SUPPORT POINTS

By Simon Mak and V. Roshan Joseph

Georgia Institute of Technology

Abstract This paper introduces a new way to compact a continuous probability distribution F into a set of representative points called support points. These points are obtained by minimizing the energy distance, a statistical potential measure initially proposed by Székely and Rizzo (2004) for testing goodness-of-fit. The energy distance has two appealing features. First, its distance-based structure allows us to exploit the duality between powers of the Euclidean distance and its Fourier transform for theoretical analysis. Using this duality, we show that support points converge in distribution to F, and enjoy an improved error rate to Monte Carlo for integrating a large class of functions. Second, the minimization of the energy distance can be formulated as a difference-of-convex program, which we manipulate using two algorithms to efficiently generate representative point sets. In simulation studies, support points provide improved integration performance to both Monte Carlo and a specific Quasi-Monte Carlo method. Two important applications of support points are then highlighted: (a) as a way to quantify the propagation of uncertainty in expensive simulations, and (b) as a method to optimally compact Markov chain Monte Carlo (MCMC) samples in Bayesian computation.

1. Introduction. This paper explores a new method for compacting a continuous probability distribution F into a set of representative points (rep-points) for F, which we call *support points*. Support points have many important applications in a wide array of fields, because these point sets provide an improved representation of F compared to a random sample. One such application is to the "small-data" problem of uncertainty propagation, where the use of support points as simulation inputs can allow engineers to quantify the propagation of input uncertainty onto system output at minimum cost. Another important application is to "big-data" problems encountered in Bayesian computation, specifically as a tool for compacting large posterior sample chains from Markov chain Monte Carlo (MCMC) methods [19]. In this paper, we demonstrate the theoretical and practical effectiveness of support points for the general problem of integration, and illustrate its usefulness for the two applications above.

1

We first outline two classes of existing methods for rep-points. The first class consists of the so-called *mse-rep-points* (see, e.g., Chapter 4 of [15]), which minimize the expected distance from a random point drawn from F to its closest rep-point. Also known as principal points [16], mse-rep-points have been employed in a variety of statistical and engineering applications, including quantizer design [22, 49] and optimal stratified sampling [7, 6]. In practice, these rep-points can be generated by first performing k-means clustering [36] on a large batch sample from F, then taking the converged cluster centers as rep-points. One weakness of mse-rep-points, however, is that they do not necessarily converge to F (see, e.g., [71, 60]). The second class of rep-points, called *energy rep-points*, aims to find a point set which minimizes some measure of statistical potential. Included here are the minimum-energy designs in [27] and the minimum Riesz energy points in [3]. While the above point sets converge in distribution to F, its convergence rate is quite slow, both theoretically and in practice [3]. Moreover, the construction of such point sets can be computationally expensive in high dimensions.

The key idea behind support points is that it optimizes a specific potential measure called the *energy distance*, which makes such point sets a type of energy rep-point. First introduced in [62], the energy distance was proposed as a computationally efficient way to evaluate goodness-of-fit (GOF), compared to the classical Kolmogorov-Smirnov (K-S) statistic [30], which is difficult to evaluate in high-dimensions. Similar to the existing energy rep-points above, we show in this paper that support points indeed converge in distribution to F. In addition, we demonstrate the improved error rate of support points over Monte Carlo for integrating a large class of functions. The minimization of this distance can also be formulated as a difference-of-convex (d.c.) program, which allows for efficient generation of support points.

Indeed, the *reverse-engineering* of a GOF test forms the basis for state-of-the-art integration techniques called Quasi-Monte Carlo (QMC) methods (see [10] and [9] for a modern overview). To see this, first let g be a differentiable integrand, and let $\{\mathbf{x}_i\}_{i=1}^n$ be the point set (with empirical distribution, or e.d.f., F_n) used to approximate the desired integral $\int g(\mathbf{x}) \, dF(\mathbf{x})$ with the sample average $\int g(\mathbf{x}) \, dF_n(\mathbf{x})$. For simplicity, assume for now that $F = U[0,1]^p$ is the uniform distribution on the p-dimensional hypercube $[0,1]^p$, the typical setting for QMC. The Koksma-Hlawka inequality (see, e.g., [44]) provides the following upper bound on the integration error I:

$$(1) \quad I(g; F, F_n) \equiv \left| \int g(\mathbf{x}) \, d[F - F_n](\mathbf{x}) \right| \le V_q(g) D_r(F, F_n), \;\; 1/q + 1/r = 1,$$
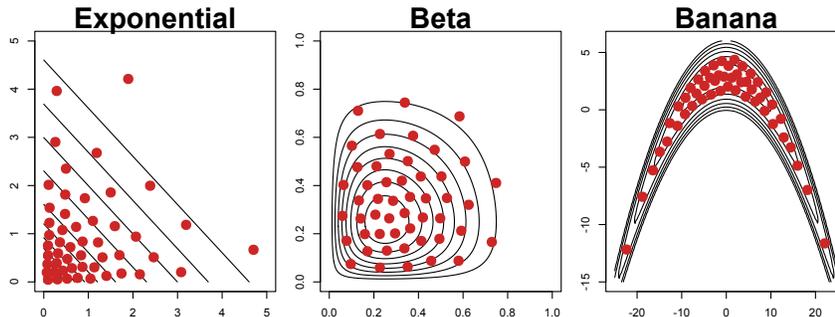
Figure 1: $n = 50$ *support points for 2-d i.i.d.* $Exp(1)$, $Beta(2,4)$ *and the banana-shaped distribution in [21]. Lines represent density contours.*

where $V_q(g) = \|\partial^p g/\partial \mathbf{x}\|_{L_q}$, and $D_r(F, F_n)$ is the L_r-*discrepancy*:

$$(2) \qquad D_r(F, F_n) = \left(\int |F_n(\mathbf{x}) - F(\mathbf{x})|^r \ d\mathbf{x} \right)^{1/r}.$$

