O}(n^{-1/d})$ seems reasonable; see the examples of Section 5. ◁

Remark 4.1 (improved bound on $\mathsf{CR}_\infty(\mathbf{X}_n)$)**.** In the proof of Theorem 4.2 we consider that all points $\mathbf{x}_i \in \mathbf{X}_n$ can be at ℓ_∞ distance r_n from $\mathbf{x}_0$, whereas some design points are necessarily further away. This consideration yields a tighter upper bound on r_n. For instance, when $d = 1$, if we take $\xi_n = \xi_{n,e}$, the empirical measure associated with $\mathbf{X}_n$, we obtain that $P_{\xi_n}(x_0) \leq P_{\tilde{\xi}_n}(1/2)$, where $\tilde{\xi}_n = \frac{1}{n}\left[\sum_{k=1}^{k_{r_n}} \delta_{(2k-1)r_n} + \delta_{1-(2k-1)r_n} + (n - 2k_{r_n})\delta_{1/2-r_n}\right]$, with $k_{r_n} = \lfloor 1/(4r_n) \rfloor$, which gives the inequality

$$P_{\tilde{\xi}_n}(1/2) = \left[1 - \frac{2}{n} k_{r_n}\right] \psi(r_n) + \frac{2}{n} \sum_{k=0}^{k_{r_n}} \psi[1/2 - (2k-1)r_n] \geq P_{0,\mu} - \gamma_K(\xi_n, \mu);$$

compare with (4.3) with $d = 1$. In practice, however, the improvement is negligible and the upper bound on $\mathsf{CR}_\infty(\mathbf{X}_n)$ remains pessimistic. ◁

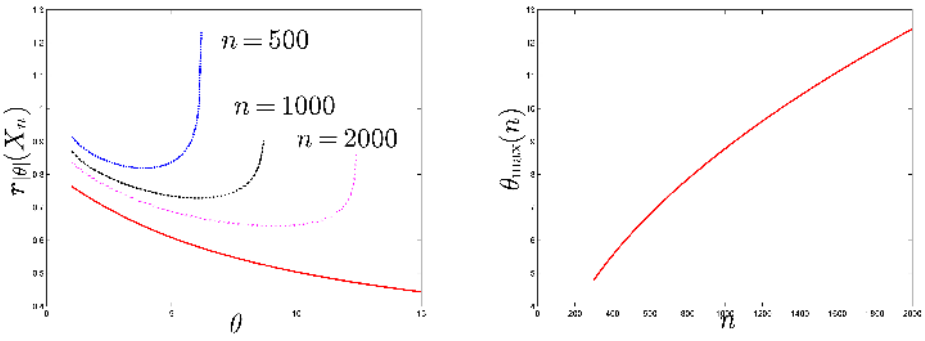

Figure 4.1: Squared exponential kernel $K_\theta(\mathbf{x}, \mathbf{x}') = \exp(-\theta\|\mathbf{x}-\mathbf{x}'\|^2)$, $d = 2$. Left: $\overline{r}_{[\theta]}(\mathbf{X}_n)$ (4.5) as a function of θ when $\gamma_K(\xi_n, \mu) = \sqrt{8/n}$, for $n = 500$, 1 000 and 2 000. Right: $\theta_{\max}(n)$ such that $P^d_{0,\mu} = \sqrt{8/n}$.

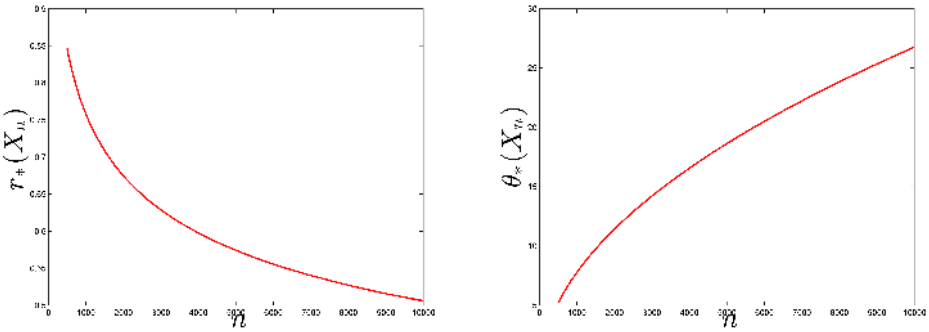

Figure 4.2: Squared exponential kernel $K_\theta(\mathbf{x}, \mathbf{x}') = \exp(-\theta\|\mathbf{x} - \mathbf{x}'\|^2)$, $d = 2$. Left: $\overline{r}_*(\mathbf{X}_n) = \min_\theta \overline{r}_{[\theta]}(\mathbf{X}_n)$ as a function of n. Right: $\theta_*(\mathbf{X}_n) = \arg\min_\theta \overline{r}_{[\theta]}(\mathbf{X}_n)$ as a function of n.

Remark 4.2. When the triangular kernel defined by $\psi_\theta(r) = \max\{1 - \theta r, 0\}$ is used for kernel herding, we have $\gamma_K(\xi_n, \mu) < \sqrt{8/n}$, see Appendix A, and (ii) implies that for any $r_0 = 1/\theta \leq 1$, $\mathsf{CR}_\infty(\mathbf{X}_n) < r_0$ for $n > 8(2/r_0)^{2d}$. This rate of decrease of $\mathsf{CR}_\infty(\mathbf{X}_n)$ is much slower than the best achievable rate $n^{-1/d}$. The existence of extensible point sequences achieving the optimal rate $n^{-1/d}$ on a smooth Riemannian manifold is established in [14]. The construction relies on the consideration of another function than $f = K_{\mathbf{x}_0}$ in (3.2), having support in $\mathscr{B}_d(\mathbf{x}_0, r_n)$ when $\mathbf{X}_n$ satisfies $\|\mathbf{x}_i - \mathbf{x}_0\| \geq r_n$ for all $\mathbf{x}_i$, with a large integration error $I_\mu(f) - I_{\xi_n}(f) = I_\mu(f)$ and a small norm $\|f\|_{\mathcal{H}}$, where $\mathcal{H}$ is a particular Sobolev space.

Denote by $e_p(\mathbf{X}_n, \mu) = (\mathsf{E}_\mu\{\min_{1\leq i\leq n} \|\mathbf{X} - \mathbf{x}_i\|^p\})^{1/p}$, $p > 0$ the L^p mean quantization error induced by $\mathbf{X}_n$. From Zador theorem, $\lim_{n\to\infty} n^{1/d} \min_{\mathbf{X}_n} e_p(\mathbf{X}_n, \mu)$ exists (and equals the infimum over n when μ is uniform on $[0,1]^d$); see [42]. The fact that the greedy construction of a design $\mathbf{X}^*_n$, recursively optimal step by step for the L^p quantization error, achieves $\limsup_{n\to\infty} n^{1/d} e_p(\mathbf{X}^*_n, \mu) < \infty$, is proved in [61]. For fixed n, a design minimizing $e_p(\mathbf{X}_n, \mu)$ can be constructed via clustering (see, e.g.,

[26]), with the famous k-means algorithm for $p = 2$ [59], k-medians for $p = 1$ [17], or with any general optimization algorithm. A combination of clustering with particle swarm optimization is used in [63] for arbitrary $p \geq 2$; clustering is considered in [74] for the limiting case $p = \infty$, with $e_\infty(\mathbf{X}_n, \mu) = \mathsf{CR}(\mathbf{X}_n)$. In general, these constructions are far more complicated than those using the methods of Section 4.3 and 4.4. ◁

4.2. Design criteria based on Bayesian quadrature. Since s_n^2 given by (2.7) does not depend on the function f considered, a design $\mathbf{X}_n$ for Bayesian integration can in principle be chosen beforehand, by direct minimization of s_n^2. This corresponds to the approach followed in [72] where several quadrature rules are tabulated for several values of n. The next theorem collects several results from the literature (part (i) appeared in [48], part (ii) is a particular case in [52] called normalized Bayesian cubature) and in particular shows the connection between the minimum of $\mathscr{E}_K(\xi_n - \mu)$ with respect to weights $\mathbf{w}_n$ and the posterior variances s_n^2 and $s_{n,0}^2$ respectively given by (2.7) and (2.11). The extension of model (2.1) to models including a linearly parameterized mean function is considered in [52]; the extension to the estimation of several integrals is treated in [72]; see also Appendix B. We assume that all points in $\mathbf{X}_n$ are pairwise different and μ is not fully supported on $\mathbf{X}_n$.

Theorem 4.3. *Let K be an SPD kernel and let $\mu \in \mathscr{M}^+(1) \cap \mathscr{M}_K$.*

(i) The optimal unconstrained weights $\mathbf{w}_n^*$ that minimize $\mathscr{E}_K(\xi_n - \mu)$ are $\mathbf{w}_n^* = \mathbf{K}_n^{-1}\mathbf{p}_n(\mu)$ and the corresponding measure ξ_n^*, with weights $\mathbf{w}_n^*$, satisfies

$$\mathscr{E}_K(\xi_n^* - \mu) = s_{n,0}^2\,, \tag{4.6}$$

with $s_{n,0}^2$ given by (2.11).

(ii) The optimal weights $\hat{\mathbf{w}}_n$ that minimize $\mathscr{E}_K(\xi_n - \mu)$ under the constraint $\mathbf{w}_n^\top \mathbf{1}_n = \sum_{i=1}^n w_i = 1$ are

$$\hat{\mathbf{w}}_n = \left(\mathbf{K}_n^{-1} - \frac{\mathbf{K}_n^{-1}\mathbf{1}_n\mathbf{1}_n^\top\mathbf{K}_n^{-1}}{\mathbf{1}_n^\top\mathbf{K}_n^{-1}\mathbf{1}_n}\right)\mathbf{p}_n(\mu) + \frac{\mathbf{K}_n^{-1}\mathbf{1}_n}{\mathbf{1}_n^\top\mathbf{K}_n^{-1}\mathbf{1}_n}\,, \tag{4.7}$$

and the corresponding measure $\hat{\xi}_n$, with weights $\hat{\mathbf{w}}_n$, satisfies

$$\mathscr{E}_K(\hat{\xi}_n - \mu) = s_n^2\,, \tag{4.8}$$

with s_n^2 given by (2.7); the estimator (2.4) of the integral $I_\mu(f)$ is $\widehat{I}_n = \hat{\mathbf{w}}_n^\top \mathbf{y}_n$.

(iii) For any bounded signed measure $\xi_n = \sum_{i=1}^n w_i\,\delta_{\mathbf{x}_i}$ we can write

$$\mathscr{E}_K(\xi_n - \mu) = (\mathbf{w}_n - \mathbf{w}_n^*)^\top \mathbf{K}_n (\mathbf{w}_n - \mathbf{w}_n^*) + \mathscr{E}_K(\xi_n^* - \mu)\,, \tag{4.9}$$

and when the weights w_i sum to one, we have

$$\mathscr{E}_K(\xi_n - \mu) = (\mathbf{w}_n - \hat{\mathbf{w}}_n)^\top \mathbf{K}_n (\mathbf{w}_n - \hat{\mathbf{w}}_n) + \mathscr{E}_K(\hat{\xi}_n - \mu). \tag{4.10}$$

Proof. The expression for $\mathbf{w}_n^*$, (4.6) and (4.9) directly follow from the fact that $\mathscr{E}_K(\xi_n - \mu)$ is quadratic in $\mathbf{w}_n$, see (4.1). Since K is SPD, straightforward calculation using Lagrangian theory indicates that the minimization of $\mathscr{E}_K(\xi_n - \mu)$ under the constraint $\mathbf{w}_n^\top \mathbf{1}_n = 1$ gives (4.7) and (4.8). Suppose that $\mathbf{w}_n^\top \mathbf{1}_n = 1$, then $\mathscr{E}_K(\xi_n - \mu) = (\mathbf{w}_n - \hat{\mathbf{w}}_n + \hat{\mathbf{w}}_n - \mathbf{w}_n^*)^\top \mathbf{K}_n (\mathbf{w}_n - \hat{\mathbf{w}}_n + \hat{\mathbf{w}}_n - \mathbf{w}_n^*) + \mathscr{E}_K(\xi_n^* - \mu)$ gives (4.10) since $\mathbf{K}_n(\hat{\mathbf{w}}_n - \mathbf{w}_n^*)$ is proportional to $\mathbf{1}_n$ and $(\mathbf{w}_n - \hat{\mathbf{w}}_n)^\top \mathbf{1}_n = 0$. ■

Remark 4.3. Equation (4.6) is simply related to the fact that, for f the realization of a Gaussian RF with zero mean and covariance $\sigma^2 K(\cdot,\cdot)$, we have

$$\begin{aligned}\mathbb{E}\{[I_\mu(f) - I_{\xi_n}(f)]^2\} &= \sigma^2 \int_{\mathscr{X}^2} K(\mathbf{x},\mathbf{x}')\,\mathrm{d}(\xi_n-\mu)(\mathbf{x})\,\mathrm{d}(\xi_n-\mu)(\mathbf{x}') \\ &= \sigma^2 \left\| P_\mu - P_{\xi_n} \right\|^2_{\mathcal{H}_K} = \sigma^2 \mathscr{E}_K(\xi_n - \mu)\,,\end{aligned}$$

the minimum being attained for the Bayes predictor $\widehat{I}_n = \mathbf{p}_n^T(\mu)\mathbf{K}_n^{-1}\mathbf{y}_n$, that is, for the weights $\mathbf{w}_n^*$. (Note that we cannot use the reproducing property $f(\mathbf{x}) = \langle f, K_\mathbf{x}\rangle_K$ since w.p. 1 f does not belong to $\mathcal{H}_K$; compare with [48, Proposition 1].)

It is shown in [53] that polynomial-based quadrature rules can be interpreted as Bayesian quadrature in a model with zero mean for a suitably chosen (polynomial) kernel; the optimal n-point set (with $n = p+1$ for polynomials of degree p) minimizing the posterior variance (4.6) realizes the cubature rule. One may refer to the discussion in Remark B.1 of Appendix B for models that include a linearly parameterized mean: any cubature rules can be interpreted as Bayesian integration; see [52]. ◁

In the discrete case considered here, the minimum-energy signed measure $\hat{\xi}_n$ with total mass one always exists, but $\hat{\xi}_n$ is not necessarily a probability measure; that is, some weights $\hat{w}_i$ may be negative. Theorem 4.3 can be extended to the case where K is only conditionally SPD, but the computation of optimal weights $\hat{\mathbf{w}}_n$ is more involved when $\mathbf{K}_n$ is singular; see Remark 4.4.

Denote by $\widetilde{\mathbf{K}}_n$ the $n \times n$ matrix with elements $\{\widetilde{\mathbf{K}}_n\}_{i,j} = K_\mu(\mathbf{x}_i,\mathbf{x}_j)$, where K_μ is the reduced kernel (3.15); the corresponding vector of potential values at the $\mathbf{x}_i$ is then $\widetilde{\mathbf{p}}_n = (\widetilde{P}_\mu(\mathbf{x}_1),\dots,\widetilde{P}_\mu(\mathbf{x}_n))^\top = \mathbf{0}$. For measures ξ_n in $\mathscr{M}(1)$, in complement of (ii) of Theorem 4.3, we also have the following property. (Similar expressions for the posterior mean and variance are obtained for a kernel having zero potential (kernel imbedding); see for instance [70] where the Stein operator is used in a more general framework; see also Remark 3.3.)

Theorem 4.4. *For K an SPD kernel, $\mu \in \mathscr{M}^+(1) \cap \mathscr{M}_K$ and $\xi_n \in \mathscr{M}(1)$, we have*

$$\mathscr{E}_K(\xi_n - \mu) = \mathscr{E}_{K_\mu}(\xi_n) = \mathbf{w}_n^\top \widetilde{\mathbf{K}}_n \mathbf{w}_n\,. \tag{4.11}$$

The posterior mean (2.4) and variance (2.7) of $I_\mu(f)$ are respectively given by

$$\widehat{I}_n = \frac{\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1}\mathbf{y}_n}{\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n}\,, \tag{4.12}$$

$$\sigma^2 s_n^2 = \sigma^2 (\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n)^{-1}\,. \tag{4.13}$$

Proof. Equation (4.11) follows from Theorem 3.5. Since we assumed that μ is not fully supported on $\mathbf{X}_n$ and K is SPD, (4.11) gives $\inf_{\|\mathbf{w}_n\|=1} \mathbf{w}_n^\top \widetilde{\mathbf{K}}_n \mathbf{w}_n > 0$, which implies that $\widetilde{\mathbf{K}}_n$ has full rank. Direct calculation using (3.15) gives $\widetilde{\mathbf{K}}_n = \mathbf{K}_n - \mathbf{p}_n(\mu)\mathbf{1}_n^\top - \mathbf{1}_n\mathbf{p}_n^\top(\mu) + \mathscr{E}_K(\mu)\,\mathbf{1}_n\mathbf{1}_n^\top$. The expression for $\widetilde{\mathbf{K}}_n^{-1}$ then yields $\mathbf{1}_n^\top\widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n = 1/s_n^2$, with s_n^2 given by (2.7), proving (4.13). The expansion of $(\mathbf{1}_n^\top\widetilde{\mathbf{K}}_n^{-1}\mathbf{y}_n)/(\mathbf{1}_n^\top\widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n)$ gives (2.4), which proves (4.12). ■

Equations (4.12) and (4.13) indicate that $\widehat{I}_n$ is the BLUE of β_0' and $\sigma^2 s_n^2$ is its variance in the model (3.17), $f(\mathbf{x}) = \beta_0' + \widetilde{Z}_x$, see Sections 3.4.2 and 3.5.2. The reason is that predictions are not modified when using the reduced kernel K_μ instead

of K, that is, when considering the model $f(\mathbf{x}) = \beta_0' + \widetilde{Z}_x$ instead of (2.1); see [37, Sect. 5.4]. It implies that the expressions (2.4) and (2.7) of $\widehat{I}_n$ and s_n^2 are unchanged when replacing K by K_μ. Since, by construction, $\widetilde{\mathbf{p}}_n(\mu) = \mathbf{0}$ and $\mathscr{E}_{K_\mu}(\mu) = 0$ (as $(\mathrm{Id}_{L^2} - \mathcal{P}_1)Z_x$ has no contribution to the integral of f), we directly obtain (4.12) and (4.13).

A further consequence is that the substitution of K_μ for K leaves the mean-squared error (2.2) unchanged, which yields a bound on the IMSPE of a design $\mathbf{X}_n$.