The discrepancy $D_r(F, F_n)$ measures how close the e.d.f. F_n is to F, with a smaller value suggesting a better fit. Setting $r = \infty$, the L_∞-discrepancy (or simply *discrepancy*) becomes the classical K-S statistic for testing GOF. In other words, a point set with good fit to F also provides reduced integration errors for a large class of integrands. A more general discussion of this connection in terms of kernel discrepancies can be found in [24].

For a general distribution F, the optimization of $D_r(F, F_n)$ can be a difficult problem. In the uniform setting $F = U[0,1]^p$, there has been some work on directly minimizing the discrepancy $D_\infty(F, F_n)$, including the cdf-rep-points in [15] and the uniform designs in [13]. Such methods, however, are quite computationally expensive, and are applicable only for small point sets on $U[0,1]^p$ (see [14]). Because of this computational burden, modern QMC methods typically use number-theoretic techniques to generate point sets which achieve an *asymptotically* quick decay rate for discrepancy. These include the randomly-shifted lattice rules [58] using the component-by-component implementation of [45] (see also [43]), and the randomly scrambled Sobol' sequences due to [59] and [47]. While most QMC methods consider integration on the uniform hypercube $U[0,1]^p$, there are several ways to map point sets on $U[0,1]^p$ to non-uniform F. One such map is the inverse Rosenblatt transformation [53]; however, it can be computed in closed-form only for a small class of distributions. When the density of F is known up to a proportional constant, the Markov chain Quasi-Monte Carlo (MCQMC) approach [48] can also be used to generate QMC points on F.

Viewed in this light, the energy distance can be seen as a kernel dis-

crepancy [23] for non-uniform distributions, with the specific kernel choice being the negative Euclidean norm. However, in contrast with the typical number-theoretic construction of QMC point sets, support points are instead generated by optimizing the underlying d.c. formulation for the energy distance. This explicit optimization can have both advantages and disadvantages. On one hand, support points can be viewed as *optimal* sampling points of F (in the sense of minimum energy) for any desired sample size n. This optimality is evident in the three examples of support points plotted in Figure 1 – the points are concentrated in regions with high densities, but is sufficiently spread out to maximize the representativeness of each point. Such a "space-filling" property can allow for improved integration performance over existing QMC techniques, which we demonstrate in Section 4. On the other hand, the computational work for optimization can grow quickly when the desired sample size or dimension increases. To this end, we propose two algorithms which exploit the appealing d.c. formulation to efficiently generate point sets as large as 10,000 points in dimensions as large as 500.

This paper is organized as follows. Section 2 proves several important theoretical properties of support points. Section 3 proposes two algorithms for efficiently generating support points. Section 4 outlines several simulations comparing the integration performance of support points with MC and an existing QMC method. Section 5 gives two important applications of support points in uncertainty propagation and Bayesian computation. Section 6 concludes with directions for future research.

2. Support points.

2.1. *Definition.* Let us first define the energy distance between two distributions F and G:

DEFINITION 1 (Energy distance; Def. 1 of [63]). *Let F and G be two distribution functions (d.f.s) on $\varnothing \neq \mathcal{X} \subseteq \mathbb{R}^p$ with finite means, and let $\mathbf{X}, \mathbf{X}' \overset{i.i.d.}{\sim} G$ and $\mathbf{Y}, \mathbf{Y}' \overset{i.i.d.}{\sim} F$. The* energy distance *between F and G is defined as:*

$$(3) \qquad E(F, G) \equiv 2\mathbb{E}\|\mathbf{X} - \mathbf{Y}\|_2 - \mathbb{E}\|\mathbf{X} - \mathbf{X}'\|_2 - \mathbb{E}\|\mathbf{Y} - \mathbf{Y}'\|_2.$$

When $G = F_n$ is the e.d.f. for $\{\mathbf{x}_i\}_{i=1}^n \subseteq \mathcal{X}$, this energy distance becomes:

$$(4) \quad E(F, F_n) = \frac{2}{n} \sum_{i=1}^n \mathbb{E}\|\mathbf{x}_i - \mathbf{Y}\|_2 - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \|\mathbf{x}_i - \mathbf{x}_j\|_2 - \mathbb{E}\|\mathbf{Y} - \mathbf{Y}'\|_2.$$

For brevity, F is assumed to be a continuous d.f. on $\varnothing \neq \mathcal{X} \subseteq \mathbb{R}^p$ with finite mean for the remainder of the paper.

The energy distance $E(F, F_n)$ was originally proposed in [62] as an efficient GOF test for high-dimensional data. In this light, support points are defined as the point set with best GOF under $E(F, F_n)$:

DEFINITION 2 (Support points). *Let* $\mathbf{Y} \sim F$. *For a fixed point set size* $n \in \mathbb{N}$, *the* support points *of F are defined as:*
(O)

$$\{\boldsymbol{\xi}_i\}_{i=1}^n \in \operatorname*{Argmin}_{\mathbf{x}_1,\cdots,\mathbf{x}_n} E(F, F_n) = \operatorname*{Argmin}_{\mathbf{x}_1,\cdots,\mathbf{x}_n} \left\{ \frac{2}{n} \sum_{i=1}^n \mathbb{E}\|\mathbf{x}_i - \mathbf{Y}\|_2 - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \|\mathbf{x}_i - \mathbf{x}_j\|_2 \right\}.$$

The minimization of $E(F, F_n)$ is justified by the following metric property:

THEOREM 1 (Energy distance, Prop. 2 of [63]). $E(F, G) \geq 0$, *with equality holding if and only if $F=G$.*

This theorem shows that the energy between two distributions is always non-negative, and equals zero if and only if these distributions are the same. In this sense, $E(F, G)$ can be viewed as a metric on the space of distribution functions. Support points, being the point set which minimizes such a metric, can then be interpreted as optimal sampling points which best represent F.

The choice of the energy distance $E(F, F_n)$ as an optimization objective is similar to its appeal in GOF testing. As mentioned in the Introduction, $E(F, F_n)$ was originally proposed as an efficient alternative to classical K-S statistic. However, not only is $E(F, F_n)$ easy-to-evaluate, it also has a desirable formulation as a d.c. program. We present in Section 3 two algorithms which exploits this structure to efficiently generate support points.

In the univariate setting of $p = 1$, an interesting equivalence can be established between support points and optimal L_2-discrepancy points:

PROPOSITION 1 (Optimal L_2-discrepancy). *For a univariate d.f. F, the support points of F are equal to the point set with minimal L_2-discrepancy.*

PROOF. It can be shown [61] that $E(F, F_n) = 2D_2^2(F, F_n)$, where F_n is the e.d.f. of $\{x_i\}_{i=1}^n \subseteq \mathcal{X} \subseteq \mathbb{R}$ and $D_2(F, F_n)$ is the one-dimensional L_2-discrepancy in (2). This proves the assertion. $\square$

Unfortunately, such an equivalence fails to hold for $p > 1$, since the L_2-discrepancy is not rotation-invariant. Support points and optimal L_2-discrepancy points can therefore behave quite differently in the multivariate setting.

2.2. *Theoretical properties.*   While the notion of reverse engineering the energy distance is intuitively appealing, some theory is needed to demonstrate why the resulting points are appropriate for (a) representing the desired distribution F, and (b) integrating under F. To this end, we provide three theorems: the first proves the distributional convergence of support points to F, the second establishes a Koksma-Hlawka-like bound connecting integration error with $E(F, F_n)$, and the last provides an existence result for the resulting error convergence rate. The proofs of these results rely on the important property that, for generalized functions, the Fourier transform of the Euclidean norm $\| \cdot \|_2$ is proportional to the same norm raised to some power (see pg. 173-174 in [17]). We refer to various forms of this *duality* property throughout the proofs.

2.2.1. *Convergence in distribution.*   We first address the distributional convergence of support points to the desired distribution F:

THEOREM 2 (Distributional convergence).   *Let $\mathbf{X} \sim F$ and $\mathbf{X}_n \sim F_n$, where F_n is the e.d.f. of the support points in* (O). *Then $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$.*

This relies on the following lemma, which slightly extends the Lévy continuity theorem to the almost-everywhere (a.e.) pointwise convergence setting.

LEMMA 1.   *Let $(F_n)_{n=1}^{\infty}$ be a sequence of d.f.s with characteristic functions (c.f.s) $(\phi_n(\mathbf{t}))_{n=1}^{\infty}$, and let F be a d.f. with c.f. $\phi(\mathbf{t})$. If $\mathbf{X}_n \sim F_n$ and $\mathbf{X} \sim F$, with $\lim_{n \to \infty} \phi_n(\mathbf{t}) = \phi(\mathbf{t})$ a.e. (in the Lebesgue sense), then $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$.*

PROOF. (Lemma 1) See Appendix A.1 of supplementary materials.   □

PROOF. (Theorem 2) Define the sequence of random variables $(\mathbf{Y}_i)_{i=1}^{\infty} \overset{i.i.d.}{\sim} F$, and let $\tilde{F}_n$ denote the e.d.f. of $\{\mathbf{Y}_i\}_{i=1}^{n}$. By the Glivenko-Cantelli lemma, $\lim_{n \to \infty} \sup_{\mathbf{x} \in \mathbb{R}^p} |\tilde{F}_n(\mathbf{x}) - F(\mathbf{x})| = 0$ a.s., so $\tilde{F}_n(\mathbf{x}) \to F(\mathbf{x})$ a.s. for all $\mathbf{x}$. Let $\phi(\mathbf{t})$ and $\tilde{\phi}_n(\mathbf{t})$ denote the c.f.s of F and $\tilde{F}_n$, respectively. Since $|\exp(i\langle \mathbf{t}, \mathbf{x} \rangle)| \leq 1$, applying the Portmanteau theorem (Theorem 8.4.1 in [52]) and the dominated convergence theorem gives:

$$(5) \qquad\qquad \lim_{n \to \infty} \mathbb{E}[|\phi(\mathbf{t}) - \tilde{\phi}_n(\mathbf{t})|^2] = 0.$$