Theorem 4.5. *For K an SPD kernel and $\mu \in \mathscr{M}^+(1) \cap \mathscr{M}_K$, the IMSPE of an n-point design $\mathbf{X}_n$ satisfies*

$$\sigma^2\, s_n^2 \le \mathsf{IMSPE}(\mathbf{X}_n) \le \sigma^2 \left[s_n^2 + \int_{\mathscr{X}} K(\mathbf{x},\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}) - \mathscr{E}_K(\mu) \right]. \tag{4.14}$$

Proof. Replacing K by K_μ in (2.2), we get

$$\int_{\mathscr{X}} \rho_n^2(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}) = \int_{\mathscr{X}} K_\mu(\mathbf{x},\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}) + \frac{1}{\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1} \mathbf{1}_n} - \mathrm{trace}\left[\widetilde{\mathbf{K}}_n^{-1} \widetilde{\mathbf{Q}}_n^\perp \widetilde{\mathbf{H}}_n(\mu) \right],$$

where $\widetilde{\mathbf{H}}_n(\mu) = \int_{\mathscr{X}} \widetilde{\mathbf{k}}_n(\mathbf{x}) \widetilde{\mathbf{k}}_n^\top(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x})$, with $\widetilde{\mathbf{k}}_n(\mathbf{x}) = (K_\mu(\mathbf{x},\mathbf{x}_1), \ldots, K_\mu(\mathbf{x},\mathbf{x}_n))^\top$, and where $\widetilde{\mathbf{Q}}_n^\perp = \mathbf{I}_n - \mathbf{1}_n \mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1} / (\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1} \mathbf{1}_n)$, with $\mathbf{I}_n$ the n-dimensional identity matrix, is a projector onto the linear space orthogonal to $\mathbf{1}_n$. Since $\widetilde{\mathbf{K}}_n^{-1}$, $\widetilde{\mathbf{Q}}_n^\perp$ and $\widetilde{\mathbf{H}}_n(\mu)$ are non-negative definite, we obtain

$$\mathsf{IMSPE}(\mathbf{X}_n) = \sigma^2 \int_{\mathscr{X}} \rho_n^2(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}) \le \sigma^2 \left[\int_{\mathscr{X}} K_\mu(\mathbf{x},\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}) + \frac{1}{\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1} \mathbf{1}_n} \right].$$

Together with (3.15) and (4.13), this gives the right-hand side inequality in (4.14). The left-hand side inequality is a simple consequence of the convexity of $t \to t^2$. (Note that (2.10) implies that $\int_{\mathscr{X}} K(\mathbf{x},\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}) \ge \mathscr{E}_K(\mu)$.) ∎

One should notice that the upper bound in (4.14) may be rather loose for large d. For instance, when K is separable as in (3.18), with all K_i identical and $K_1(x,x) = 1$ for all x, and μ is uniform on $\mathscr{X} = [0,1]^d$, then $\int_{\mathscr{X}} K(\mathbf{x},\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}) - \mathscr{E}_K(\mu) = 1 - \mathscr{E}_{K_1}^d(\mu^{(1)})$, which can be close to one for large d.

Remark 4.4 (optimal weights for CISPD kernels)**.** Lagrangian theory indicates that the solution $\hat{\mathbf{w}}_n$ is obtained by solving the linear equation $\mathbf{R}_n (\hat{\mathbf{w}}_n^\top \ \lambda)^\top = (\mathbf{0}^\top \ 1)^\top$, where

$$\mathbf{R}_n = \begin{pmatrix} \widetilde{\mathbf{K}}_n & \mathbf{1}_n \\ \mathbf{1}_n^\top & 0 \end{pmatrix}.$$

When K is conditionally SPD, K_μ is conditionally SPD too, and the matrix $\mathbf{R}_n$ has full rank $n+1$. Indeed, $\mathbf{R}_n (\mathbf{z}_n^\top \ z)^\top = 0$ implies $\mathbf{1}_n^\top \mathbf{z}_n = 0$ and $\widetilde{\mathbf{K}}_n \mathbf{z}_n + z \mathbf{1}_n = 0$. Multiplying the second equation by $\mathbf{z}_n^\top$, we get $\mathbf{z}_n^\top \widetilde{\mathbf{K}}_n \mathbf{z}_n = 0$. Since K_μ is conditionally SPD, this is incompatible with $\mathbf{1}_n^\top \mathbf{z}_n = 0$ unless $\mathbf{z}_n = \mathbf{0}$ and $z = 0$. We obtain

$$\hat{\mathbf{w}}_n = \frac{(\widetilde{\mathbf{K}}_n + \mathbf{1}_n \mathbf{1}_n^\top)^{-1} \mathbf{1}_n}{\mathbf{1}_n^\top (\widetilde{\mathbf{K}}_n + \mathbf{1}_n \mathbf{1}_n^\top)^{-1} \mathbf{1}_n},$$

and $s_n^2 = \hat{\mathbf{w}}_n^\top \widetilde{\mathbf{K}}_n \hat{\mathbf{w}}_n = (\mathbf{1}_n^\top (\widetilde{\mathbf{K}}_n + \mathbf{1}_n \mathbf{1}_n^\top)^{-1} \mathbf{1}_n)^{-1} - 1$. When K is SPD and $\widetilde{\mathbf{K}}_n$ has full rank (Theorem 4.4), we recover $\hat{\mathbf{w}}_n = \widetilde{\mathbf{K}}_n^{-1} \mathbf{1}_n / (\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1} \mathbf{1}_n)$ and $\widehat{I}_n = \hat{\mathbf{w}}_n^\top \mathbf{y}_n$ given by (4.12). ◁

Remark 4.5 (IMSPE for separable kernels)**.** The use of a separable kernel (3.18) and a product measure $\mu = \otimes_{i=1}^{d} \mu^{(i)}$ on $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_d$ facilitates the calculations of $\widetilde{\mathbf{K}}_n$ and $\mathscr{E}_K(\xi_n - \mu)$, see (4.1), since $\mathscr{E}_K(\mu)$ and $P_\mu(\mathbf{x}_i)$ have the simple expressions (3.19, 3.20). The calculation of the IMSPE is facilitated too, but to a lesser extent. Indeed, using (2.2) we obtain

$$\int_{\mathscr{X}} \rho_n^2(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}) = \int_{\mathscr{X}} K(\mathbf{x},\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}) + \frac{1}{\mathbf{1}_n^\top \mathbf{K}_n^{-1}\mathbf{1}_n} - 2\,\frac{\mathbf{p}_n^\top(\mu)\mathbf{K}_n^{-1}\mathbf{1}_n}{\mathbf{1}_n^\top \mathbf{K}_n^{-1}\mathbf{1}_n} - \operatorname{trace}\left[\mathbf{K}_n^{-1}\mathbf{Q}_n^\perp \mathbf{H}_n(\mu)\right],$$

where $\mathbf{Q}_n^\perp = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^\top \mathbf{K}_n^{-1}/(\mathbf{1}_n^\top \mathbf{K}_n^{-1}\mathbf{1}_n)$ and

$$\{\mathbf{H}_n(\mu)\}_{j,k} = \int_{\mathscr{X}} K(\mathbf{x},\mathbf{x}_j)K(\mathbf{x},\mathbf{x}_k)\mathrm{d}\mu(\mathbf{x}) = \prod_{i=1}^{d}\int_{\mathscr{X}_i} K_i(x_i,x_{j\,i})K_i(x_i,x_{k\,i})\mathrm{d}\mu^{(i)}(x_i)\,. \triangleleft$$

4.3. One-shot designs. We consider two design constructions based on minimization of the posterior variance $\sigma^2\, s_n^2$: in the fist one, the design points are uniformly weighted; in the second one, they receive the optimal weights (4.7).

4.3.1. n**-point empirical measures.** Consider the empirical measure $\xi_{n,e} = (1/n)\sum_{i=1}^{n} \delta_{\mathbf{x}_i}$ associated with a given design $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. As indicated hereafter, the literature on space-filling design provides several examples of constructions of n-point designs through the minimization of the squared MMD $\mathscr{E}_K(\xi_{n,e} - \mu) = \mathscr{E}_{K_\mu}(\xi_{n,e})$ with respect to $\mathbf{X}_n$. Notice that $\mathscr{E}_K(\xi_{n,e} - \mu) = (\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n \mathbf{1}_n)/n^2$; see (4.11).

For $\mathscr{X} = [0,1]^d$, separable kernels based on variants of Brownian motion covariance yield L_2 discrepancies (symmetric, centred, wrap-around and so on); see, e.g., [47], [31, Chap. 3]. For instance, for $\mathscr{X} = [0,1]$ and $K(x,x') = 1 - |x - x'|$ (for which the expressions of $\mathscr{E}_K(\mu)$ and $P_\mu(x)$ are given in Table 3.1), $\mathscr{E}_{K_\mu}(\xi_{n,e})$ is twice the squared L_2 star discrepancy for $d = 1$.

The ISPD kernel $K_{s,\epsilon}^{\otimes}(\mathbf{x},\mathbf{x}') = \prod_{i=1}^{d} K_{s,\epsilon}(x_i,x_i')$, with $K_{s,\epsilon}$ given by (3.6) with $s > 0$ and $\epsilon > 0$, is called *projection kernel* in [62]. For very small ϵ, the minimization of $\mathscr{E}_{K_{1,\epsilon}^{\otimes}}(\xi_{n,e})$ corresponds to the construction of a maximum-projection design, as defined in [51]. Note that minimizing $\mathscr{E}_{K_{s,\epsilon}^{\otimes}}(\xi_{n,e})$ is not equivalent to minimizing $\mathscr{E}_{K_{s,\epsilon}^{\otimes}}(\xi_{n,e} - \mu)$: in particular, when μ is uniform on $\mathscr{X}$, which is assumed to be compact and convex, the former tends to push design points to the boundary of $\mathscr{X}$ whereas the latter keeps all points in the interior of $\mathscr{X}$; see [62].

In [64], space-filling designs in a compact set $\mathscr{X} \subset \mathbb{R}^d$ are constructed by minimizing $\mathscr{E}_{K^{(1)}}(\xi_{n,e} - \mu)$ for μ uniform on $\mathscr{X}$, see (3.7). The authors call *support points* the optimal support $\mathbf{X}_n^*$, which they determine via a majorization-minimization algorithm using the property that the problem can be formulated as a difference-of-convex optimization problem. Values of $\mathscr{E}_{K^{(1)}}(\mu)$ and $P_\mu(\mathbf{x})$ are not available even for $\mathscr{X} = [0,1]^d$ and Monte-Carlo approximation is used.

4.3.2. n**-point optimal measures.** Theorems 4.3 and 4.4 indicate that, if K is SPD, $\mathscr{E}_K(\hat{\xi}_n - \mu) = (\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n)^{-1}$ is the minimum value of $\mathscr{E}_K(\xi_n - \mu)$ for measures $\xi_n \in \mathscr{M}(1)$. Hence, we can construct space-filling designs on a compact and convex subset $\mathscr{X}$ of $\mathbb{R}^d$ by maximizing $\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n$ with respect to $\mathbf{X}_n \in \mathbb{R}^{nd}$, for a suitable K, taking μ uniform on $\mathscr{X}$. This can be performed using any unconstrained nonlinear programming algorithm, see the examples in Section 5. Note that, from the Cauchy-Schwarz inequality, $\mathscr{E}_K(\hat{\xi}_n - \mu) = (\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n)^{-1} \leq (\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n \mathbf{1}_n)/n^2 = \mathscr{E}_K(\xi_{n,e} - \mu)$.

4.4. Sequences of nested designs. There exist situations where the number n of design points ultimately used (for integration, or function approximation) differs from that initially planned, say N. It is the case in particular when function evaluations are computationally more expensive than expected, and numerical experimentation is stopped after $n < N$ simulations, or when simulations fail at some design points and testing at more than N points is required to obtain N valid evaluations in total. In such circumstances, it is convenient to have sequences of nested designs (extensible point sequences) at one's disposal. The objective is then to construct ordered sequences $\mathbf{x}_1, \mathbf{x}_2, \ldots$ of designs points such that any design $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ made of the first n points of the sequence has good space-filling properties. A typical example is given by Low Discrepancy Sequences (LDS) in $[0,1]^d$, see [69].

When K is SPD, we may exploit expression (4.13) of the conditional variance of $I_\mu(f)$ in a greedy sequential construction: at step n we choose $\mathbf{x}_{n+1}$ that minimizes s^2_{n+1}. This sequential construction, called Sequential Bayesian Quadrature in [15], is straightforward to implement compared with global minimization of s^2_n, see (2.7). Direct calculation, using formulae for the inversion of the block matrix

$$\widetilde{\mathbf{K}}_{n+1} = \begin{pmatrix} \widetilde{\mathbf{K}}_n & \widetilde{\mathbf{k}}_n(\mathbf{x}_{n+1}) \\ \widetilde{\mathbf{k}}_n^\top(\mathbf{x}_{n+1}) & K_\mu(\mathbf{x}_{n+1}, \mathbf{x}_{n+1}) \end{pmatrix},$$

where $\{\widetilde{\mathbf{K}}_n\}_{i,j} = K_\mu(\mathbf{x}_i, \mathbf{x}_j)$ and $\{\widetilde{\mathbf{k}}_n(\mathbf{x})\}_i = K_\mu(\mathbf{x}, \mathbf{x}_i)$, $i, j = 1, \ldots, n$, $\mathbf{x} \in \mathscr{X}$, gives

$$s^2_{n+1} = \left[\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1} \mathbf{1}_n + \frac{(1 - \widetilde{\mathbf{k}}_n(\mathbf{x}_{n+1})^\top \widetilde{\mathbf{K}}_n^{-1} \mathbf{1}_n)^2}{K_\mu(\mathbf{x}_{n+1}, \mathbf{x}_{n+1}) - \widetilde{\mathbf{k}}_n^\top(\mathbf{x}_{n+1}) \widetilde{\mathbf{K}}_n^{-1} \widetilde{\mathbf{k}}_n(\mathbf{x}_{n+1})} \right]^{-1}. \quad (4.15)$$

The sequential construction is thus

$$\mathbf{x}_{n+1} \in \operatorname{Arg\,max}_{\mathbf{x} \in \mathscr{X}} \frac{(1 - \widetilde{\mathbf{k}}_n(\mathbf{x})^\top \widetilde{\mathbf{K}}_n^{-1} \mathbf{1}_n)^2}{K_\mu(\mathbf{x}, \mathbf{x}) - \widetilde{\mathbf{k}}_n^\top(\mathbf{x}) \widetilde{\mathbf{K}}_n^{-1} \widetilde{\mathbf{k}}_n(\mathbf{x})}. \quad (4.16)$$

The conditional gradient algorithm of [33] yields a simpler construction, particularly well adapted to the situation (and also applicable when K is unbounded). It relies on the sequential selection of points that minimize the current directional derivative of $\mathscr{E}_K(\xi - \mu) = \mathscr{E}_{K_\mu}(\xi)$, with ξ supported on design points previously selected. The algorithm is initialized at a measure $\xi^{(n_0)}$ supported on $\mathbf{X}_{n_0} \in \mathscr{X}^{n_0}$ (with for instance $n_0 = 1$ and $\xi^{(1)} = \delta_{\mathbf{x}_1}$ for some $\mathbf{x}_1 \in \mathscr{X}$). Let $\xi^{(n)}$ denote the measure associated with the current design $\mathbf{X}_n$ of iteration n, with weights $w_i^{(n)}$, i.e., $\xi^{(n)} = \sum_{i=1}^n w_i^{(n)} \delta_{\mathbf{x}_i}$. The next design point is chosen in $\operatorname{Arg\,min}_{\mathbf{x} \in \mathscr{X}} F_{K_\mu}(\xi^{(n)}, \delta_{\mathbf{x}})$, with F_{K_μ} the directional derivative (3.12) (any minimizer can be selected in case there are several). Straightforward calculation using (3.12) gives $\mathbf{x}_{n+1} \in \operatorname{Arg\,min}_{\mathbf{s} \in \mathscr{X}} [P_{\xi^{(n)}}(\mathbf{x}) - P_\mu(\mathbf{x})]$, that is,

$$\mathbf{x}_{n+1} \in \operatorname{Arg\,min}_{\mathbf{x} \in \mathscr{X}} \left[\sum_{i=1}^n w_i^{(n)} K(\mathbf{x}, \mathbf{x}_i) - P_\mu(\mathbf{x}) \right]. \quad (4.17)$$

After choosing $\mathbf{x}_{n+1}$, the measure $\xi^{(n)}$ is updated into

$$\xi^{(n+1)} = (1 - \alpha_n) \xi^{(n)} + \alpha_n \delta_{\mathbf{x}_{n+1}} \quad (4.18)$$

for some $\alpha_n \in [0,1]$, so that $\xi^{(n+1)} \in \mathscr{M}^+(1)$ when $\xi^{(n)} \in \mathscr{M}^+(1)$. When $\xi^{(n_0)}$ is the empirical (uniform) measure on $\mathbf{X}_{n_0}$, the choice $\alpha_n = 1/(n+1)$ implies that $\xi^{(n)}$ remains uniform on its support $\mathbf{X}_n$ for all n, see [96] for an early contribution in the design context. The method is called *kernel herding* in the machine-learning literature, see [5, 19, 48]. It is shown in [19] that $\mathscr{E}_K(\xi^{(n)} - \mu) = \mathcal{O}(1/n^2)$ when $\mathcal{H}_K$ is finite dimensional, but we only have the weaker result $\mathscr{E}_K(\xi^{(n)} - \mu) = \mathcal{O}(1/n)$ when $\mathcal{H}_K$ is infinite dimensional, see [5]; see also Appendix A.