Using Prop. 1 of [63] (this is a *duality* result connecting the energy distance with c.f.s), the expected energy between $\tilde{F}_n$ and F becomes:

$$(6) \quad \mathbb{E}[E(F, \tilde{F}_n)] = \frac{1}{a_p} \mathbb{E}\left[\int \frac{|\phi(\mathbf{t}) - \tilde{\phi}_n(\mathbf{t})|^2}{\|\mathbf{t}\|_2^{p+1}} \, d\mathbf{t} \right] = \frac{1}{a_p} \int \frac{\mathbb{E}\left[|\phi(\mathbf{t}) - \tilde{\phi}_n(\mathbf{t})|^2 \right]}{\|\mathbf{t}\|_2^{p+1}} \, d\mathbf{t},$$

where a_p is some constant depending on p, with the last step following from Fubini's theorem. Note that $\mathbb{E}\left[|\phi(\mathbf{t}) - \tilde{\phi}_n(\mathbf{t})|^2\right] = \frac{1}{n}\text{Var}\left[\exp(\text{i}\langle\mathbf{t}, \mathbf{Y}_1\rangle)\right]$, so $\mathbb{E}\left[|\phi(\mathbf{t}) - \tilde{\phi}_n(\mathbf{t})|^2\right]$ is monotonically decreasing in n. By the monotone convergence theorem and (5), we have:

$$(7) \qquad \lim_{n\to\infty}\mathbb{E}[E(F, \tilde{F}_n)] = \frac{1}{a_p}\int\lim_{n\to\infty}\frac{\mathbb{E}[|\phi(\mathbf{t}) - \tilde{\phi}_n(\mathbf{t})|^2]}{\|\mathbf{t}\|_2^{p+1}}\,d\mathbf{t} = 0.$$

Consider now the e.d.f.s $(F_n)_{n=1}^\infty$ and c.f.s $(\phi_n)_{n=1}^\infty$ for support points. By Definition 2, $E(F, F_n) \leq \mathbb{E}[E(F, \tilde{F}_n)]$, so $\lim_{n\to\infty} E(F, F_n) = 0$ by (7) and the squeeze theorem. Take any subsequence $(n_k)_{k=1}^\infty \subseteq \mathbb{N}_+$, and note that:

$$\lim_{k\to\infty} E(F, F_{n_k}) = \lim_{k\to\infty}\int\frac{|\phi(\mathbf{t}) - \phi_{n_k}(\mathbf{t})|^2}{\|\mathbf{t}\|_2^{p+1}}\,d\mathbf{t} = 0.$$

We know by the Riesz-Fischer Theorem (pg. 148 in [54]) that a sequence of functions (f_n) which converge to f in L_2 has a subsequence which converges pointwise a.e. to f. Applied here, this suggests the existence of a further subsequence $(n_k')_{k=1}^\infty \subseteq (n_k)_{k=1}^\infty$ satisfying $\phi_{n_k'}(\mathbf{t}) \overset{k\to\infty}{\to} \phi(\mathbf{t})$ a.e., so by Lemma 1, $\mathbf{X}_{n_k'} \overset{d}{\to} \mathbf{X}$. Since $(n_k)_{k=1}^\infty$ was arbitrarily chosen, it follows by the proof of Corollary 1 in Chapter 9 of [57] that $\mathbf{X}_n \overset{d}{\to} \mathbf{X}$, which is as desired. $\square$

In words, this theorem shows that support points are indeed representative of the desired distribution F when the number of points n grows large. From this, the *consistency* of support points can be established:

COROLLARY 1 (Consistency).   *Let $\mathbf{X} \sim F$ and $\mathbf{X}_n \sim F_n$, with F_n as in Theorem 2. (a) If $g : \mathcal{X} \to \mathbb{R}$ is continuous, then $g(\mathbf{X}_n) \overset{d}{\to} g(\mathbf{X})$. (b) If g is continuous and bounded, then $\lim_{n\to\infty}\mathbb{E}[g(\mathbf{X}_n)] = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n g(\boldsymbol{\xi}_i) = \mathbb{E}[g(\mathbf{X})].$*

PROOF. Part (a) follows from the continuous mapping theorem and Theorem 2. Part (b) follows by the Portmanteau theorem. $\square$

The purpose of this corollary is two-fold: it demonstrates the consistency of support points for integration, and justifies the use of these point sets for a variety of other applications. Specifically, part (a) shows that support points are appropriate for performing uncertainty propagation in stochastic simulations, an application further explored in Section 4.2. Part (b) shows that any continuous and bounded integrand g can be consistently estimated using support points, i.e., its sample average converges to the desired integral.