Remark 4.6. If we take $\mathbf{x}_{n+1} \in \operatorname{Arg\,min}_{\mathbf{x}\in\mathscr{X}} \sum_{i=1}^n w_i^{(n)} K(\mathbf{x}, \mathbf{x}_i)$ instead of (4.17), then the algorithm minimizes $\mathscr{E}_K(\xi)$; see Example 3.5 and Remark 3.2. The presence of $P_\mu(\mathbf{x})$ in (4.17) permits to keep the design points $\mathbf{x}_i$ in the interior of $\mathscr{X}$; see also [62]. The first point $\mathbf{x}_1$ can be chosen as a minimizer of $P_\mu(\mathbf{x})$ (that is, $\mathbf{1}_d/2$ when μ is uniform on $[0,1]^d$). The construction (4.17) is well-defined even if K is singular: in that case, it guarantees that all design points are different ($\mathbf{x}_i \neq \mathbf{x}_j$ for all i, j); the same is true for all one-dimensional canonical projections when K is the product of singular kernels. More generally, the addition of $\delta_{\mathbf{x}_{n+1}}$ to the current design measure creates a local maximum of the function $\sum_{i=1}^{n+1} w_i^{(n)} K(\mathbf{x}, \mathbf{x}_i) - P_\mu(\mathbf{x})$ in the neighborhood of $\mathbf{x}_{n+1}$, so that the next point $\mathbf{x}_{n+2}$ is chosen at some distance from all previous ones. The choice of an adequate kernel has therefore some importance: its decrease should be fast enough to ensure that points are well spread apart (the correlation length should be small enough when K corresponds to a correlation function); a translation-invariant kernel with bounded support leaves some arbitrariness in the choice of $\mathbf{X}_n$ until the union of the supports of the $K(\mathbf{x}_i, \cdot)$ covers $\mathscr{X}$, and is not necessarily suitable. We thus recommend using a (translation-invariant) kernel with unbounded support; if $n_{\max}$ is the maximum design size considered, a correlation function with correlation length $L \approx n_{\max}^{-1/d}$ is appropriate. Choosing K differentiable facilitates the minimization of $\mathscr{E}_K(\xi_{n,e} - \mu) = \mathbf{1}_n^\top \widetilde{\mathbf{K}}_n \mathbf{1}_n / n^2$ (Section 4.3.1), or the maximization of $\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1} \mathbf{1}_n$ (Section 4.3.2). ◁

Remark 4.7 (greedy MMD minimization)**.** Denote $\xi^{(n+)}(\mathbf{x}) = [n/(n+1)]\xi^{(n)} + [1/(n+1)]\delta_{\mathbf{x}}$. The direct minimization of $\mathscr{E}_K(\xi^{(n+)}(\mathbf{x}) - \mu)$ with respect to $\mathbf{x}$ yields

$$\mathbf{x}^{(n+1)} \in \operatorname*{Arg\,min}_{\mathbf{x}\in\mathscr{X}} \left[\frac{1}{n+1} \sum_{i=1}^n K(\mathbf{x}, \mathbf{x}_i) - P_\mu(\mathbf{x}) + \frac{1}{2(n+1)} K(\mathbf{x}, \mathbf{x}) \right], \quad (4.19)$$

that is, a selection very close to (4.17) when $K(\mathbf{x}, \mathbf{x})$ is constant (Matérn kernel for instance). Note that this construction requires $K(\mathbf{x}, \mathbf{x}) < \infty$ for all $\mathbf{x} \in \mathscr{X}$, contrary to (4.17). ◁

In practice n is always smaller than some given $n_{\max}$, and to facilitate the construction we can restrict the choice of the $\mathbf{x}_i$ to a finite subset $\mathscr{X}_\Omega = \{\mathbf{s}_1, \ldots, \mathbf{s}_\Omega\}$ of $\mathscr{X}$, with $\Omega \gg n_{\max}$ (when $\mathscr{X} = [0,1]^d$, $\mathscr{X}_\Omega$ can be given by a regular grid, or by the first Ω points of a LDS). For any $n \leq n_{\max}$, we can write $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} = \{\mathbf{s}_{i_1}, \ldots, \mathbf{s}_{i_n}\}$, the construction being initialized at some n_0-point design $\mathbf{X}_{n_0} \subset \mathscr{X}_\Omega$. A measure ξ supported on $\mathbf{X}_n$ can thus be written as $\xi = \sum_{i=1}^\Omega \omega_i \delta_{\mathbf{s}_i}$, with $\omega_i = 0$ when $\mathbf{s}_i \notin \mathbf{X}_n$. Therefore, for all n, $\xi^{(n)}$ is fully characterized by a Ω-dimensional vector $\boldsymbol{\omega}^{(n)} = (\omega_1^{(n)}, \ldots, \omega_\Omega^{(n)})^\top$, with $\boldsymbol{\omega}^{(n)}$ in the probability simplex $\mathbb{P}_\Omega$ when $\xi^{(n)} \in \mathscr{M}^+(1)$. The updating equations (4.17, 4.18) then imply that $\boldsymbol{\omega}^{(n+1)}$ is obtained by moving $\boldsymbol{\omega}^{(n)}$ in the direction of a vertex of $\mathbb{P}_\Omega$, hence the name *vertex-direction* given to methods based on (4.18) in the literature on optimal design, see, e.g., [75, Chap. 9] and the references therein. The cost of the determination of $\mathbf{x}_{n+1}$ in (4.17) is $\mathcal{O}(\Omega)$:

we compute $K(\mathbf{x}, \mathbf{x}_n)$ for all $\mathbf{x} \in \mathscr{X}_\Omega$ and update the sum $\sum_{i=1}^{n-1} K(\mathbf{x}, \mathbf{x}_i)$; the cost for n iterations scales as $\mathcal{O}(n\Omega)$, including the initial cost for the computation of $P_\mu(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{X}_\Omega$.

The measure $\xi^{(n)}$ constructed by (4.18) with $\alpha_n = 1/(n+1)$ is uniform on its support. The minimum-norm-point algorithm of [5] replaces $\xi^{(n)}$ by the measure having the same support but optimal weights in $\mathbb{P}_\Omega$. The strategy that consists in optimizing the weights on the current support of the design measure at each iteration is known to be efficient also in other design contexts, see Algorithm 1 in [77]. Here the weights are solution of a convex quadratic programming problem, which facilitates their determination. In the examples of Section 5, we consider a still simplified version where $\xi^{(n)}$ is replaced by $\hat{\xi}^{(n)}$ having weights $\hat{w}_i^{(n)}$ given by (4.7). This modification of $\xi^{(n)}$ at each iteration induces an additional computational cost of $\mathcal{O}(n^3)$ at iteration n ($\mathcal{O}(n^2)$ if rank-one updating is used to compute $\mathbf{K}_n^{-1}$) and requires the storage of all $K(\mathbf{x}, \mathbf{x}_i)$, $i = 1, \ldots, n$, $\mathbf{x} \in \mathscr{X}_\Omega$, in order to compute $\sum_{i=1}^n \hat{w}_i^{(n)} K(\mathbf{x}, \mathbf{x}_i)$. When K is a correlation function with small correlation length L and $\mathsf{PR}(\mathbf{X}_n) \gg L$, all design points have similar influence on $\mathscr{E}_K(\xi_n - \mu)$ and the associated optimal weights $\hat{w}_i^{(n)}$ are nearly identical: for n small enough, the resulting design is then similar to that obtained when $\xi^{(n)}$ is forced to be uniform.

We have also considered several variants of (4.18), where the step-size α_n is optimized instead of being fixed to $1/(n+1)$, or using a vertex-exchange method based on the true steepest–descent direction; see Appendix A for details: the performances, in terms of decrease of $\mathscr{E}_K(\xi^{(n)} - \mu)$ or in terms of space-filling properties of its support, $\mathsf{CR}(\mathbf{X}_n)$ and $\mathsf{PR}(\mathbf{X}_n)$, were not significantly better than those obtained with (4.18). The same observation holds for the Sequential Bayesian Quadrature (4.16) and the greedy MMD minimization (4.19).

Finally, note that the n-th design in a sequence of nested designs can be used as initialization for the (unconstrained) minimization of $\mathscr{E}_K(\xi_{n,e} - \mu) = \mathbf{1}_n^\top \widetilde{\mathbf{K}}_n \mathbf{1}_n / n^2$ (Section 4.3.1), or the maximization of $\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1} \mathbf{1}_n$ (Section 4.3.2), with respect to $\mathbf{X}_n$.

5. Illustrative examples.

5.1. Nested designs, $d = 2$. We take $\mathscr{X} = [0,1]^2$, μ is uniform on $\mathscr{X}$ and the candidate set $\mathscr{X}_\Omega$ is given by a regular 64×64 grid in $\mathscr{X}$; K is the product of uni-dimensional Matérn 3/2 covariance functions $K_{3/2,\theta}$, see (3.5). We consider nested designs of size up to $n_{\max} = 140$ and compare designs $\mathbf{X}_n^{\mathrm{VD}}$ generated by the vertex-direction method (4.17, 4.18) with more classical design sequences: $\mathbf{X}_n^{\mathrm{S}}$, given by the first n points of Sobol' LDS; $\mathbf{X}_n^{\mathrm{sS}}$ obtained by application of a random linear scramble, see [65]; and an extensible lattice sequence $\mathbf{X}_n^{\mathrm{EL}}$, where the n-th point is given by $\{n\mathbf{g}\}$, with $\mathbf{g}$ having irrational components independent over the rationals and $\{t\}$ denoting the fractional part, applied componentwise. Choosing a suitable generator $\mathbf{g}$ is a delicate matter; see, e.g., [55], [69, Chap. 5] and the references therein. We use the construction suggested in http://extremelearning.com.au/, with $\mathbf{g} = (1/\varphi_d, 1/\varphi_d^2, \ldots, 1/\varphi_d^d)^\top$ and φ_d the unique positive root of $x^{d+1} = x + 1$, which seems quite effective for small enough d. We initialize (4.17, 4.18) by $\mathbf{X}_1 = \{(0.5, 0.5)\}$ ($n_0 = 1$) and take $\theta = 10$. We also consider a variant of the minimum-norm-point algorithm of [5], where $\hat{\xi}^{(n)}$ having the optimal weights (4.7) is substituted for $\xi^{(n)}$ in (4.17, 4.18); we denote by $\mathbf{X}_n^{\mathrm{MN}}$ the corresponding designs. The choice $n_{\max} = 140$ is arbitrary; we take $n_{\max} > 2^7 = 128$ to be more fair with $\mathbf{S}_n$ which is known to have appealing properties when n is a power of two.

The four designs $\mathbf{X}_{140}^{\mathrm{VD}}$, $\mathbf{X}_{140}^{\mathrm{MN}}$, $\mathbf{X}_{140}^{\mathrm{sS}}$ and $\mathbf{X}_{140}^{\mathrm{EL}}$ are presented in Figure 5.1. Visually,

they are all reasonably space filling, with a slightly better behavior for $\mathbf{X}_{140}^{\mathrm{MN}}$ (top right) and $\mathbf{X}_{140}^{\mathrm{EL}}$ (bottom right); $\mathbf{X}_{140}^{\mathrm{S}}$ (not shown) has a few nearly coincident points that appear after $n = 110$ (this will be revealed by Figure 5.2-right).

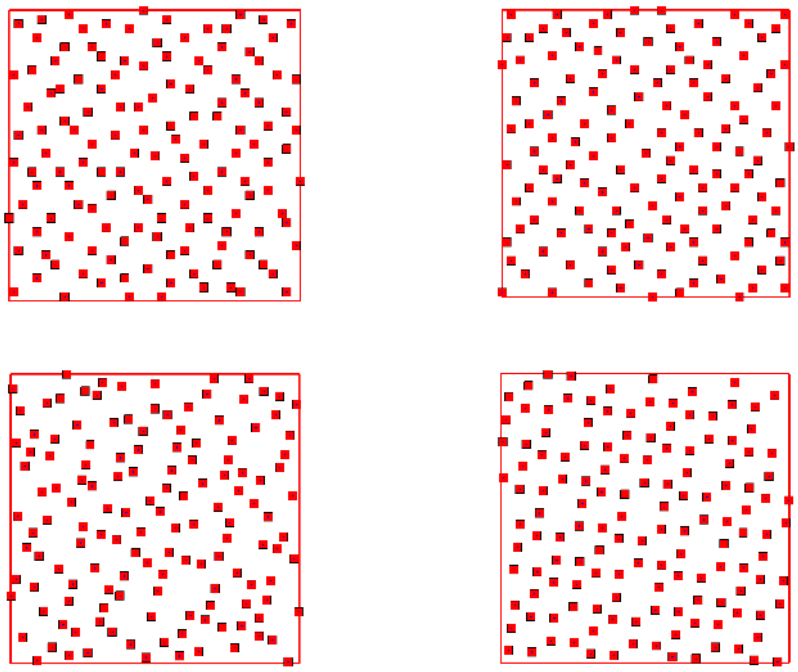

Figure 5.1: Top left: $\mathbf{X}_{140}^{\mathrm{VD}} = \mathrm{Supp}(\xi^{(140)})$ generated by (4.17, 4.18) with $\alpha_n = 1/(n+1)$ and $\theta = 10$ in $K_{3/2,\theta}$. Top right: $\mathbf{X}_{140}^{\mathrm{MN}} = \mathrm{Supp}(\hat{\xi}^{(140)})$ (minimum-norm-point variant where $\hat{\xi}^{(n)}$ with optimal weights is substituted for $\xi^{(n)}$ at each iteration). Bottom left: $\mathbf{X}_{140}^{\mathrm{sS}}$ (first 140 points of a scrambled Sobol' LDS). Bottom right: $\mathbf{X}_{140}^{\mathrm{EL}}$ (extensible lattice sequence).

Figure 5.2 shows the scaled values $n^{1/d}\,\mathsf{CR}(\mathbf{X}_n)$ (left, small values are preferred) and $n^{1/d}\,\mathsf{PR}(\mathbf{X}_n)$ (right, large values are preferred) for the five sequences of nested designs considered, $\mathbf{X}_n^{\mathrm{VD}}$, $\mathbf{X}_n^{\mathrm{MN}}$, $\mathbf{X}_n^{\mathrm{S}}$, $\mathbf{X}_n^{\mathrm{sS}}$ and $\mathbf{X}_n^{\mathrm{EL}}$, for $n = 2, \ldots, n_{\max} = 140$. The behavior of $\mathsf{CR}(\mathbf{X}_n^{\mathrm{S}})$ on the left panel illustrates the fact that Sobol' sequence has suitable space-filling properties for n equal to a power of two (notice the jump downwards at $n = 128$) but may perform rather poorly otherwise (and the situation deteriorates as d increases); $\mathsf{PR}(\mathbf{X}_n^{\mathrm{S}})$ on the right panel reveals the inclusion of nearly coincident points after $n = 110$. Overall, the scrambled sequence $\mathbf{X}_n^{\mathrm{sS}}$ performs better than $\mathbf{X}_n^{\mathrm{S}}$ but significantly worse than $\mathbf{X}_n^{\mathrm{VD}}$ and $\mathbf{X}_n^{\mathrm{MN}}$, both in terms of covering and packing radius. The extensible lattice sequence $\mathbf{X}_n^{\mathrm{EL}}$ performs slightly better than $\mathbf{X}_n^{\mathrm{MN}}$ in terms of packing radius, but $\mathsf{CR}(\mathbf{X}_n^{\mathrm{EL}})$ is significantly larger than $\mathsf{CR}(\mathbf{X}_n^{\mathrm{MN}})$ all along the sequence; $\mathbf{X}_n^{\mathrm{MN}}$ performs consistently better than $\mathbf{X}_n^{\mathrm{VD}}$ for both criteria.

The evolution of $\mathscr{E}_K(\xi_{n,e} - \mu)$ as a function of n, with $\xi_{n,e}$ the empirical measure associated with $\mathbf{X}_n$, is shown in Figure 5.3 for $\mathbf{X}_n^{\mathrm{VD}}$, $\mathbf{X}_n^{\mathrm{MN}}$ and $\mathbf{X}_n^{\mathrm{S}}$; $\mathscr{E}_K(\hat{\xi}_n - \mu)$, with $\hat{\xi}_n$ having optimal weights (4.7), is shown on the right panel, for the same designs (note that $\mathscr{E}_K(\hat{\xi}_n - \mu) \leq \mathscr{E}_K(\xi_{n,e} - \mu)$). $\mathbf{X}_n^{\mathrm{VD}}$ performs slightly better than $\mathbf{X}_n^{\mathrm{MN}}$ in

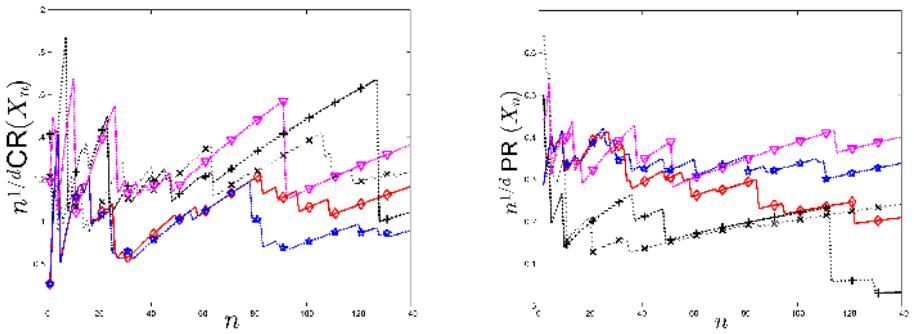

Figure 5.2: $n^{1/d}\,\mathsf{CR}(\mathbf{X}_n)$ (left) and $n^{1/d}\,\mathsf{PR}(\mathbf{X}_n)$ (right) for the designs $\mathbf{X}_n^{\mathrm{VD}}$ (red, diamonds), $\mathbf{X}_n^{\mathrm{MN}}$ (blue, stars), $\mathbf{X}_n^{\mathrm{S}}$ (black, pluses), $\mathbf{X}_n^{\mathrm{sS}}$ (black, x-marks) and $\mathbf{X}_n^{\mathrm{EL}}$ (magenta, triangles).

terms of $\mathscr{E}_K(\xi_{n,e}-\mu)$ but slightly worse in terms of $\mathscr{E}_K(\hat{\xi}_n-\mu)$; both perform better than $\mathbf{X}_n^{\mathrm{S}}$ for the two criteria. $\mathbf{X}_n^{\mathrm{sS}}$ (not shown) performs similarly to $\mathbf{X}_n^{\mathrm{S}}$; $\mathbf{X}_n^{\mathrm{EL}}$ (not shown) is between $\mathbf{X}_n^{\mathrm{S}}$ and $\mathbf{X}_n^{\mathrm{VD}}$. An important observation here is that designs may have significantly distinct space-filling properties (Figure 5.2) although they perform almost similarly in terms of MMD (Figure 5.3), in particular in terms of decrease rate. Figure 5.4 illustrates the fact that a faster decrease of MMD does not mean that the design points have a better distribution: there, $\theta = 1$ in $K_{3/2,\theta}$, so that MMD considers the integration of functions much smoother than previously when we had $\theta = 10$. The left panel shows $\mathscr{E}_K(\xi_{n,e}-\mu)$ for $\mathbf{X}_n^{\mathrm{VD}}$ and $\mathbf{X}_n^{\mathrm{S}}$. On the one hand, both constructions yield a much faster decrease of MMD than on the left panel of Figure 5.3, with a significantly smaller MMD for $\mathbf{X}_n^{\mathrm{VD}}$ than for $\mathbf{X}_n^{\mathrm{S}}$. On the other hand, the design $\mathbf{X}_{140}^{\mathrm{VD}}$ shown on the right panel of the figure has very poor space-filling properties; compare with the top-left panel of Figure 5.2.