2.2.2. *A Koksma-Hlawka-like bound.* Next, we present a theorem which upper bounds the squared integration error $I^2(g; F, F_n)$ by a term proportional to $E(F, F_n)$ for a large class of integrands. Such a result provides some justification on why the energy distance may be a *good* criterion for integration. Here, we first provide a brief review of conditionally positive definite (c.p.d.) kernels, its native spaces, and their corresponding reproducing kernels, three ingredients which will be used for proving the desired theorem.

Consider the following definition of a *conditionally positive definite kernel*:

DEFINITION 3 (c.p.d. kernel; Def. 8.1 of [67]). *A continuous function* $\Phi : \mathbb{R}^p \to \mathbb{R}$ *is a* c.p.d. kernel *of order m if, for all pairwise distinct* $\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathbb{R}^p$ *and all* $\boldsymbol{\zeta} \in \mathbb{R}^N \setminus \{0\}$ *satisfying* $\sum_{j=1}^N \zeta_j p(\mathbf{x}_j) = 0$ *for all polynomials of degree less than m, the quadratic form* $\sum_{j=1}^N \sum_{k=1}^N \zeta_j \zeta_k \Phi(\mathbf{x}_j - \mathbf{x}_k)$ *is positive.*

Similar to the theory of positive definite kernels (see, e.g., Section 10.1 and 10.2 of [67]), one can use a c.p.d. kernel Φ to construct a reproducing kernel Hilbert space (RKHS) along with its reproducing kernel. This is achieved using the so-called *native space* of Φ:

DEFINITION 4 (Native space; Def. 10.16 of [67]). *Let* $\Phi : \mathbb{R}^p \to \mathbb{R}$ *be a c.p.d. kernel of order $m \geq 1$, and let* $\mathcal{P} = \pi_{m-1}(\mathbb{R}^p)$ *be the space of polynomials with degree less than m. Define the linear space:*

$$\mathcal{F}_\Phi(\mathbb{R}^p) = \left\{ f(\cdot) = \sum_{j=1}^N \zeta_j \Phi(\mathbf{x}_j - \cdot) \; : \; \begin{array}{l} N \in \mathbb{N}; \; \boldsymbol{\zeta} \in \mathbb{R}^N; \; \mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathbb{R}^p, \\ \sum_{j=1}^N \zeta_j p(\mathbf{x}_j) = 0 \text{ for all } p \in \mathcal{P} \end{array} \right\},$$

endowed with the inner product:

$$\left\langle \sum_{j=1}^N \zeta_j \Phi(\mathbf{x}_j - \cdot), \sum_{k=1}^M \zeta'_k \Phi(\mathbf{y}_k - \cdot) \right\rangle_\Phi = \sum_{j=1}^N \sum_{k=1}^M \zeta_j \zeta'_k \Phi(\mathbf{x}_j - \mathbf{y}_k).$$

Let $\{\boldsymbol{\psi}_1, \cdots, \boldsymbol{\psi}_m\} \subseteq \mathbb{R}^p, m = \dim(\mathcal{P})$ *be a $\mathcal{P}$-unisolvent subset[1], and let* $\{p_1, \cdots, p_m\} \subseteq \mathcal{P}$ *be a Lagrange basis of $\mathcal{P}$ for such a subset. Furthermore, define the projective map* $\Pi_\mathcal{P} : C(\mathbb{R}^p)[2] \to \mathcal{P}$ *as* $\Pi_\mathcal{P}(f) = \sum_{k=1}^m f(\boldsymbol{\psi}_k) p_k$, *and the map* $\mathcal{R} : \mathcal{F}_\Phi(\mathbb{R}^p) \to C(\mathbb{R}^p)$ *as* $\mathcal{R}f(\mathbf{x}) = f(\mathbf{x}) - \Pi_\mathcal{P} f(\mathbf{x})$. *The* native space *for Φ is then defined as:*

$$\mathcal{N}_\Phi(\mathbb{R}^p) = \mathcal{R}(\mathcal{F}_\Phi(\mathbb{R}^p)) + \mathcal{P},$$

---

[1]See Definition 2.6 of [67].

[2]$C(\mathbb{R}^p)$ is the space of continuous functions on $\mathbb{R}^p$.

and is equipped with the semi-inner product:

$$\langle f, g \rangle_{\mathcal{N}_\Phi(\mathbb{R}^p)} = \langle \mathcal{R}^{-1}(f - \Pi_\mathcal{P} f), \mathcal{R}^{-1}(g - \Pi_\mathcal{P} g) \rangle_\Phi.$$

After obtaining the native space $\mathcal{N}_\Phi(\mathbb{R}^p)$, one can then define an appropriate inner product on $\mathcal{N}_\Phi(\mathbb{R}^p)$ to transform it into a RKHS:

THEOREM 3 (Native space to RKHS; Thm. 10.20 of [67]). *The native space $\mathcal{N}_\Phi(\mathbb{R}^p)$ for a c.p.d. kernel Φ carries the inner product $\langle f, g \rangle = \langle f, g \rangle_{\mathcal{N}_\Phi(\mathbb{R}^p)} + \sum_{k=1}^m f(\boldsymbol{\psi}_k) g(\boldsymbol{\psi}_k)$. With this inner product, $\mathcal{N}_\Phi(\mathbb{R}^p)$ becomes a reproducing kernel Hilbert space with reproducing kernel:*

$$k(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x} - \mathbf{y}) - \sum_{k=1}^m p_k(\mathbf{x}) \Phi(\boldsymbol{\psi}_k - \mathbf{y}) - \sum_{l=1}^m p_l(\mathbf{y}) \Phi(\mathbf{x} - \boldsymbol{\psi}_l)$$
$$+ \sum_{k=1}^m \sum_{l=1}^m p_k(\mathbf{x}) p_l(\mathbf{y}) \Phi(\boldsymbol{\psi}_k - \boldsymbol{\psi}_l) + \sum_{k=1}^m p_k(\mathbf{x}) p_k(\mathbf{y}).$$

The following *generalized Fourier transform* (GFT) will also be useful:

DEFINITION 5 (GFT; Defs. 8.8, 8.9 of [67]). *Suppose $f : \mathbb{R}^p \to \mathbb{C}$ is continuous and slowly increasing. A measurable function $\hat{f} \in {}^3 L_2^{loc}(\mathbb{R}^p \setminus \{0\})$ is called the* generalized Fourier transform *of f if $\exists m \in \mathbb{N}_0/2$ such that $\int_{\mathbb{R}^p} f(\mathbf{x}) \hat{\gamma}(\mathbf{x}) \, d\mathbf{x} = \int_{\mathbb{R}^p} \hat{f}(\omega) \gamma(\omega) \, d\omega$ is satisfied for all $\gamma \in \mathcal{S}_{2m}$, where $\hat{\gamma}$ denotes the standard Fourier transform of γ. Here, $\mathcal{S}_{2m} = \{\gamma \in \mathcal{S} \, : \, \gamma(\omega) = \mathcal{O}(\|\omega\|_2^{2m}) \text{ for } \|\omega\|_2 \to 0\}$, where $\mathcal{S}$ is the Schwartz space.*

Specific definitions for slowly increasing functions and Schwartz spaces can be found in Definitions 5.19 and 5.17 of [67]. Here, the order of the GFT $\hat{f}$ refers to the value m in Definition 5, which can reside on the half-integers $\mathbb{N}_0/2$ since the index of the underlying space $\mathcal{S}_{2m}$ will still be an integer.

With these concepts in hand, we now present the Koksma-Hlawka-like bound. As demonstrated below, the choice of the negative distance kernel $\Phi = -\|\cdot\|_2$ is important for connecting integration error with the distance-based energy distance $E(F, F_n)$.

THEOREM 4 (Koksma-Hlawka). *Let $\{\mathbf{x}_i\}_{i=1}^n \subseteq \mathcal{X} \subseteq \mathbb{R}^p$ be a point set with e.d.f. F_n, and let $\Phi(\mathbf{x}) = -\|\mathbf{x}\|_2$. Then Φ is a c.p.d. kernel of order 1. Moreover:*

${}^3 L_2^{loc}$ denotes the space of locally L_2-integrable functions.

(a) *The native space of* Φ, $\mathcal{N}_\Phi(\mathbb{R}^p)$, *can be explicitly written as:*

(8)
$$
\mathcal{N}_\Phi(\mathbb{R}^p) = \left\{ f \in C(\mathbb{R}^p): \begin{array}{l} \text{(G1)}\ \exists m \in \mathbb{N}_0\ s.t.\ f(\mathbf{x}) = \mathcal{O}(\|\mathbf{x}\|_2^m)\ for\ \|\mathbf{x}\|_2 \to \infty \\ \text{(G2)}\ f\ has\ a\ GFT\ \hat{f}\ of\ order\ 1/2 \\ \text{(G3)}\ \int \|\omega\|_2^{p+1}|\hat{f}(\omega)|^2\ d\omega < \infty \end{array} \right\},
$$

with semi-inner product given by:

(9)
$$
\langle f, g \rangle_{\mathcal{N}_\Phi(\mathbb{R}^p)} = \left\{ \Gamma((p+1)/2)2^p \pi^{(p-1)/2} \right\}^{-1} \int \hat{f}(\omega)\overline{\hat{g}(\omega)}\|\omega\|_2^{p+1}\ d\omega,
$$

(b) *Consider the function space* $\mathcal{G}_p = \mathcal{N}_\Phi(\mathbb{R}^p)$, *equipped with inner product* $\langle f, g \rangle_{\mathcal{G}_p} = \langle f, g \rangle_{\mathcal{N}_\Phi(\mathbb{R}^p)} + f(\boldsymbol{\psi})g(\boldsymbol{\psi})$ *for a fixed choice of* $\boldsymbol{\psi} \in \mathcal{X}$. *Then* $(\mathcal{G}_p, \langle \cdot, \cdot \rangle_{\mathcal{G}_p})$ *is a RKHS, and for any integrand* $g \in \mathcal{G}_p$, *the integration error in* (1) *is bounded by:*