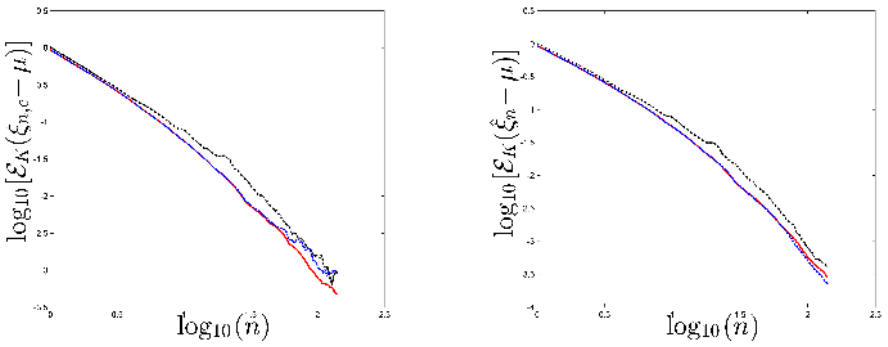

Figure 5.3: $\mathscr{E}_K(\xi_{n,e}-\mu)$ (left) and $\mathscr{E}_K(\hat{\xi}_n-\mu)$ (right) as functions of n (log scale), with $\xi_{n,e}$ the empirical measure associated with $\mathbf{X}_n$ and $\hat{\xi}_n$ the measure with optimal weights $\hat{\mathbf{w}}_n$ (4.7), for $\mathbf{X}_n^{\mathrm{VD}}$ (red solid line), $\mathbf{X}_n^{\mathrm{MN}}$ (blue dashed line) and $\mathbf{X}_n^{\mathrm{S}}$ (black dotted line); $\theta = 10$ in $K_{3/2,\theta}$.

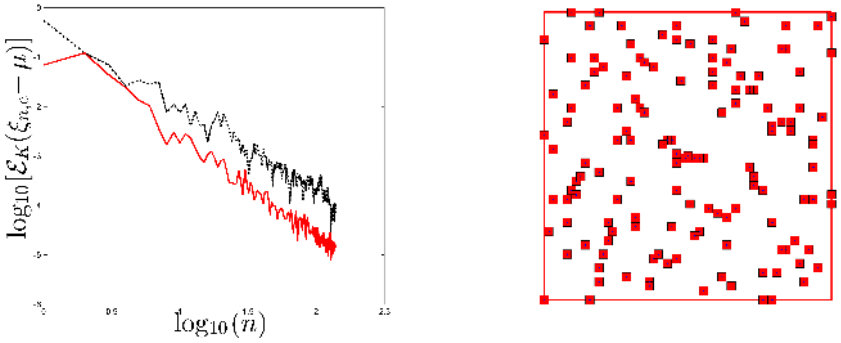

Figure 5.4: Left: $\mathscr{E}_K(\xi_{n,e}-\mu)$ as a function of n (log scale), with $\xi_{n,e}$ the empirical measure associated with $\mathbf{X}_n^{\mathrm{VD}}$ (red solid line) and $\mathbf{X}_n^{\mathrm{S}}$ (black dotted line); right: $\mathbf{X}_{140}^{\mathrm{VD}}$; $\theta = 1$ in $K_{3/2,\theta}$.

We construct now designs $\mathbf{X}_n^{\mathrm{VD}}$ and $\mathbf{X}_n^{\mathrm{MN}}$ for different values of θ in $K_{3/2,\theta}$. Figure 5.5 top-left (respectively, top-right) shows $\mathsf{CR}(\mathbf{X}_n^{\mathrm{VD}})$ (respectively, $\mathsf{PR}(\mathbf{X}_n^{\mathrm{VD}})$) for $\theta = 1$ (dashed line), $\theta = 1/n_{\max}^{1/d} \simeq 11.8$ (black dash-dotted line) and $\theta = 50$ (solid line); the coloured region is the envelope of the curves obtained when $\theta = 1, 5, 10, 15, \ldots, 50$. The second row gives the same information for the design $\mathbf{X}_n^{\mathrm{MN}}$. One may notice that the curves in solid line coincide on the two rows: when $K_{3/2,\theta}$ has a small correlation length (θ is large), the two constructions give similar designs since the optimal weights in $\hat{\xi}^{(n)}$ are nearly all equal to $1/n$ at each n ($\max_i\{\hat{w}_i^{(n)}\} - \min_i\{\hat{w}_i^{(n)}\} < 3 \times 10^{-4}$ for $n = 140$ and $\theta = 50$). Weight optimization makes the construction less sensitive to the choice of θ: in some sense it permits to compensate for the loss in space-filling performance incurred by choosing a kernel with excessively large correlation length (θ is too small) and forcing all weights to be equal during the construction. The choice $\theta = n_{\max}^{1/d} \simeq 11.8$ appears to yield good performance, with larger θ leading to larger packing radii but worse covering behavior.

5.2. One-shot designs, $d = 2, 3$. Here we use the n-th design in a sequence of nested designs to initialize the search for an MMD design. In [64], the MMD associated with energy distance (kernel (3.7)) is minimized with a Majorization-Minimization (MM) algorithm; the corresponding optimal designs are called support points, denoted by $\mathbf{X}_n^{\mathrm{supp}}$ is what follows. Using the explicit form of the kernel $K_{3/2,\theta}(\cdot,\cdot)$ and potential $P_\mu(\cdot)$, we can also construct a convex majorant for the (squared) MMD (4.1), for any set of weights $\mathbf{w}_n$. An MM algorithm can then be used to directly minimize $\mathscr{E}_K(\xi_{n,e} - \mu)$ ($\xi_{n,e}$ having all weights equal to $1/n$), or $\mathscr{E}_K(\hat{\xi}_n - \mu) = (\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1} \mathbf{1}_n)^{-1}$, with respect to $\mathbf{X}_n$. In the second case, we alternate MM steps for the minimization with respect to $\mathbf{X}_n$ with fixed weights, and (explicit) weight optimization through (4.7). We denote by $\mathbf{X}_n^{\mathrm{MM\text{–}MMD}}$ the corresponding designs. Alternatively, we can also directly minimize $\mathscr{E}_K(\xi_{n,e} - \mu)$, or maximize $\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1} \mathbf{1}_n$, with respect to $\mathbf{X}_n$ using any nonlinear programming algorithm. Note that we do not need to impose the constraints $\mathbf{x}_i \in \mathscr{X}$ for all i thanks to the presence of potentials $P_\mu(\mathbf{x}_i)$ in (4.1) acting as penalty functions. Since derivatives are available, we use Conjugate Gradient (CG); the corresponding designs are denoted $\mathbf{X}_n^{\mathrm{CG\text{–}MMD}}$. MM and CG only give locally optimal solutions, which therefore depend on the initialization. Table 5.1 gives an illustration for $n = 100$, $d = 2, 3$ and K the product of Matérn 3/2 covariance functions

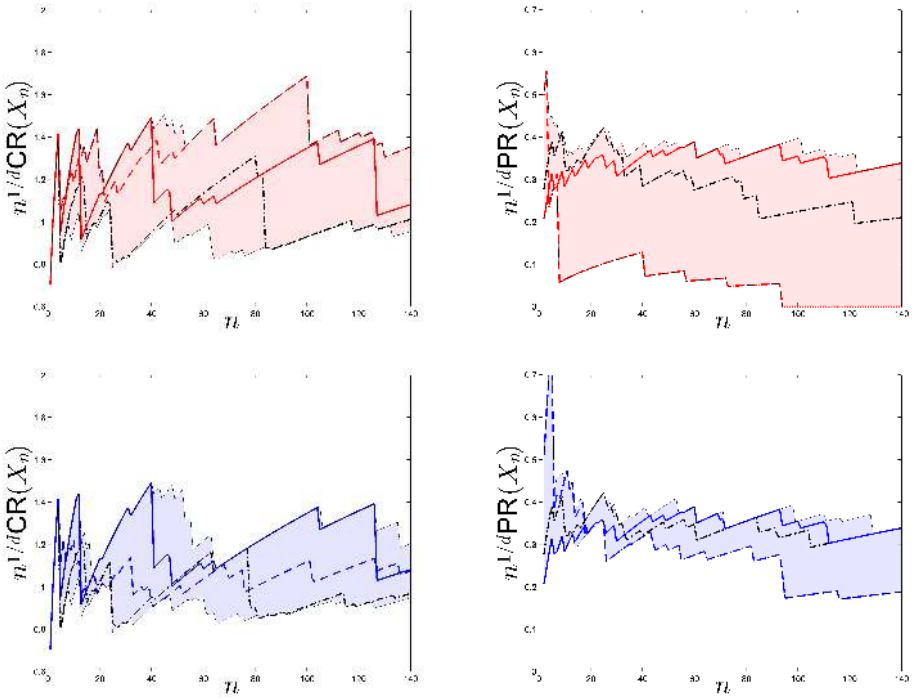

Figure 5.5: $n^{1/d}\,\mathsf{CR}(\mathbf{X}_n)$ (left column) and $n^{1/d}\,\mathsf{PR}(\mathbf{X}_n)$ (right column) as functions of n, for $\mathbf{X}_n^{\mathrm{VD}}$ (top row) and $\mathbf{X}_n^{\mathrm{MN}}$ (bottom row), when θ varies between 1 and 50 in $K_{3/2,\theta}$: $\theta = 1$ in dashed line, $\theta = 1/n_{\max}^{1/d}$ in dashed-dotted line, $\theta = 50$ in solid line.

$K_{3/2,\theta}$ with $\theta = n^{1/d}$. $\mathbf{X}_{100}^{\mathrm{sS}}$ corresponds to the first 100 points of a scrambled Sobol' sequence, $\mathbf{X}_{100}^{\mathrm{VD}}$ is obtained with (4.17, 4.18) and is used to initialize the optimization for the other designs in the table: the minimization of $\mathscr{E}_K(\xi_{n,e} - \mu)$ yields $\mathbf{X}_{100,e}^{\mathrm{MM\text{–}MMD}}$ and $\mathbf{X}_{100,e}^{\mathrm{CG\text{–}MMD}}$, the maximization of $\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1} \mathbf{1}_n$ yields $\mathbf{X}_{100}^{\mathrm{MM\text{–}MMD}}$ and $\mathbf{X}_{100}^{\mathrm{CG\text{–}MMD}}$. Initialization at $\mathbf{X}_{100}^{\mathrm{sS}}$ gives designs with worse covering and packing performances for all constructions considered. Table 5.1 shows the very good space-filling performance of designs based on MMD minimization for the energy distance (the support points of [64]), but the construction is computationally more demanding in high dimensions due to the necessity to approximate μ by a discrete measure (we use a 64×64 and a $16 \times 16 \times 16$ regular grid for $d = 2, 3$ respectively, with thus 4,096 points in both cases).

Table 5.1: Covering and packing performances of various fixed-size designs (d=2,3; n=100).

		$\mathbf{X}_{100}^{\mathrm{sS}}$	$\mathbf{X}_{100}^{\mathrm{VD}}$	$\mathbf{X}_{100}^{\mathrm{supp}}$	$\mathbf{X}_{100,e}^{\mathrm{MM\text{–}MMD}}$	$\mathbf{X}_{100}^{\mathrm{MM\text{–}MMD}}$	$\mathbf{X}_{100,e}^{\mathrm{CG\text{–}MMD}}$	$\mathbf{X}_{100}^{\mathrm{CG\text{–}MMD}}$
$d = 2$	CR	0.1377	0.0925	0.0839	0.0889	0.0874	0.0889	0.0845
	PR	0.0204	0.0262	0.0419	0.0365	0.0238	0.0369	0.0302
$d = 3$	CR	0.3054	0.2645	0.2032	0.2673	0.2624	0.2681	0.2886
	PR	0.0415	0.0751	0.1042	0.0886	0.0896	0.0938	0.0922

5.3. $d = 10$**.** In this section, we replace $\mathsf{CR}(\mathbf{X}_n)$ by its under approximation $\max_{\mathbf{x}\in\mathscr{X}_Q} \min_{1\le i\le n} \|\mathbf{x}-\mathbf{x}_i\|$, where $\mathscr{X}_Q$ corresponds to the first 2^{19} points of a scrambled Sobol' sequence complemented with a 3^d full factorial design. On purpose, we choose a candidate set for kernel herding having the same size as above, despite $d = 10$: $\mathscr{X}_\Omega$ is given by 4,096 points of a scrambled Sobol' sequence in $[0,1]^d$. In general, enlarging Ω improves the performance of MMD-based designs.

Figure 5.6 is similar to Figure 5.2 and shows the scaled values $n^{1/d}\,\mathsf{CR}(\mathbf{X}_n)$ (left) and $n^{1/d}\,\mathsf{PR}(\mathbf{X}_n)$ (right) for a scrambled Sobol' sequence $\mathbf{X}_n^{\mathrm{sS}}$ and three sequences of nested designs for $n = 1,\dots,n_{\max} = 100$: $\mathbf{X}_n^{\mathrm{VD\text{–}log}}$, $\mathbf{X}_n^{\mathrm{VD\text{–}M}}$ and $\mathbf{X}_n^{\mathrm{MN\text{–}M}}$. $\mathbf{X}_n^{\mathrm{VD\text{–}log}}$ is generated by (4.17, 4.18) with K the product of the uni-dimensional (singular) logarithmic kernel $K_{(0)}$ in (3.8); $\mathbf{X}_n^{\mathrm{VD\text{–}M}}$ is generated by the same vertex-direction method but for the product of Matérn kernels $K_{3/2,\theta}$, with $\theta = n_{\max}^{1/d}$, $\mathbf{X}_n^{\mathrm{MN\text{–}M}}$ is for the minimum-norm-point algorithm with the same kernel.

The three MMD related nested designs perform significantly better than the scrambled Sobol' sequence; like we observed in smaller dimensions, $\mathbf{X}_n^{\mathrm{MN\text{–}M}}$ is performing consistently better than $\mathbf{X}_n^{\mathrm{VD\text{–}M}}$; $\mathbf{X}_n^{\mathrm{VD\text{–}log}}$ performs slightly worse than $\mathbf{X}_n^{\mathrm{VD\text{–}M}}$ but has the advantage of not requiring the tuning of a length-scale parameter θ. Note that, since the computational cost only scales as $\mathcal{O}(n\Omega)$, one can easily generate many designs, for different kernels, different correlation length parameters, or different candidate sets $\mathscr{X}_\Omega$, and then select the best one according to the values of a particular criterion of interest over a particular range of design sizes.

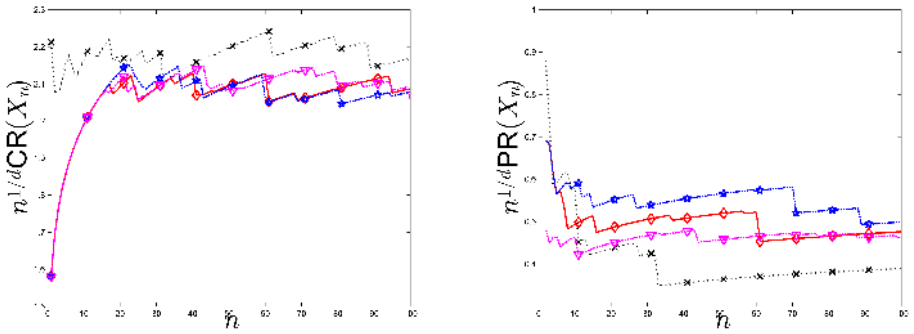

Figure 5.6: $n^{1/d}\,\mathsf{CR}(\mathbf{X}_n)$ (left) and $n^{1/d}\,\mathsf{PR}(\mathbf{X}_n)$ (right) for the designs $\mathbf{X}_n^{\mathrm{sS}}$ (black, x-marks), $\mathbf{X}_n^{\mathrm{VD\text{–}log}}$ (magenta, triangles), $\mathbf{X}_n^{\mathrm{VD\text{–}M}}$ (red, diamonds) and $\mathbf{X}_n^{\mathrm{MN\text{–}M}}$ (blue, stars).

Table 5.2 presents the covering and packing radii for different designs with $n = 100$. $\mathbf{X}_{100}^{\mathrm{sS}}$, $\mathbf{X}_{100}^{\mathrm{VD\text{–}log}}$ and $\mathbf{X}_{100}^{\mathrm{VD\text{–}M}}$ are like in Figure 5.6; $\mathbf{X}_{100}^{\mathrm{supp}}$ correspond to support points obtained by minimizing the MMD associated with energy distance (μ is approximated by the uniform measure on a 4^d full-factorial design), $\mathbf{X}_{100,e}^{\mathrm{CG\text{–}MMD}}$ and $\mathbf{X}_{100}^{\mathrm{CG\text{–}MMD}}$ respectively minimize $\mathscr{E}_K(\xi_{n,e}-\mu)$ and maximize $\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n$; these three optimizations are initialized at $\mathbf{X}_{100}^{\mathrm{VD\text{–}M}}$. $\mathbf{X}_{100}^{\mathrm{Lh}}$ is a maximin-optimal Latin hypercube design (downloaded from https://spacefillingdesigns.nl/) and $\mathbf{X}_{100}^{\mathrm{Lh\text{–}supp}}$ is obtained by minimizing the MMD associated with energy distance, with initialization at $\mathbf{X}_{100}^{\mathrm{Lh}}$. The three designs $\mathbf{X}_{100}^{\mathrm{supp}}$, $\mathbf{X}_{100,e}^{\mathrm{CG\text{–}MMD}}$ and $\mathbf{X}_{100}^{\mathrm{CG\text{–}MMD}}$ improve $\mathbf{X}_{100}^{\mathrm{VD\text{–}M}}$ both in terms of covering and packing radius. The much better performance of $\mathbf{X}_{100}^{\mathrm{Lh\text{–}supp}}$ compared with $\mathbf{X}_{100}^{\mathrm{supp}}$ illustrates the importance of a good initialization for MMD minimization.

Notice that $\mathsf{CR}(\mathbf{X}_{100}^{\text{Lh–supp}}) > \mathsf{CR}(\mathbf{X}_{100}^{\text{Lh}})$, whereas $\mathsf{PR}(\mathbf{X}_{100}^{\text{Lh–supp}}) > \mathsf{PR}(\mathbf{X}_{100}^{\text{Lh}})$.

Table 5.2: Covering and packing performances of various fixed-size designs ($d=10$; $n=100$).