(10)
$$
I(g; F, F_n) \le \|g\|_{\mathcal{G}_p}\sqrt{E(F, F_n)}, \quad \|g\|_{\mathcal{G}_p}^2 \equiv \langle g, g \rangle_{\mathcal{G}_p}.
$$

PROOF. (Theorem 4) Consider first part (a). Let $\Phi(\cdot) = -\|\cdot\|_2$, and let $\hat{\Phi}$ be its GFT of order 1. From Theorem 8.16 of [67], we have the following *duality* representation:

$$
\hat{\Phi}(\omega) = \frac{2^{p/2}\Gamma((p+1)/2)}{\sqrt{\pi}}\|\omega\|_2^{-p-1}, \quad \omega \in \mathbb{R}^p \setminus \{0\}.
$$

By Corollary 8.18 of [67], $\Phi(\cdot)$ is also c.p.d. of order 1. Using the fact that $\Phi(\cdot)$ is even along with the continuity of $\hat{\Phi}(\omega)$ on $\mathbb{R}^p \setminus \{0\}$, an application of Theorem 10.21 in [67] completes the proof for part (a).

Consider now part (b). By Prop. 3 of [63], the kernel $\Phi(\cdot)$ is c.p.d. with respect to the space of constant functions $\mathcal{P} = \{f(\mathbf{x}) \equiv C \text{ for some } C \in \mathbb{R}\}$, with $\dim \mathcal{P} = 1$. Note that any choice of $\boldsymbol{\psi} \in \mathcal{X}$ provides a $\mathcal{P}$-unisolvent subset, with the Lagrange basis for the single point $\boldsymbol{\psi}$ being the unit function $p(\cdot) \equiv 1$. Hence, by Theorem 3, the native space $\mathcal{N}_\Phi(\mathbb{R}^p)$ can be transformed into a RKHS $\mathcal{G}_p$ by equipping it with a new inner product $\langle f, g \rangle_{\mathcal{G}_p} = \langle f, g \rangle_{\mathcal{N}_\Phi(\mathbb{R}^p)} + f(\boldsymbol{\psi})g(\boldsymbol{\psi})$. From the same theorem, the corresponding reproducing kernel for the RKHS $(\mathcal{G}_p, \langle \cdot, \cdot \rangle_{\mathcal{G}_p})$ becomes $\tilde{k}(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x} - \mathbf{y}) - \Phi(\boldsymbol{\psi} - \mathbf{y}) - \Phi(\boldsymbol{\psi} - \mathbf{x}) + 1$.

Next, let $\tilde{k}_\mathbf{x}(\mathbf{z}) = \tilde{k}(\mathbf{x}, \mathbf{z})$. We claim the function $\int \tilde{k}_\mathbf{x}(\cdot)\ d[F - F_n](\mathbf{x})$ belongs in $\mathcal{G}_p$. To see this, define the linear operator $\mathcal{L} : \mathcal{G}_p \to \mathbb{R}$ as $\mathcal{L}f =$

$\int f(\mathbf{x})\, dF(\mathbf{x})$. Note that $\mathcal{L}$ is a bounded operator, because for all $f \in \mathcal{G}_p$:

$$
\begin{aligned}
|\mathcal{L}f| = \left| \int f(\mathbf{x})\, dF(\mathbf{x}) \right| &\leq \int |f(\mathbf{x})|\, dF(\mathbf{x}) \\
&= \int |\langle f(\cdot), \tilde{k}_\mathbf{x}(\cdot) \rangle_{\mathcal{G}_p}|\, dF(\mathbf{x}) \quad \text{(RKHS reproducing property)} \\
&\leq \int \|f\|_{\mathcal{G}_p} \|\tilde{k}_\mathbf{x}(\cdot)\|_{\mathcal{G}_p}\, dF(\mathbf{x}) \quad \text{(Cauchy-Schwarz)} \\
&= \|f\|_{\mathcal{G}_p} \int \tilde{k}^{1/2}(\mathbf{x}, \mathbf{x})\, dF(\mathbf{x}), \quad \text{(RKHS kernel trick)}
\end{aligned}
$$

and the last expression must be bounded because $\int \tilde{k}^{1/2}(\mathbf{x}, \mathbf{x})\, dF(\mathbf{x}) \leq [\int \tilde{k}(\mathbf{x}, \mathbf{x})\, dF(\mathbf{x})]^{1/2}$, the latter of which is finite due to the earlier finite mean assumption on F. By the Riesz Representation Theorem (Theorem 8.12, [25]), there exists a unique $\tilde{f} \in \mathcal{G}_p$ satisfying $\mathcal{L}f = \int f(\mathbf{x})\, dF(\mathbf{x}) = \langle f, \tilde{f} \rangle_{\mathcal{G}_p}$ for all $f \in \mathcal{G}_p$. Setting $f(\mathbf{x}) = \tilde{k}_\mathbf{z}(\mathbf{x})$ in this expression, we get $\int \tilde{k}_\mathbf{z}(\mathbf{x})\, dF(\mathbf{x}) = \langle \tilde{k}_\mathbf{z}(\cdot), \tilde{f} \rangle_{\mathcal{G}_p} = \tilde{f}(\mathbf{z})$ by the RKHS reproducing property, so $\tilde{f} = \int \tilde{k}_\mathbf{x}(\cdot)\, dF(\mathbf{x}) \in \mathcal{G}_p$. Finally, note that $\int \tilde{k}_\mathbf{x}(\cdot)\, dF_n(\mathbf{x}) \in \mathcal{G}_p$ because a RKHS is closed under addition, so $\int \tilde{k}_\mathbf{x}(\cdot)\, d[F - F_n](\mathbf{x}) \in \mathcal{G}_p$, as desired.