	$\mathbf{X}_{100}^{\text{sS}}$	$\mathbf{X}_{100}^{\text{VD–log}}$	$\mathbf{X}_{100}^{\text{VD–M}}$	$\mathbf{X}_{100}^{\text{supp}}$	$\mathbf{X}_{100,e}^{\text{CG–MMD}}$	$\mathbf{X}_{100}^{\text{CG–MMD}}$	$\mathbf{X}_{100}^{\text{Lh}}$	$\mathbf{X}_{100}^{\text{Lh–supp}}$
CR	1.3684	1.2990	1.3168	1.3083	1.2594	1.2893	1.2515	1.2762
PR	0.2456	0.2921	0.3004	0.5332	0.3993	0.4109	0.5109	0.6337

For any $d' \in \{1,\dots,d\}$ and any $r \in \{1,\dots,\binom{d}{d'}\}$, let $P_{d',r}$ denote one of the $\binom{d}{d'}$ distinct projections on an axis-aligned d' dimensional sub-space. The following criteria measure the worst-case projection performance of a design in terms of its covering and packing radii in dimension d':

$$\mathsf{CR}_{d'}(\mathbf{X}_n) = \max_{r=1,\dots,\binom{d}{d'}} \max_{\mathbf{x}\in[-1,1]^{d'}} \|\mathbf{x}, P_{d',r}(\mathbf{X}_n)\|,$$

$$\mathsf{PR}_{d'}(\mathbf{X}_n) = \frac{1}{2} \min_{r=1,\dots,\binom{d}{d'}} \min_{i\neq j} \|P_{d',r}(\mathbf{x}_i) - P_{d',r}(\mathbf{x}_j)\|.$$

Figure 5.7 shows the ratios $\mathsf{CR}_{d'}(\mathbf{X}_{100})/\,\mathsf{CR}_{d'}(\mathbf{X}_{100}^{\text{Lh}})$ and $\mathsf{PR}_{d'}(\mathbf{X}_{100})/\,\mathsf{PR}_{d'}(\mathbf{X}_{100}^{\text{Lh}})$ for $\mathbf{X}_{100} = \mathbf{X}_{100}^{\text{sS}}$, $\mathbf{X}_{100}^{\text{VD–log}}$, $\mathbf{X}_{100}^{\text{supp}}$, $\mathbf{X}_{100,e}^{\text{CG–MMD}}$, $\mathbf{X}_{100}^{\text{Lh–supp}}$ and $d' = 2,\dots,10$.

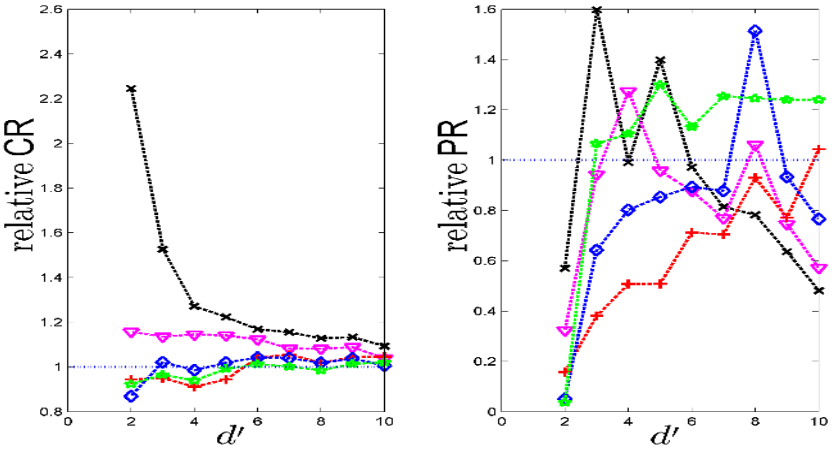

Figure 5.7: Relative performances $\mathsf{CR}_{d'}(\mathbf{X}_{100})/\,\mathsf{CR}_{d'}(\mathbf{X}_{100}^{\text{Lh}})$ (left) and $\mathsf{PR}_{d'}(\mathbf{X}_{100})/\,\mathsf{PR}_{d'}(\mathbf{X}_{100}^{\text{Lh}})$ (right) for $\mathbf{X}_{100}^{\text{sS}}$ (black, x-marks), $\mathbf{X}_{100}^{\text{VD–log}}$ (magenta, triangles), $\mathbf{X}_{100}^{\text{supp}}$ (red, pluses), $\mathbf{X}_{100,e}^{\text{CG–MMD}}$ (blue, diamonds) and $\mathbf{X}_{100}^{\text{Lh–supp}}$ (green, stars).

Although the size of the candidate set used for the construction of $\mathbf{X}_{100}^{\text{VD–log}}$ is very small ($\Omega = 4,096$) relative to the dimension ($d = 10$), this sequentially constructed design performs significantly better than the scrambled Sobol' sequence also in projections to smaller dimensions. The one-shot designs $\mathbf{X}_{100}^{\text{supp}}$ and $\mathbf{X}_{100,e}^{\text{CG–MMD}}$ are only slightly worse than the (best available) Latin hypercube design $\mathbf{X}_{100}^{\text{Lh}}$. MMD minimization for the kernel (3.7) associated with energy-distance, initialized at $\mathbf{X}_{100}^{\text{Lh}}$, yields the overall best design $\mathbf{X}_{100}^{\text{Lh–supp}}$ among those considered.

6. Conclusion. Optimal designs for Bayesian integration of an unknown function considered as a realization of a Gaussian RF with covariance K, with respect to a measure μ, that minimize the posterior integration variance, are also optimal designs for the BLUE in a location model with correlated errors, with their correlation kernel depending on K and μ, and minimize the MMD, a kernel discrepancy to μ. The fact that this squared discrepancy takes the form of a quadratic energy, depending on K, for the difference between μ and the design measure, permits to use all the classical machinery of optimal design, including theory (convexity, directional derivatives, optimality theorems) and algorithms. When μ is uniform, MMD minimization appears to be a natural way of constructing space-filling designs: the quadratic form of the criterion makes the algorithms simple and intuitive; one-step-ahead constructions allow the fast generation of sequences of nested designs with good properties for any size.

For μ uniform, the space-filling properties of designs obtained by MMD minimization depend on the choice of the kernel K. The paper has focused on two classical space-filling characteristics, the covering and packing radii. Finding the most suitable kernel for any of these characteristics remains an open issue. For instance, the last two columns of Table 5.2 indicate that the isotropic kernel (3.7) associated with energy-distance favours packing on expense of covering. On the one hand, separable translation-invariant kernels peaked enough at the origin ensure that designs points are well spread in all projections. Singular kernels, which do not define RKHS and present significant theoretical challenges, also have great potential in this respect (see [79]). On the other hand, support points that minimize MMD for a particular isotropic kernel provided best results for the whole d-dimensional set in our numerical examples.

Our main intention with this paper is to promote the general use of MMD minimization for the construction of space-filling designs. We hope that the stimulating connections between Bayesian integration and other areas, such as potential theory and BLUE, will be of general interest and will attract attention to this type of design approaches.

Acknowledgements. The first author thanks the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme UQ for inverse problems in complex systems where work on this paper was partly undertaken (EPSRC grant no EP/K032208/1). Part of this work was supported by the project INDEX (INcremental Design of EXperiments) ANR-18-CE91-0007 of the French National Research Agency (ANR). We gratefully acknowledge Sébastien Da Veiga (Safran, Paris) for drawing our attention to the machine learning literature on kernel herding and to connection with minimization of L_2 discrepancy. Discussions with Nicolas Durrande (PROWLER.io, Cambridge, UK) on the RKHS of Matérn kernels were very helpful. We are also very grateful to the two anonymous referees for their careful reading and their numerous comments and suggestions that helped us to improve the paper.

Appendix A. Some convergence properties of conditional gradient algorithms.

We consider a conditional gradient algorithm with iterations given by (4.18). K is a bounded ISPD kernel (and is thus SPD); in contrast with [15], we do not assume that $\mathcal{H}_K$ is finite dimensional. In the context of MMD minimization, the criterion is quadratic, which facilitates the developments to follow, but the results presented

are more general than that and rely mainly on convexity. We focus our attention on the case when $\mathscr{X}$ is replaced by a finite set $\mathscr{X}_\Omega = \{\mathbf{s}_1, \ldots, \mathbf{s}_\Omega\}$, so that a probability measure ξ on $\mathscr{X}_\Omega$ is characterized by a vector of weights $\boldsymbol{\omega}$ in the probability simplex $\mathbb{P}_\Omega$. However, after the proof of Theorem A.1 we indicate why a similar analysis applies to the infinite-dimensional situation.

Denote $J_K(\boldsymbol{\omega}) = \|\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}\|^2_{\mathbf{K}} = (\boldsymbol{\omega} - \hat{\boldsymbol{\omega}})^\top \mathbf{K}(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}})$, with $\mathbf{K} = \mathbf{K}_\Omega$ a non-negative definite $\Omega \times \Omega$ matrix and $\hat{\boldsymbol{\omega}}$ in $\mathbb{P}_\Omega$ (and $\hat{\boldsymbol{\omega}} = \mathbf{1}_\Omega$ when the target measure is uniform on $\mathscr{X}_\Omega$). Denote by B_Ω an upper bound on $\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|^2_{\mathbf{K}}$ for $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$ in $\mathbb{P}_\Omega$. Denoting by $\lambda_{\max}(\mathbf{K})$ the largest eigenvalue of $\mathbf{K}$, we can always take $B_\Omega = 2\,\lambda_{\max}(\mathbf{K})$, the bound used in the developments below. When $K(\mathbf{x}, \mathbf{x}) = 1$ and $K(\mathbf{x}, \mathbf{x}') \geq 0$ for all $\mathbf{x}, \mathbf{x}'$, we can take $B_\Omega = 2$ (and replace $\lambda_{\max}(\mathbf{K})$ by 1 in Theorems A.1, A.2 and A.3).

For $i = 1, \ldots, \Omega$, we denote by $\mathbf{e}_i$ the i-th basis vector, with component number i equal to one. Iteration (4.18) has the form

$$\boldsymbol{\omega}^{(n+1)} = \boldsymbol{\omega}^{(n)} + \alpha_n \Delta_n$$

for some step-size α_n and direction $\Delta_n = \mathbf{e}_{i_n^+} - \boldsymbol{\omega}^{(n)}$, with the index i_n^+ taken in $\operatorname{Arg\,min}_{i=1,\ldots,\Omega} \mathbf{e}_i^\top \nabla J_K(\boldsymbol{\omega}^{(n)})$, where the gradient $\nabla J_K(\boldsymbol{\omega})$ is given by

$$\nabla J_K(\boldsymbol{\omega}) = 2\mathbf{K}(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}})\,.$$

This is equivalent to $\mathbf{s}_{i_n^+} \in \operatorname{Arg\,min}_{\mathbf{s} \in \mathscr{X}_\Omega} \big[P_{\xi^{(n)}}(\mathbf{s}) - P_\mu(\mathbf{s})\big]$, see (4.17).

A.1. Vertex-direction, predefined step-size. Take $\alpha_n = 1/(n+1)$ in (4.18). We first mention a simple result indicating that $\|\boldsymbol{\omega}^{(n)} - \mathbf{1}_\Omega/\Omega\|^2_{\mathbf{K}} = \mathcal{O}(1/n)$ during the initial $n_1 \leq \Omega$ iterations when all i_n^+ are distinct for $n \leq n_1$.

Lemma A.1. *Algorithm (4.18) with $\alpha_n = 1/(n+1)$, initialized at $\boldsymbol{\omega}^{(1)} = \mathbf{e}_{i_0}$ for some $i_0 \in \{1, \ldots, \Omega\}$, satisfies*

$$\|\boldsymbol{\omega}^{(n)} - \mathbf{1}_\Omega/\Omega\|^2_{\mathbf{K}} \leq \frac{\lambda_{\max}(\mathbf{K})}{n}\,, \quad 1 \leq n \leq n_1 \leq \Omega\,,$$

where n_1 is such that all i_n^+ are distinct for $n \leq n_1$.

Proof. For $n \leq n_1$, after a suitable reordering of indices we have $\boldsymbol{\omega}^{(n)} = (1/n, \ldots, 1/n, 0 \ldots, 0)^\top$. Therefore, $\|\boldsymbol{\omega}^{(n)} - \mathbf{1}_\Omega/\Omega\|^2_{\mathbf{K}} \leq \lambda_{\max}(\mathbf{K})\|\boldsymbol{\omega}^{(n)} - \mathbf{1}_\Omega/\Omega\|^2 = \lambda_{\max}(\mathbf{K})\,(\Omega - n)/(n\Omega) \leq \lambda_{\max}(\mathbf{K})/n$. ∎

Note that this property is independent of the order in which the vertices of $\mathbb{P}_\Omega$ (the $\mathbf{e}_{i_n}$) are selected. It is therefore also valid for MC sampling without replacement within $\mathscr{X}_\Omega$. Also note that the optimal step-size $\hat{\alpha}_n$ at iteration n for the minimization of $\|\boldsymbol{\omega}^{(n)} - \mathbf{1}_\Omega/\Omega\|^2$ equals $\alpha_n = 1/(n+1)$.

The following lemma shows that (4.18) with $\alpha_n = 1/(n+1)$ ensures that $\|\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}\|^2_{\mathbf{K}} = \mathcal{O}(\log n/n)$, independently of Ω and of the positions of $\boldsymbol{\omega}^{(1)}$ and $\hat{\boldsymbol{\omega}}$ in $\mathbb{P}_\Omega$.

Theorem A.1. *Algorithm (4.18) with $\alpha_n = 1/(n+1)$, initialized at any $\boldsymbol{\omega}^{(1)}$ in $\mathbb{P}_\Omega$, satisfies*

$$\|\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}\|^2_{\mathbf{K}} \leq 2\,\lambda_{\max}(\mathbf{K})\,\frac{1 + 2\,\log(n+1)}{n}\,, \quad n \geq 1\,. \tag{A.1}$$

Proof. The proof follows the same lines as in [20, Sect. 3]. Denote $g(\boldsymbol{\omega}) = \|\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}\|^2_{\mathbf{K}}$ and $\boldsymbol{\omega}^{(n+)}(\alpha) = \boldsymbol{\omega}^{(n)} + \alpha\Delta_n$. Notice that $\boldsymbol{\omega}^{(n)} \in \mathbb{P}_\Omega$ for all $n \geq 1$. We have

$$\begin{aligned} g[\boldsymbol{\omega}^{(n+)}(\alpha)] &= g(\boldsymbol{\omega}^{(n)}) + 2\alpha\Delta_n^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}) + \alpha^2\|\Delta_n\|^2_{\mathbf{K}} \\ &\leq g(\boldsymbol{\omega}^{(n)}) + 2\alpha\Delta_n^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}) + \alpha^2\lambda_{\max}(\mathbf{K})\|\Delta_n\|^2 \\ &\leq g(\boldsymbol{\omega}^{(n)}) + 2\alpha\Delta_n^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}) + 2\,\alpha^2\lambda_{\max}(\mathbf{K})\,. \end{aligned}$$

The convexity of $g(\cdot)$ and the definition of Δ_n imply that

$$g(\boldsymbol{\omega}^{(n)}) \geq g(\hat{\boldsymbol{\omega}}) = 0 \geq g(\boldsymbol{\omega}^{(n)}) + (\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}^{(n)})^\top \nabla J_K(\boldsymbol{\omega}^{(n)}) \geq g(\boldsymbol{\omega}^{(n)}) + \Delta_n^\top \nabla J_K(\boldsymbol{\omega}^{(n)}) \,.$$

Therefore, $\Delta_n^\top \nabla J_K(\boldsymbol{\omega}^{(n)}) = 2\Delta_n^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}) \leq -g(\boldsymbol{\omega}^{(n)})$ and

$$g[\boldsymbol{\omega}^{(n+)}(\alpha)] \leq (1-\alpha) g(\boldsymbol{\omega}^{(n)}) + 2\,\alpha^2 \lambda_{\max}(\mathbf{K}) \,. \tag{A.2}$$

The rest of the proof is by induction on n. The bound (A.1) is valid for $n = 1$ since $\|\boldsymbol{\omega}^{(1)} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}^2 \leq 2\,\lambda_{\max}(\mathbf{K})$. Suppose that it is satisfied by $\boldsymbol{\omega}^{(n)}$; (A.2) gives

$$\begin{aligned}
\|\boldsymbol{\omega}^{(n+1)} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}^2 &\leq 2\,\lambda_{\max}(\mathbf{K}) \left\{ \frac{1 + 2\,\log(n+2)}{n+1} - \frac{2(n+1)\log[1 + 1/(n+1)] - 1}{(n+1)^2} \right\} \\
&\leq 2\,\lambda_{\max}(\mathbf{K})\, \frac{1 + 2\,\log(n+2)}{n+1}
\end{aligned}$$

since $\log(1+t) \geq t/2$ for $t \in [0,1]$. ■

Using $\alpha = 2/(n+3)$ in (A.2), one can easily prove by induction that $g(\boldsymbol{\omega}^{(n)}) \leq 8\,\lambda_{\max}(\mathbf{K})/(n+3)$ for all n, see [20], which means that (4.18) with $\alpha_n = 2/(n+3)$ instead of $1/(n+1)$ satisfies $\|\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}^2 \leq 8\,\lambda_{\max}(\mathbf{K})/(n+3)$, $n \geq 1$, with thus a much faster decrease than (A.1). Using a different approach, it is shown in [28] that a rate of decrease of $\mathcal{O}(1/n)$ is also obtained when α_n corresponds to the sequence $\alpha_{n+1} = \alpha_n - \alpha_n^2/2$ with $\alpha_0 = 1$.

Remark A.1 (the infinite-dimensional situation)**.** A property similar to Theorem A.1 remains valid in the infinite-dimensional case, when working directly in the set $\mathscr{M}^+(1)$ of probability measures on $\mathscr{X}$. For the sake of simplicity, here we only consider the case when K is uniformly bounded, with moreover $K(\mathbf{x},\mathbf{x}) = 1$ and $K(\mathbf{x},\mathbf{x}') \geq 0$ for all $\mathbf{x}, \mathbf{x}'$. One may refer to [18] for a deeper analysis, including in particular results in the situation where approximate minimization over a finite set is conducted to select $\mathbf{x}_{n+1}$ in (4.17). The assumption above on K implies $\gamma_K^2(\xi,\nu) = \mathscr{E}_K(\xi - \nu) = \|P_\xi - P_\nu\|_{\mathcal{H}_K}^2 \leq 2$ for any ξ, ν in $\mathscr{M}^+(1)$ and we obtain that Algorithm (4.18) with $\alpha_n = 1/(n+1)$, initialized at any $\xi^{(1)}$ in $\mathbb{P}_\Omega$, satisfies

$$\gamma_K^2(\xi^{(n)}, \mu) \leq 2\,\frac{1 + 2\,\log(n+1)}{n} \,, \quad n \geq 1 \,.$$

The proof is similar to that of Theorem A.1. Denoting $\xi^{(n+)}(\alpha) = \xi^{(n)} + \alpha\Delta_n$, with $\Delta_n = \delta_{\mathbf{x}_{n+1}} - \xi^{(n)}$, we obtain

$$\begin{aligned}
\gamma_K^2[\xi^{(n+)}(\alpha), \mu] &= \gamma_K^2(\xi^{(n)}, \mu) + 2\alpha \mathscr{E}_K(\Delta_n, \xi^{(n)} - \mu) + \alpha^2 \mathscr{E}_K(\Delta_n) \\
&\leq \gamma_K^2(\xi^{(n)}, \mu) + 2\alpha \mathscr{E}_K(\Delta_n, \xi^{(n)} - \mu) + 2\alpha^2 \,.
\end{aligned}$$

The convexity of $\mathscr{E}_K(\cdot)$ implies

$$\begin{aligned}
\mathscr{E}_K(\xi^{(n)} - \mu) \geq 0 &\geq \mathscr{E}_K(\xi^{(n)} - \mu) + F_K(\xi^{(n)} - \mu, \mu) \\
&\geq \mathscr{E}_K(\xi^{(n)} - \mu) + \min_{\nu \in \mathscr{M}^+(1)} F_K(\xi^{(n)} - \mu, \nu) \\
&= \mathscr{E}_K(\xi^{(n)} - \mu) + 2 \min_{\mathbf{x} \in \mathscr{X}} [P_{\xi^{(n)}}(\mathbf{x}) - P_\mu(\mathbf{x}) - \mathscr{E}_K(\xi^{(n)} - \mu)] \\
&= \mathscr{E}_K(\xi^{(n)} - \mu) + 2\mathscr{E}_K(\Delta_n, \xi^{(n)} - \mu) \,.
\end{aligned}$$

Therefore, $\gamma_K^2[\xi^{(n+)}(\alpha), \mu] \leq (1-\alpha)\,\gamma_K^2(\xi^{(n)}, \mu) + 2\alpha^2$, and the rest of the proof is by induction on n, using $\xi^{(n+1)} = \xi^{(n+)}[1/(n+1)]$. Similarly, we get $\gamma_K^2(\xi^{(n)}, \mu) \leq 8/(n+3)$ when $\alpha_n = 2/(n+3)$. ◁

Next lemma, based on [19], shows that $\|\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}^2$ decreases as C/n^2 when $\hat{\boldsymbol{\omega}}$ lies in the interior of $\mathbb{P}_\Omega$. Here, contrary to [19], we do not assume that $\mathcal{H}$ is finite dimensional and use instead the finite dimensionality of $\boldsymbol{\omega}$.

Lemma A.2. *When $\hat{\boldsymbol{\omega}}$ is in the interior of $\mathbb{P}_\Omega$, (4.18) with $\alpha_n = 1/(n+1)$, initialized at any $\boldsymbol{\omega}^{(1)}$ in $\mathbb{P}_\Omega$, satisfies*

$$\|\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}^2 \leq 4R_*^2 \left(1 + \frac{R_*^2}{\alpha_*^2}\right) \frac{1}{n^2}\,, \quad n \geq 1\,,$$

where $R_* = [\lambda_{\max}(\mathbf{K})\,(1 - 1/\Omega)]^{1/2}$ and $\alpha_* = w_*/L$, with $w_* = \min_{i=1,\ldots,\Omega}\{\hat{\mathbf{w}}\}_i$ (so that $w_* \leq 1/\Omega$) and $L = (\max_{i=1,\ldots,\Omega}\{\mathbf{K}^{-1}\}_{ii})^{1/2}$.

Proof. Denote $\mathbf{v}(\alpha) = \hat{\boldsymbol{\omega}} - \alpha(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}})/\|\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}$, $\alpha > 0$. Then, for any $i = 1, \ldots, \Omega$,

$$\begin{aligned}\frac{|\mathbf{e}_i^\top(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}})|}{\|\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}} &\leq \max_{\mathbf{u}:\, \mathbf{u}^\top \mathbf{1}_\Omega = 0,\, \mathbf{u}^\top \mathbf{K}\mathbf{u} = 1} |\mathbf{e}_i^\top \mathbf{u}| \\ &\leq \max_{\mathbf{u}:\, \mathbf{u}^\top \mathbf{K}\mathbf{u} = 1} |\mathbf{e}_i^\top \mathbf{u}| = \sqrt{\mathbf{e}_i^\top \mathbf{K}^{-1} \mathbf{e}_i} \leq L = w_*/\alpha_*\,,\end{aligned}$$

so that $\{\mathbf{v}(\alpha)\}_i \geq \{\hat{\boldsymbol{\omega}}\}_i - w_* \geq 0$, and $\mathbf{v}(\alpha) \in \mathbb{P}_\Omega$, for any $\alpha \leq \alpha_*$. The definition of $\mathbf{e}_{i_n^+}$ then implies that

$$(\mathbf{e}_{i_n^+} - \hat{\boldsymbol{\omega}})^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}) \leq [\mathbf{v}(\alpha_*) - \hat{\boldsymbol{\omega}})]^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}) = -\alpha_* \|\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}\,. \quad \text{(A.3)}$$

The rest of the proof is based on [19]. Denote $\mathbf{e}_{i_0}^+ = \boldsymbol{\omega}^{(1)}$ and $\mathbf{z}_n = \sum_{i=1}^n (\mathbf{e}_{i_{n-1}^+} - \hat{\boldsymbol{\omega}})$. We can write $\boldsymbol{\omega}^{(n)} = (1/n)\sum_{i=1}^n \mathbf{e}_{i_{n-1}^+}$, so that $\mathbf{z}_n = n(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}})$, $n^2\|\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}^2 = \|\mathbf{z}_n\|_{\mathbf{K}}^2$, and we only need to bound $\|\mathbf{z}_n\|_{\mathbf{K}}^2$. We have

$$\|\mathbf{z}_n\|_{\mathbf{K}}^2 - \|\mathbf{z}_{n+1}\|_{\mathbf{K}}^2 = -2(\mathbf{e}_{i_n^+} - \hat{\boldsymbol{\omega}})^\top \mathbf{K}\mathbf{z}_n - \|\mathbf{e}_{i_n^+} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}^2\,,$$

where $\|\mathbf{e}_{i_n^+} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}} \leq 2\,R_*$ and $(\mathbf{e}_{i_n^+} - \hat{\boldsymbol{\omega}})^\top \mathbf{K}\mathbf{z}_n \leq -\alpha_* \|\mathbf{z}_n\|_{\mathbf{K}}$ from (A.3). Therefore,

$$\|\mathbf{z}_{n+1}\|_{\mathbf{K}}^2 \leq \|\mathbf{z}_n\|_{\mathbf{K}}^2 - 2\alpha_* \left(\|\mathbf{z}_n\|_{\mathbf{K}} - 2R_*^2/\alpha_*\right)\,.$$

Suppose that $\|\mathbf{z}_n\|_{\mathbf{K}} > 2R_*^2/\alpha_*$. Then, $\|\mathbf{z}_{n+1}\|_{\mathbf{K}}^2 \leq \|\mathbf{z}_n\|_{\mathbf{K}}^2$, and $\|\mathbf{z}_n\|_{\mathbf{K}}^2$ decreases until some n_0 when $\|\mathbf{z}_{n_0}\|_{\mathbf{K}} \leq 2R_*^2/\alpha_*$. But then,

$$\|\mathbf{z}_{n_0+1}\|_{\mathbf{K}}^2 \leq \|\mathbf{z}_{n_0}\|_{\mathbf{K}}^2 - 2\alpha_* \left(\|\mathbf{z}_{n_0}\|_{\mathbf{K}} - 2R_*^2/\alpha_*\right) \leq 4R_*^2(1 + R_*^2/\alpha_*^2)\,,$$

so that $\|\mathbf{z}_n\|_{\mathbf{K}}^2 \leq 4R_*^2(1 + R_*^2/\alpha_*^2)$ for all $n > n_0$. ∎

Lemma A.2 indicates that $\|\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}^2 \leq C/n^2$. However, for large Ω the constant C grows like $\mathcal{O}(\Omega^2)$ (since $\alpha_* \leq 1/(L\Omega)$) which makes this result of theoretical interest only. Note that applications typically concern situations where Ω is very large.

A.2. Vertex-direction, optimal step-size. The choice of a predefined step-size $\alpha_n = 1/(n+1)$ in (4.18) does not ensure a monotonic decrease of $\mathscr{E}_K(\xi^{(n)} - \mu)$. An alternative option is to choose α_n that minimizes $\mathscr{E}_K[\xi^{(n+)}(\alpha) - \mu]$ with respect to $\alpha \in [0, 1]$, with $\xi^{(n+)}(\alpha) = (1-\alpha)\,\xi^{(n)} + \alpha\,\delta_{\mathbf{x}_{n+1}}$ and $\mathbf{x}_{n+1}$ given by (4.17). Straightforward calculation gives $\alpha_n = \min\{1, \hat{\alpha}_n\}$, with

$$\begin{aligned}\hat{\alpha}_n &= \frac{\langle P_{\xi^{(n)}} - P_\mu, P_{\xi^{(n)}} - P_{\delta_{\mathbf{x}_{n+1}}}\rangle_K}{\|P_{\xi^{(n)}} - P_{\delta_{\mathbf{x}_{n+1}}}\|_{\mathcal{H}_K}^2} \\ &= \frac{\mathscr{E}_K(\xi^{(n)}) - P_{\xi^{(n)}}(\mathbf{x}_{n+1}) - \sum_{i=1}^n w_i^{(n)} P_\mu(\mathbf{x}_i) + P_\mu(\mathbf{x}_{n+1})}{\mathscr{E}_K(\xi^{(n)}) - 2P_{\xi^{(n)}}(\mathbf{x}_{n+1}) + K(\mathbf{x}_{n+1}, \mathbf{x}_{n+1})}\end{aligned}$$

(which requires that $\mathscr{E}_K(\xi^{(n)}) < \infty$), and $\hat{\alpha}_n$ satisfies

$$\hat{\alpha}_n = \frac{(\mathbf{e}_{i_n^+} - \boldsymbol{\omega}^{(n)})^\top \mathbf{K}(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}^{(n)})}{\|\mathbf{e}_{i_n^+} - \boldsymbol{\omega}^{(n)}\|_{\mathbf{K}}^2}\,. \tag{A.4}$$

Next Lemma indicates that $\hat{\alpha}_n \leq 1$ when $\hat{\mathbf{w}} \in \mathbb{P}_\Omega$, so that setting $\alpha_n = \hat{\alpha}_n$ in (4.18) ensures that $\boldsymbol{\omega}^{(n)}$ remains in $\mathbb{P}_\Omega$ for all n. It should be noticed that the global decrease of $\|\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}^2$ over *many* iterations with this optimal α_n is not necessarily better that with the predefined step-size $\alpha_n = 1/(n+1)$ of Section A.1; see in particular [28] for such considerations; see also [5]. One may refer to [27] for the infinite-dimensional situation.

Lemma A.3. *When $\hat{\boldsymbol{\omega}} \in \mathbb{P}_\Omega$, $\hat{\alpha}_n$ given by (A.4) is less than or equal to one.*

Proof. We can write $(\mathbf{e}_{i_n^+} - \boldsymbol{\omega}^{(n)})^\top \mathbf{K}(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}^{(n)}) = \|\mathbf{e}_{i_n^+} - \boldsymbol{\omega}^{(n)}\|_{\mathbf{K}}^2 + (\mathbf{e}_{i_n^+} - \hat{\boldsymbol{\omega}})^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}) - \|\mathbf{e}_{i_n^+} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}^2$. When $\hat{\boldsymbol{\omega}} \in \mathbb{P}_\Omega$, $\hat{\omega}_i \geq 0$ for all i, and $\sum_{i=1}^\Omega \hat{\omega}_i(\mathbf{e}_i - \hat{\boldsymbol{\omega}}) = \mathbf{0}$ implies that $\sum_{i=1}^\Omega \hat{\omega}_i(\mathbf{e}_i - \hat{\boldsymbol{\omega}})^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}) = 0$. Therefore $\min_{i=1,\ldots,\Omega}(\mathbf{e}_i - \hat{\boldsymbol{\omega}})^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}) = (\mathbf{e}_{i_n^+} - \hat{\boldsymbol{\omega}})^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}) \leq 0$, which gives $\hat{\alpha}_n \leq 1$. ∎

Theorem A.2. *Algorithm (4.18) with $\alpha_n = \hat{\alpha}_n$ given by (A.4), initialized at any $\boldsymbol{\omega}^{(1)} \in \mathbb{P}_\Omega$, satisfies*

$$\|\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}^2 \leq 8\,\lambda_{\max}(\mathbf{K})\,\frac{1}{n+3}\,, \quad n \geq 1\,. \tag{A.5}$$

Proof. The proof follows [20, Sect. 2] and uses the same notation as in the proof of Theorem A.1. The right-hand side of (A.2) is minimum for $\hat{\alpha} = g(\boldsymbol{\omega}^{(n)})/[4\,\lambda_{\max}(\mathbf{K})] \leq 1/2$. Therefore,

$$\begin{aligned} g(\boldsymbol{\omega}^{(n+1)}) = \min_{\alpha\in[0,1]} g[\boldsymbol{\omega}^{(n+)}(\alpha)] &\leq (1-\hat{\alpha})g(\boldsymbol{\omega}^{(n)}) + 2\,\hat{\alpha}^2\lambda_{\max}(\mathbf{K}) \\ &= g(\boldsymbol{\omega}^{(n)})\left[1 - \frac{g(\boldsymbol{\omega}^{(n)})}{8\,\lambda_{\max}(\mathbf{K})}\right]. \end{aligned}$$

Since $1 - t \leq 1/(1+t)$ for all $t > -1$, we obtain

$$g(\boldsymbol{\omega}^{(n+1)})/[8\,\lambda_{\max}(\mathbf{K})] \leq \frac{1}{1 + \{g(\boldsymbol{\omega}^{(n)})/[8\,\lambda_{\max}(\mathbf{K})]\}^{-1}}$$

which, by induction, implies that $g(\boldsymbol{\omega}^{(n)}) \leq 8\,\lambda_{\max}(\mathbf{K})/(n+3)$; that is, (A.5). ∎

Lemma A.4. *When $\hat{\boldsymbol{\omega}}$ is in the interior of $\mathbb{P}_\Omega$, (4.18) with $\alpha_n = \hat{\alpha}_n$ given by (A.4), initialized at any $\boldsymbol{\omega}^{(1)} \in \mathbb{P}_\Omega$, satisfies*

$$\|\boldsymbol{\omega}^{(n+1)} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}^2 \leq \|\boldsymbol{\omega}^{(1)} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}^2 \,\exp\left(-\frac{\alpha_*^2\, n}{4R_*^2}\right), \quad n \geq 1\,, \tag{A.6}$$

where $R_* = [\lambda_{\max}(\mathbf{K})\,(1 - 1/\Omega)]^{1/2}$ and $\alpha_* = w_*/L$, with $w_* = \min_{i=1,\ldots,\Omega}\{\hat{\mathbf{w}}\}_i$ (so that $w_* \leq 1/\Omega$) and $L = (\max_{i=1,\ldots,\Omega}\{\mathbf{K}^{-1}\}_{ii})^{1/2}$.

Proof. We use the same approach as in [6] and use the same notation as in the proof of Theorem A.1. We can write $g(\boldsymbol{\omega}^{(n+1)}) = g[\boldsymbol{\omega}^{(n+)}(\hat{\alpha}_n)]$, with $\hat{\alpha}_n$ given by (A.4). Therefore,

$$\begin{aligned} g(\boldsymbol{\omega}^{(n+1)}) &= g(\boldsymbol{\omega}^{(n)}) + 2\hat{\alpha}_n\Delta_n^\top\mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}) + \alpha_n^2\|\Delta_n\|_{\mathbf{K}}^2 \\ &= g(\boldsymbol{\omega}^{(n)}) - \frac{[\Delta_n^\top\mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}})]^2}{\|\Delta_n\|_{\mathbf{K}}^2}\,. \end{aligned}$$

Equation (A.3) implies that $[\Delta_n^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}})]^2 \geq \alpha_*^2 g(\boldsymbol{\omega}^{(n)})$, and thus

$$g(\boldsymbol{\omega}^{(n+1)}) \leq g(\boldsymbol{\omega}^{(n)}) \left[1 - \frac{\alpha_*^2}{\|\Delta_n\|_{\mathbf{K}}^2}\right] \leq g(\boldsymbol{\omega}^{(n)}) \left[1 - \frac{\alpha_*^2}{4R_*^2}\right] .$$

This implies $g(\boldsymbol{\omega}^{(n+1)}) \leq g(\boldsymbol{\omega}^{(1)}) \exp[-\alpha_*^2\, n/(4R_*^2)]$, that is, (A.6). ∎

Similarly to Lemma A.2, the small value of the constant α_* makes the linear convergence rate in (A.6) of theoretical interest only.

A.3. Vertex-exchange. Following [66, 67], one may also use a vertex-exchange method based on the true steepest-descent direction, see also [12, 13]. The iterations are then

$$\xi^{(n+1)} = \xi^{(n)} + \alpha_n \left(\delta_{\mathbf{x}_{n+1}} - \delta_{\mathbf{x}_n^-}\right), \tag{A.7}$$

where $\mathbf{x}_{n+1}$ is given by (4.17) and

$$\mathbf{x}_n^- \in \operatorname{Arg} \max_{\mathbf{x} \in \operatorname{Supp}(\xi^{(n)})} \left[P_{\xi^{(n)}}(\mathbf{x}) - P_\mu(\mathbf{x})\right] , \tag{A.8}$$

with $\operatorname{Supp}(\xi^{(n)}) = \mathbf{X}_n$ the support of $\xi^{(n)}$. The step-size α_n is then given by $\min\{\hat{\alpha}_n, \xi^{(n)}(\mathbf{x}_n^-)\}$, where $\hat{\alpha}_n$ minimizes $\mathscr{E}_K[\{\xi^{(n)} + \alpha\,(\delta_{\mathbf{x}_{n+1}} - \delta_{\mathbf{x}_n^-})\} - \mu]$ with respect to α (the constraint $\alpha_n \leq \xi^{(n)}(\mathbf{x}_n^-)$ ensures that $\xi^{(n+1)} \in \mathscr{M}^+(1)$ when $\xi^{(n)} \in \mathscr{M}^+(1)$). Direct calculation gives

$$\begin{aligned} \hat{\alpha}_n &= \frac{\langle P_{\xi^{(n)}} - P_\mu, P_{\delta_{\mathbf{x}_n^-}} - P_{\delta_{\mathbf{x}_{n+1}}} \rangle_K}{\|P_{\delta_{\mathbf{x}_n^-}} - P_{\delta_{\mathbf{x}_{n+1}}}\|_{\mathcal{H}_K}^2} \\ &= \frac{[P_{\xi^{(n)}}(\mathbf{x}_n^-) - P_\mu(\mathbf{x}_n^-)] - [P_{\xi^{(n)}}(\mathbf{x}_{n+1}) - P_\mu(\mathbf{x}_{n+1})]}{K(\mathbf{x}_n^-, \mathbf{x}_n^-) + K(\mathbf{x}_{n+1}, \mathbf{x}_{n+1}) - 2K(\mathbf{x}_n^-, \mathbf{x}_{n+1})} . \end{aligned} \tag{A.9}$$

For the algorithm defined by (A.7, A.8), we have $\boldsymbol{\omega}^{(n+1)} = \boldsymbol{\omega}^{(n)} + \alpha_n \Delta_n$ with now $\Delta_n = \mathbf{e}_{i_n^+} - \mathbf{e}_{i_n^-}$, where we take $i_n^+ \in \operatorname{Arg}\min_{i=1,\dots,\Omega} \mathbf{e}_i^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}})$ and $i_n^- \in \operatorname{Arg}\max_{i:\, \mathbf{e}_i^\top \boldsymbol{\omega}^{(n)} > 0} \mathbf{e}_i^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}})$. The step size (A.9) equals

$$\hat{\alpha}_n = \frac{\Delta_n^\top \mathbf{K}(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}^{(n)})}{\|\Delta_n\|_{\mathbf{K}}^2} . \tag{A.10}$$

Take $\alpha_n = \min\{\omega_{i_n^-}^{(n)}, \hat{\alpha}_n\}$ in (A.7), so that $\boldsymbol{\omega}^{(n)}$ remains in $\mathbb{P}_\Omega$ for all n. Using the same notation as in the proof of Theorem A.1, we have

$$g[\boldsymbol{\omega}^{(n+)}(\alpha)] \leq g(\boldsymbol{\omega}^{(n)}) + 2\alpha \Delta_n^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}) + 2\,\alpha^2 \lambda_{\max}(\mathbf{K}) ,$$

and, since $\boldsymbol{\omega}^{(n)} \in \mathbb{P}_\Omega$, the convexity of $g(\cdot)$ and the definition of Δ_n imply that

$$g(\hat{\boldsymbol{\omega}}) = 0 \geq g(\boldsymbol{\omega}^{(n)}) + 2(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}^{(n)})^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}) \geq g(\boldsymbol{\omega}^{(n)}) + 2\Delta_n^\top \mathbf{K}(\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}) .$$

We obtain the following property; the proof is identical to that of Theorem A.2.

Theorem A.3. *Suppose that $\hat{\boldsymbol{\omega}}$ and $\mathbf{K}$ are such that $\hat{\alpha}_n \leq \omega_{i_n^-}^{(n)}$ for any $\boldsymbol{\omega}^{(n)} \in \mathbb{P}_\Omega$. Then, algorithm (A.7, A.8) with $\alpha_n = \hat{\alpha}_n$ given by (A.10), initialized at any $\boldsymbol{\omega}^{(1)} \in \mathbb{P}_\Omega$, satisfies*

$$\|\boldsymbol{\omega}^{(n)} - \hat{\boldsymbol{\omega}}\|_{\mathbf{K}}^2 \leq \frac{8\,\lambda_{\max}(\mathbf{K})}{n+3} , \quad n \geq 1 .$$

There exist situations where the condition $\hat{\alpha}_n \le \omega_{i_n^-}^{(n)}$ is not satisfied. Take for instance $\Omega = 3$, $\mathbf{K}$ the identity matrix and $\hat{\boldsymbol{\omega}} = (0,0,1)^\top$, $\boldsymbol{\omega}^{(n)} = (1/3,1/3,1/3)^\top$; then $\hat{\alpha}_n = 1/2 > \omega_{i_n^-}^{(n)} = 1/3$). On the other hand, the condition is satisfied for instance for $\hat{\boldsymbol{\omega}} = \mathbf{1}_\Omega/\Omega$ and $\mathbf{K}$ the identity matrix (we have $i_{n^-} = \operatorname{Arg\,max}_{i:\,\omega_i^{(n)}>0}(\omega_i^{(n)} - \hat{\omega}_i)$, and $\sum_{i=1}^{\Omega}(\omega_i^{(n)} - \hat{\omega}_i) = 0$ implies that $\omega_{i_n^-}^{(n)} > \hat{\omega}_{i_n^-}$ and similarly $\omega_{i_n^+}^{(n)} < \hat{\omega}_{i_n^+}$; we get $\hat{\alpha}_n = (\omega_{i_n^-}^{(n)} - \omega_{i_n^+}^{(n)})/2 \le \omega_{i_n^-}^{(n)}/2 < \omega_{i_n^-}^{(n)}$), and numerical experiments indicate that it holds true in most situations.

Appendix B. Bayesian quadrature: several integrals.

Following [72], consider a generalization of the situation considered in Section 2.2 where one wishes to estimate

$$\mathbf{I}_\mu(f) = \mathsf{E}_\mu\{f(\mathbf{X})\mathbf{r}(\mathbf{X})\} = \int_{\mathscr{X}} f(\mathbf{x})\mathbf{r}(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x})\,,$$

with $\mathbf{r}(\mathbf{x}) = (r_0(\mathbf{x}),\dots,r_p(\mathbf{x}))^\top$ a vector of $p+1$ known functions of $\mathbf{x}$, such that the $(p+1)\times(p+1)$ matrix

$$\mathbf{M}_r = \mathsf{E}_\mu\{\mathbf{r}(\mathbf{X})\mathbf{r}^\top(\mathbf{X})\}$$

exists and is nonsingular. See also [58]. Without any loss of generality, we may assume that $r_0(\mathbf{x}) \equiv 1$.

We also slightly generalize the model (2.1) by introducing a linear trend $\mathbf{h}^\top(\mathbf{x})\boldsymbol{\beta}$; that is, we consider

$$f(\mathbf{x}) = \mathbf{h}^\top(\mathbf{x})\boldsymbol{\beta} + Z_x\,, \tag{B.1}$$

where $\mathbf{h}(\mathbf{x}) = (h_0(\mathbf{x}),\dots,h_{p'}(\mathbf{x}))^\top$ is a vector of $p'+1$ known functions of $\mathbf{x}$ and $\boldsymbol{\beta} \in \mathbb{R}^{p'+1}$ has the normal prior $\mathscr{N}(\hat{\boldsymbol{\beta}}^0, \sigma^2\mathbf{A})$, non-informative so that we can replace $\mathbf{A}^{-1}$ by the null matrix $\mathbf{0}$ in all calculations (the choice of $\hat{\boldsymbol{\beta}}^0$ being then irrelevant). We assume that the matrix $\mathsf{E}_\mu\{\mathbf{h}(\mathbf{X})\mathbf{h}^\top(\mathbf{X})\}$ is well-defined. For reasons that will become clear below, we shall consider in particular the case where $\mathbf{h} = \mathbf{r}$.

The posterior mean and variance of $f(\mathbf{x})$, conditional on σ^2 and K, are now, respectively,

$$\begin{aligned}
\hat{\eta}_n(\mathbf{x}) &= \mathbf{h}^\top(\mathbf{x})\hat{\boldsymbol{\beta}}^n + \mathbf{k}_n^\top(\mathbf{x})\mathbf{K}_n^{-1}(\mathbf{y}_n - \mathbf{H}_n\hat{\boldsymbol{\beta}}^n)\,,\\
\sigma^2\rho_n^2(\mathbf{x}) &= \sigma^2\big\{K(\mathbf{x},\mathbf{x}) - \mathbf{k}_n^\top(\mathbf{x})\mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x})\\
&\qquad + [\mathbf{h}(\mathbf{x}) - \mathbf{H}_n^\top\mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x})]^\top(\mathbf{H}_n^\top\mathbf{K}_n^{-1}\mathbf{H}_n)^{-1}[\mathbf{h}(\mathbf{x}) - \mathbf{H}_n^\top\mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x})]\big\}\,,
\end{aligned}$$

where $\{\mathbf{H}_n(\mathbf{x})\}_{i,j} = h_j(\mathbf{x}_i)$, $i = 1,\dots,n$, $j = 0,\dots,p'$, and

$$\hat{\boldsymbol{\beta}}^n = (\mathbf{H}_n^\top\mathbf{K}_n^{-1}\mathbf{H}_n)^{-1}\mathbf{H}_n^\top\mathbf{K}_n^{-1}\mathbf{y}_n\,.$$

The posterior mean and covariance matrix of $\mathbf{I}_\mu(f)$ are

$$\widehat{\mathbf{I}}_n = \mathbf{B}(\mu)\hat{\boldsymbol{\beta}}^n + \mathbf{P}_n(\mu)\mathbf{K}_n^{-1}(\mathbf{y}_n - \mathbf{H}_n\hat{\boldsymbol{\beta}}^n)\,, \tag{B.2}$$

$$\begin{aligned}
\sigma^2\,\mathbf{V}_n &= \sigma^2\big\{\mathbf{U}_r(\mu) - \mathbf{P}_n(\mu)\mathbf{K}_n^{-1}\mathbf{P}_n^\top(\mu)\\
&+ \big[\mathbf{B}(\mu) - \mathbf{P}_n(\mu)\mathbf{K}_n^{-1}\mathbf{H}_n\big]\,(\mathbf{H}_n^\top\mathbf{K}_n^{-1}\mathbf{H}_n)^{-1}\,\big[\mathbf{B}(\mu) - \mathbf{P}_n(\mu)\mathbf{K}_n^{-1}\mathbf{H}_n\big]^\top\big\}\,,
\end{aligned} \tag{B.3}$$

where $\mathbf{B}(\mu) = \mathsf{E}_\mu\{\mathbf{r}(\mathbf{X})\mathbf{h}^\top(\mathbf{X})\}$, $\mathbf{P}_n(\mu) = \mathsf{E}_\mu\{\mathbf{r}(\mathbf{X})\mathbf{k}_n^\top(\mathbf{X})\}$ and

$$\mathbf{U}_r(\mu) = \mathsf{E}_\mu\{\mathbf{r}(\mathbf{X})\mathbf{r}^\top(\mathbf{X}')K(\mathbf{X},\mathbf{X}')\}\,,$$

with $\mathbf{X}$ and $\mathbf{X}'$ i.i.d. $\sim \mu$.

Remark B.1 (Reproduction of cubature rules)**.** Consider the case where $p = 1$ ($\mathbf{r}(\mathbf{x}) \equiv 1$), $n = p' + 1$, so that $\mathbf{H}_n$ is $n \times n$, and suppose that $\mathbf{X}_n$ is such that $\mathbf{H}_n$ is nonsingular. Then, direct calculation shows that $\widehat{\mathbf{I}}_n = \hat{\mathbf{w}}_n^\top \mathbf{y}_n$, with $\mathbf{w}_n^\top = \mathbf{H}_n^\top \mathsf{E}_\mu\{\mathbf{h}(\mathbf{X})\}$ and $\mathbf{V}_n = \mathscr{E}_K(\xi_n - \mu)$, where ξ_n has weights $\mathbf{w}_n$. The weights $\mathbf{w}_n$ are independent of the choice of the kernel K and every function f in the linear space spanned by $\mathbf{h}(\cdot)$ is integrated exactly ($\widehat{\mathbf{I}}_n = I_\mu(f)$ when $f(\mathbf{x}) = \boldsymbol{\gamma}^\top \mathbf{h}(\mathbf{x})$ for some vector $\boldsymbol{\gamma}$), see [52, Th. 2.10]. In the same paper, these results are used to show that for any n-point cubature rule there exists n functions $h_i(\cdot)$ such that the rule corresponds to Bayesian integration for model (B.1). One may also refer to [53] for the relation between polynomial-based quadrature rules and Bayesian quadrature (for a suitably chosen polynomial kernel) when $\boldsymbol{\beta}$ in (B.1) is considered as a vector of known constants (for instance, zero), so that the posterior variance is given by (2.11). ◁

Suppose that $\mathbf{M}_h = \mathsf{E}_\mu\{\mathbf{h}(\mathbf{X})\mathbf{h}^\top(\mathbf{X})\}$ is nonsingular. Following Section 3.5.2, we can write $f(\mathbf{x}) = \mathbf{h}^\top(\mathbf{x})\boldsymbol{\beta} + \mathcal{P}_h Z_x + (\mathrm{Id}_{L^2} - \mathcal{P}_h)Z_x$, where $\mathcal{P}_h$ denotes the orthogonal projection of $L^2(\mathscr{X}, \mu)$ onto the linear space spanned by $\mathbf{h}(\cdot)$; that is, $\mathcal{P}_h g(\mathbf{x}) = \mathbf{h}^\top(\mathbf{x})\mathbf{M}_h^{-1} \int_{\mathscr{X}} \mathbf{h}(\mathbf{x}')g(\mathbf{x}')\,\mathrm{d}\mu(\mathbf{x}')$ for all $g \in L^2(\mathscr{X}, \mu)$. This gives

$$\mathcal{P}_h Z_x = \mathbf{h}^\top(\mathbf{x})\mathbf{M}_h^{-1} \int_{\mathscr{X}} \mathbf{h}(\mathbf{x}')Z_{x'}\,\mathrm{d}\mu(\mathbf{x}')\,.$$

In absence of prior information on $\boldsymbol{\beta}$ ($\mathbf{A}^{-1} = \mathbf{0}$), the prior on the parameters $\boldsymbol{\beta}' = \boldsymbol{\beta} + \mathbf{M}_h^{-1} \int_{\mathscr{X}} \mathbf{h}(\mathbf{x}')Z_{x'}\,\mathrm{d}\mu(\mathbf{x}')$ remains non-informative, and the covariance kernel of $\widetilde{Z}_x = (\mathrm{Id}_{L^2} - \mathcal{P}_h)Z_x$ is

$$\begin{aligned} K_\mu(\mathbf{x},\mathbf{x}') = K(\mathbf{x},\mathbf{x}') &- \mathbf{u}_\mu^\top(\mathbf{x})\mathbf{M}_h^{-1}\mathbf{h}(\mathbf{x}') - \mathbf{h}^\top(\mathbf{x})\mathbf{M}_h^{-1}\mathbf{u}_\mu(\mathbf{x}') \\ &+ \mathbf{h}^\top(\mathbf{x})\mathbf{M}_h^{-1}\mathbf{U}_h(\mu)\mathbf{M}_h^{-1}\mathbf{h}(\mathbf{x}')\,, \end{aligned}$$

where $\mathbf{U}_h(\mu) = \mathsf{E}_\mu\{\mathbf{h}(\mathbf{X})\mathbf{r}^\top(\mathbf{X}')K(\mathbf{X},\mathbf{X}')\}$ and $\mathbf{u}_\mu(\mathbf{x}) = \mathsf{E}_\mu\{\mathbf{h}(\mathbf{X})K(\mathbf{X},\mathbf{x})\}$, $\mathbf{x} \in \mathscr{X}$.

Similarly to Section 4.2 (see [37, Sect. 5.4]), this kernel reduction does not modify predictions, and direct calculation shows that $\mathsf{E}_\mu\{\mathbf{h}(\mathbf{X})\mathbf{h}^\top(\mathbf{X}')K_\mu(\mathbf{X},\mathbf{X}')\} = \mathbf{0}$ and $\mathsf{E}_\mu\{\mathbf{h}(\mathbf{X})\widetilde{\mathbf{k}}_n^\top(\mathbf{X})\} = \mathbf{0}$, with $\widetilde{\mathbf{k}}_n(\mathbf{x}) = (K_\mu(\mathbf{x},\mathbf{x}_1),\ldots,K_\mu(\mathbf{x},\mathbf{x}_n))^\top$. Taking $\mathbf{h} = \mathbf{r}$, we thus obtain the following property, where $\mathbf{R}_n$ is the $n \times (p+1)$ matrix $\{\mathbf{R}_n(\mathbf{x})\}_{i,j} = r_j(\mathbf{x}_i)$, $i = 1,\ldots,n$, $j = 0,\ldots,p$, and $\{\widetilde{\mathbf{K}}_n\}_{i,j} = K_\mu(\mathbf{x}_i,\mathbf{x}_j)$, $i,j = 1,\ldots,n$.

Lemma B.1. *When K is SPD and $\mathbf{h} = \mathbf{r}$ in (B.1), $\widehat{\mathbf{I}}_n$ given by (B.2) satisfies*

$$\widehat{\mathbf{I}}_n = \mathbf{M}_r(\mathbf{R}_n^\top \widetilde{\mathbf{K}}_n^{-1}\mathbf{R}_n)^{-1}(\mathbf{R}_n^\top \widetilde{\mathbf{K}}_n^{-1}\mathbf{y}_n)\,,$$

and the posterior covariance matrix (B.3) satisfies

$$\mathbf{V}_n = \mathbf{M}_r(\mathbf{R}_n^\top \widetilde{\mathbf{K}}_n^{-1}\mathbf{R}_n)^{-1}\mathbf{M}_r\,. \tag{B.4}$$

To ensure a precise estimation of $\mathbf{I}_\mu(f)$, we may select a design $\mathbf{X}_n$ that minimizes $\mathcal{J}(\mathbf{V}_n)$, with $\mathcal{J}(\cdot)$ a Loewner increasing function defined on the set of symmetric non-negative define matrices. Typical choices are $\mathcal{J}(\mathbf{V}_n) = \det(\mathbf{V}_n)$ (D-optimality) and $\mathcal{J}(\mathbf{V}_n) = \mathrm{trace}(\mathbf{V}_n)$ (A-optimality). Greedy minimization of $\mathcal{J}(\mathbf{V}_n)$ corresponds to

Sequential Bayesian Quadrature, see Section 4.4. Using (B.4) and formulae for the inversion of a block matrix, we obtain the following expressions for $\det(\mathbf{V}_{n+1})$ and $\mathrm{trace}(\mathbf{V}_{n+1})$:

$$\det(\mathbf{V}_{n+1}) = \det(\mathbf{V}_n)\,\frac{K_\mu(\mathbf{x},\mathbf{x}) - \widetilde{\mathbf{k}}_n^\top(\mathbf{x})\widetilde{\mathbf{K}}_n^{-1}\widetilde{\mathbf{k}}_n(\mathbf{x})}{\widetilde{\rho}_n^2(\mathbf{x})}\,,$$

$$\mathrm{trace}(\mathbf{V}_{n+1}) = \mathrm{trace}(\mathbf{V}_n) - \frac{[\mathbf{r}(\mathbf{x}) - \mathbf{R}_n^\top\widetilde{\mathbf{K}}_n^{-1}\widetilde{\mathbf{k}}_n(\mathbf{x})]^\top(\mathbf{R}_n^\top\widetilde{\mathbf{K}}_n^{-1}\mathbf{R}_n)^{-1}\mathbf{M}_r^2(\mathbf{R}_n^\top\widetilde{\mathbf{K}}_n^{-1}\mathbf{R}_n)^{-1}[\mathbf{r}(\mathbf{x}) - \mathbf{R}_n^\top\widetilde{\mathbf{K}}_n^{-1}\widetilde{\mathbf{k}}_n(\mathbf{x})]}{\widetilde{\rho}_n^2(\mathbf{x})}\,,$$

with

$$\widetilde{\rho}_n^2(\mathbf{x}) = \left[K_\mu(\mathbf{x},\mathbf{x}) - \widetilde{\mathbf{k}}_n^\top(\mathbf{x})\widetilde{\mathbf{K}}_n^{-1}\widetilde{\mathbf{k}}_n(\mathbf{x}) + \frac{(1 - \widetilde{\mathbf{k}}_n^\top(\mathbf{x})\widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n)^2}{\mathbf{1}_n^\top\widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n}\right].$$

When $p = 0$ ($\mathbf{r}(\mathbf{x}) \equiv 1$), $\mathbf{V}_n = s_n^2$ in (4.13) and $\det(\mathbf{V}_{n+1}) = \mathrm{trace}(\mathbf{V}_{n+1}) = s_{n+1}^2$ given by (4.15).

Appendix C. Energy and potential for the triangular kernel.

Consider the triangular kernel $K_\theta(x,x') = \max\{1-\theta|x-x'|, 0\}$, $\theta > 0$, with μ uniform on $[0,1]$. The expressions of $\mathscr{E}_K(\mu)$ and $P_\mu(x)$ vary depending on the range considered for θ, with in all cases $P_\mu(x) = 0$ when $x \le -1/\theta$ or $1 + 1/\theta \le x$.

$2 \le \theta$. $\mathscr{E}_K(\mu) = (3\theta - 1)/(3\theta^2)$ and

$$P_\mu(x) = \begin{cases} \theta(x+1/\theta)^2/2 & \text{if } -1/\theta \le x \le 0 \\ 1/\theta - \theta(1/\theta - x)^2/2 & \text{if } 0 \le x \le 1/\theta \\ 1/\theta & \text{if } 1/\theta \le x \le 1 - 1/\theta \\ 1/\theta - \theta(1/\theta + x - 1)^2/2 & \text{if } 1 - 1/\theta \le x \le 1 \\ \theta(1 - x + 1/\theta)^2/2 & \text{if } 1 \le x \le 1 + 1/\theta \end{cases}$$

$1 \le \theta \le 2$. $\mathscr{E}_K(\mu) = (3\theta - 1)/(3\theta^2)$ and

$$P_\mu(x) = \begin{cases} \theta(x+1/\theta)^2/2 & \text{if } -1/\theta \le x \le 0 \\ 1/\theta - \theta(1/\theta - x)^2/2 & \text{if } 0 \le x \le 1 - 1/\theta \\ 1/\theta - \theta(1/\theta - x)^2/2 - \theta(1/\theta + x - 1)^2/2 & \text{if } 1 - 1/\theta \le x \le 1/\theta \\ 1/\theta - \theta(1/\theta + x - 1)^2/2 & \text{if } 1/\theta \le x \le 1 \\ \theta(1 - x + 1/\theta)^2/2 & \text{if } 1 \le x \le 1 + 1/\theta \end{cases}$$

$0 < \theta \le 1$. $\mathscr{E}_K(\mu) = 1 - \theta/3$ and

$$P_\mu(x) = \begin{cases} \theta(x+1/\theta)^2/2 & \text{if } -1/\theta \le x \le 1 - 1/\theta \\ 1 - \theta/2 + \theta x & \text{if } 1 - 1/\theta \le x \le 0 \\ 1 - \theta/2 + \theta x - \theta x^2 & \text{if } 0 \le x \le 1 \\ 1 + \theta/2 - \theta x & \text{if } 1 \le x \le 1/\theta \\ \theta(1 - x + 1/\theta)^2/2 & \text{if } 1/\theta \le x \le 1 + 1/\theta \end{cases}$$

REFERENCES

[1] H. Aikawa and M. Essén. *Potential Theory – Selected Topics*. Springer, Berlin, 1996.

[2] A. Antoniadis. Analysis of variance on function spaces. *Math. Operationsforsch. u. Statist.*, 15(1):59–71, 1984.

[3] P. Audze and V. Eglais. New approach for planning out of experiments. *Problems of Dynamics and Strengths*, 35:104–107, 1977.

[4] Y. Auffray, P. Barbillon, and J.-M. Marin. Maximin design on non hypercube domains and kernel interpolation. *Statistics and Computing*, 22(3):703–712, 2012.

[5] F. Bach, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. *Proc. ICML 2012, arXiv preprint arXiv:1203.4523*, 2012.

[6] A. Beck and M. Teboulle. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Mathematical Methods of Operations Research*, 59(2):235–247, 2004.

[7] J. Bect, F. Bachoc, and D. Ginsbourger. A supermartingale approach to Gaussian process based sequential design of experiments. *Bernoulli*, 25(4A):2883–2919, 2019. arXiv preprint arXiv:1608.01118.

[8] J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22(3):773–793, 2012.

[9] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, Boston, 2004.

[10] S. Biedermann and H. Dette. Minimax optimal designs for nonparametric regression — a further optimality property of the uniform distribution. In P. Hackl A.C. Atkinson and W.G. Müller, editors, *mODa'6 – Advances in Model–Oriented Design and Analysis, Proceedings of the 76th Int. Workshop, Puchberg/Schneeberg (Austria)*, pages 13–20, Heidelberg, June 2001. Physica Verlag.

[11] G. Björck. Distributions of positive mass, which maximize a certain generalized energy integral. *Arkiv för Matematik*, 3(21):255–269, 1956.

[12] D. Böhning. Numerical estimation of a probability measure. *Journal of Statistical Planning and Inference*, 11:57–69, 1985.

[13] D. Böhning. A vertex-exchange-method in D-optimal design theory. *Metrika*, 33:337–347, 1986.

[14] A. Breger, M. Ether, and M. Gräf. Points on manifolds with asymptotically optimal covering radius. *Journal of Complexity*, 48:1–14, 2018.

[15] F.-X. Briol, C. Oates, M. Girolami, and M.A. Osborne. Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees. In *Advances in Neural Information Processing Systems*, pages 1162–1170, 2015.

[16] F.-X. Briol, C.J. Oates, M. Girolami, M.A. Osborne, and D. Sejdinovic. Probabilistic integration: A role in statistical computation? *Statistical Science*, 34(1):1–22, 2019.

[17] H. Cardot, P. Cénac, and J.-M. Monnez. A fast and recursive algorithm for clustering large datasets. *Comput. Statist. Data Anal.*, 56(6):1434–1449, 2012.

[18] W.Y. Chen, L. Mackey, J. Gorham, F.-X. Briol, and C.J. Oates. Stein points. *arXiv preprint arXiv:1803.10161v4, Proc. ICML*, 2018.

[19] Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. In *Proceedings 26th Conference on Uncertainty in Artificial Intelligence (UAI'10)*, pages 109–116, Catalina Island, CA, July 2010. AUAI Press Arlington, Virginia. arXiv preprint arXiv:1203.3472.

[20] K.M. Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):63, 2010.

[21] S.B. Damelin, F.J. Hickernell, D.L. Ragozin, and X. Zeng. On energy, discrepancy and group invariant measures on measurable subsets of Euclidean space. *J. Fourier Anal. Appl.*, 16:813–839, 2010.

[22] G. Detommaso, T. Cui, Y. Marzouk, A. Spantini, and R. Scheichl. A Stein variational Newton method. In *Advances in Neural Information Processing Systems*, pages 9187–9197, 2018.

[23] H. Dette, A. Pepelyshev, and A. Zhigljavsky. Best linear unbiased estimators in continuous time regression models. *The Annals of Statistics*, 47(4):1928–1959, 2019.

[24] P. Diaconis. Bayesian numerical analysis. *Statistical Decision Theory and Related Topics IV*, 1:163–175, 1988.

[25] J. Dick and F. Pillichshammer. *Digital Nets and Sequences. Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, Cambridge, 2010.

[26] Q. Du, V. Faber, and M. Gunzburger. Centroidal Voronoi tessellations: applications and algorithms. *SIAM Review*, 41(4):637–676, 1999.

[27] J.C. Dunn. Convergence rates for conditional gradient sequences generated by implicit step length rules. *SIAM J. Control and Optimization*, 18(5):473–487, 1980.

[28] J.C. Dunn and S. Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62:432–444, 1978.

[29] N. Durrande, D. Ginsbourger, O. Roustant, and L. Carraro. Anova kernels and RKHS of zero mean functions for model-based sensitivity analysis. *Journal of Multivariate Analysis*,

115:57–67, 2013.

[30] N. Durrande, J. Hensman, M. Rattray, and N.D. Lawrence. Detecting periodicities with Gaussian processes. *PeerJ Comput. Sci.*, 2:e50, 2016.

[31] K.-T. Fang, R. Li, and A. Sudjianto. *Design and Modeling for Computer Experiments*. Chapman & Hall/CRC, Boca Raton, 2006.

[32] V.V. Fedorov. *Theory of Optimal Experiments*. Academic Press, New York, 1972.

[33] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Res. Logist. Quart.*, 3:95–110, 1956.

[34] B. Fuglede. On the theory of potentials in locally compact spaces. *Acta mathematica*, 103:139–215, 1960.

[35] B. Fuglede and N. Zorii. Green kernels associated with Riesz kernels. *Annales Academiae Scientiarum Fennicae, Mathematica*, 43(to appear), 2018. arXiv preprint arXiv:1610.00268.

[36] B. Gauthier and L. Pronzato. Spectral approximation of the IMSE criterion for optimal designs in kernel-based interpolation models. *SIAM/ASA J. Uncertainty Quantification*, 2:805–825, 2014. DOI 10.1137/130928534.

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

[38] D. Ginsbourger. Sequential design of computer experiments. *Wiley StatsRef*, 99:1–11, 2017.

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

[40] T. Gneiting. Strictly and non-strictly positive define functions on spheres. *Bernoulli*, 19(4):1327–1349, 2013.

[41] A. Gorodetsky and Y. Marzouk. Mercer kernels and integrated variance experimental design: connections between Gaussian process regression and polynomial approximation. *SIAM/ASA J. Uncertainty Quantification*, 4(1):796–828, 2016.

[42] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*. Springer, Berlin, 2000.

[43] U. Grenander. Stochastic processes and statistical inference. *Arkiv för Matematik*, 1(3):195–277, 1950.

[44] D.P. Hardin and E.B. Saff. Discretizing manifolds via minimum energy points. *Notices of the AMS*, 51(10):1186–1194, 2004.

[45] P. Hennig, M.A. Osborne, and M. Girolami. Probabilistic numerics and uncertainty in computations. *Proc. Royal Soc. A*, 471(2179):20150142, 2015.

[46] J. Hensman, N. Durrande, and A. Solin. Variational Fourier features for Gaussian processes. *Journal of Machine Learning Research*, 18:1–52, 2018.

[47] F.J. Hickernell. A generalized discrepancy and quadrature error bound. *Mathematics of Computation*, 67(221):299–322, 1998.

[48] F. Huszár and D. Duvenaud. Optimally-weighted herding is Bayesian quadrature. In *Proceedings 28th Conference on Uncertainty in Artificial Intelligence (UAI'12)*, pages 377–385, Catalina Island, CA, August 2012. AUAI Press Arlington, Virginia. arXiv preprint arXiv:1204.1664.

[49] M.E. Johnson, L.M. Moore, and D. Ylvisaker. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26:131–148, 1990.

[50] V.R. Joseph, T. Dasgupta, R. Tuo, and C.F.J. Wu. Sequential exploration of complex surfaces using minimum energy designs. *Technometrics*, 57(1):64–74, 2015.

[51] V.R. Joseph, E. Gul, and S. Ba. Maximum projection designs for computer experiments. *Biometrika*, 102(2):371–380, 2015.

[52] T. Karvonen, C.J. Oates, and S. Särkkä. A Bayes-Sard cubature method. In *Advances in Neural Information Processing Systems*, pages 5886–5897, 2018.

[53] T. Karvonen and S. Särkkä. Classical quadrature rules via Gaussian processes. In *27th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2017.

[54] J. Kiefer and J. Wolfowitz. Optimum designs in regression problems. *Annals of Math. Stat.*, 30:271–294, 1959.

[55] N.M. Korobov. Properties and calculation of optimal coefficients. *Doklady Akademii Nauk SSSR*, 132(5):1009–1012, 1960.

[56] N.S. Landkof. *Foundations of Modern Potential Theory*. Springer, Berlin, 1972.

[57] F.M. Larkin. Gaussian measure in Hilbert space and applications in numerical analysis. *The Rocky Mountain Journal of Mathematics*, 2(3):379–421, 1972.

[58] F.M. Larkin. Probabilistic error estimates in spline interpolation and quadrature. In *IFIP Congress*, pages 605–609, 1974.

[59] R. Lekivetz and B. Jones. Fast flexible space-filling designs for nonrectangular regions. *Quality and Reliability Engineering International*, 31(5):829–837, 2015.

[60] Q. Liu and D. Wang. Stein variational gradient descent: a general purpose Bayesian inference algorithm. *Advances In Neural Information Processing Systems*, pages 2378–2386, 2016. arXiv preprint arXiv:1608.04471v2.

[61] H. Luschgy and G. Pagès. Greedy vector quantization. *Journal of Approximation Theory*, 198:111–131, 2015.

[62] S. Mak and V.R. Joseph. Projected support points, with application to optimal MCMC reduction. *arXiv preprint arXiv:1708.06897*, 2017.

[63] S. Mak and V.R. Joseph. Minimax and minimax projection designs using clustering. *Journal of Computational and Graphical Statistics*, 27(1):166–178, 2018.

[64] S. Mak and V.R. Joseph. Support points. *Annals of Statistics*, 46(6A):2562–2592, 2018.

[65] J. Matousek. On the L2-discrepancy for anchored boxes. *Journal of Complexity*, 14(4):527–556, 1998.

[66] I. Molchanov and S. Zuyev. Variational calculus in the space of measures and optimal design. In A. Atkinson, B. Bogacka, and A. Zhigljavsky, editors, *Optimum Design 2000*, chapter 8, pages 79–90. Kluwer, Dordrecht, 2001.

[67] I. Molchanov and S. Zuyev. Steepest descent algorithm in a space of measures. *Statistics and Computing*, 12:115–123, 2002.

[68] W. Näther. *Effective Observation of Random Fields*, volume 72. Teubner-Texte zur Mathematik, Leipzig, 1985.

[69] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, Philadelphia, 1992.

[70] C.J. Oates, A. Barp, and M. Girolami. Posterior integration on a Riemannian manifold. *arXiv preprint arXiv:1712.01793*, 2018.

[71] C.J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of Royal Statistical Society*, B79(3):695–718, 2017.

[72] A. O'Hagan. Bayes-Hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3):245–260, 1991.

[73] V.I. Paulsen. An introduction to the theory of reproducing kernel Hilbert spaces, 2009. https://www.math.uh.edu/ vern/rkhs.pdf.

[74] L. Pronzato. Minimax and maximin space-filling designs: some properties and methods for construction. *Journal de la Société Française de Statistique*, 158(1):7–36, 2017.

[75] L. Pronzato and A. Pázman. *Design of Experiments in Nonlinear Models. Asymptotic Normality, Optimality Criteria and Small-Sample Properties*. Springer, LNS 212, New York, 2013.

[76] L. Pronzato, H.P. Wynn, and A. Zhigljavsky. Extremal measures maximizing functionals based on simplicial volumes. *Statistical Papers*, 57(4):1059–1075, 2016. hal-01308116.

[77] L. Pronzato and A.A. Zhigljavsky. Algorithmic construction of optimal designs on compact sets for concave and differentiable criteria. *Journal of Statistical Planning and Inference*, 154:141–155, 2014.

[78] L. Pronzato and A.A. Zhigljavsky. Measures minimizing regularized dispersion. *J. Scientific Computing*, 78(3):1550–1570, 2019.

[79] L. Pronzato and A.A. Zhigljavsky. Minimum-energy measures for singular kernels. *Journal of Computational and Applied Mathematics*, 2020. (In revivion, hal-02495643).

[80] C.R. Rao. Diversity and dissimilarity coefficients: a unified approach. *Theoret. Popn Biol.*, 21(1):24–43, 1982.

[81] C.R. Rao and T.K. Nayak. Cross entropy, dissimilarity measures and characterizations of quadratic entropy. *IEEE Transactions on Information Theory*, 31(5):589–593, 1985.

[82] K. Ritter, G.W. Wasilkowski, and H. Woźniakowski. Multivariate integration and approximation for random fields satisfying Sacks-Ylvisaker conditions. *The Annals of Applied Probability*, 5(2):518–540, 1995.

[83] J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–435, 1989.

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

[85] R. Schaback. Error estimates and condition numbers for radial basis function interpolation. *Advances in Computational Mathematics*, 3(3):251–264, 1995.

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

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

[88] 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.

[89] 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.

[90] M.L. Stein. *Interpolation of Spatial Data. Some Theory for Kriging.* Springer, Heidelberg, 1999.

[91] I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52(10):4635–4643, 2006.

[92] Z. Szabó and B. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18:1–29, 2018.

[93] 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.

[94] E. Vazquez and J. Bect. Sequential search based on kriging: convergence analysis of some algorithms. Proc. 58th World Statistics Congress of the ISI, August 21-26, Dublin, Ireland, arXiv preprint arXiv:1111.3866v1, 2011.

[95] H. Wackernagel. *Multivariate Geostatistics. An Introduction with Applications.* Springer, Berlin, 1998.

[96] H.P. Wynn. The sequential generation of D-optimum experimental designs. *Annals of Math. Stat.*, 41:1655–1664, 1970.