With this in hand, the integration error can be bounded as follows:

$$
\begin{aligned}
I(g; F, F_n) = \left| \int g(\mathbf{x})\, d[F - F_n](\mathbf{x}) \right| \\
= \left| \int \left\langle g(\cdot), \tilde{k}_\mathbf{x}(\cdot) \right\rangle_{\mathcal{G}_p} d[F - F_n](\mathbf{x}) \right| \quad \text{(Reproducing property)} \\
= \left| \left\langle g(\cdot), \int \tilde{k}_\mathbf{x}(\cdot)\, d[F - F_n](\mathbf{x}) \right\rangle_{\mathcal{G}_p} \right| \\
\leq \|g\|_{\mathcal{G}_p} \left\| \int \tilde{k}_\mathbf{x}(\cdot)\, d[F - F_n](\mathbf{x}) \right\|_{\mathcal{G}_p}. \quad \text{(Cauchy-Schwarz)}
\end{aligned}
$$

The last term can be rewritten as:

$$
\begin{aligned}
\sqrt{\left\| \int \tilde{k}_\mathbf{x}(\cdot)\, d[F - F_n](\mathbf{x}) \right\|_{\mathcal{G}_p}^2} &= \sqrt{\left\langle \int \tilde{k}_\mathbf{x}(\cdot)\, d[F - F_n](\mathbf{x}), \int \tilde{k}_\mathbf{y}(\cdot)\, d[F - F_n](\mathbf{y}) \right\rangle_{\mathcal{G}_p}} \\
&= \sqrt{\int \int \langle \tilde{k}_\mathbf{x}(\cdot), \tilde{k}_\mathbf{y}(\cdot) \rangle_{\mathcal{G}_p}\, d[F - F_n](\mathbf{x})\, d[F - F_n](\mathbf{y})} \\
&= \sqrt{\int \int \tilde{k}(\mathbf{x}, \mathbf{y})\, d[F - F_n](\mathbf{x})\, d[F - F_n](\mathbf{y})}
\end{aligned}
$$

$$\text{(Kernel trick)}$$

$$= \sqrt{\int \int \Phi(\mathbf{x} - \mathbf{y}) \, d[F - F_n](\mathbf{x}) \, d[F - F_n](\mathbf{y})}$$

$$= \sqrt{E(F, F_n)}, \qquad\qquad \text{(Equation (4))}$$

where the second-last step follows because $\int \Phi(\boldsymbol{\psi} - \mathbf{y}) \, d[F - F_n](\mathbf{x}) = \int \Phi(\boldsymbol{\psi} - \mathbf{x}) \, d[F - F_n](\mathbf{y}) = \int d[F - F_n](\mathbf{x}) = 0$. This completes the proof. $\square$

The appeal of Theorem 4 is that it connects the integration error $I(g; F, F_n)$ with the energy distance $E(F, F_n)$ for all integrands g in the function space $\mathcal{G}_p$. Similar to the usual Koksma-Hlawka inequality, such a theorem justifies the use of support points for integration, because the integration error for all functions in $\mathcal{G}_p$ can be sufficiently bounded by minimizing $E(F, F_n)$.

A natural question to ask is how large $\mathcal{G}_p$ is compared with the commonly-used Sobolev space $W_{s,2}$, i.e., the set of functions whose s-th order differentials have finite L_2 norm. Such a comparison is particularly important in light of the fact that an anchored variant of the Sobolev space is typically employed in QMC analysis (see, e.g., [9]). Recall that s can be extended to the non-negative real numbers using fractional calculus, in which case $W_{s,2}$ becomes the fractional Sobolev space. By comparing the definition of the fractional Sobolev space in the Fourier domain (see (3.7) in [8]), one can show that $W_{(p+1)/2,2}$ is contained within $\mathcal{G}_p$. Moreover, using the fact that $W_{s,2}$ is a decreasing family as $s > 0$ increases (see paragraph prior to Prop. 1.52 in [2]), it follows that $W_{\lceil (p+1)/2 \rceil,2} \subseteq W_{(p+1)/2,2} \subseteq \mathcal{G}_p$. In fact, for odd dimensions p, Theorem 10.43 of [67] shows that $\mathcal{G}_p$ is indeed equal to the Sobolev space $W_{\lceil (p+1)/2 \rceil,2} = W_{(p+1)/2,2}$, so the embedding result becomes an equality.

Viewing this embedding now in terms of Theorem 4, it follows that all integrands g with square-integrable $\lceil (p+1)/2 \rceil$-th order differentials enjoy the upper bound in (10). Hence, as dimension p grows, an increasing order of smoothness is required for integration using support points, which appears to be a necessary trade-off for the appealing d.c. formulation in (O). This is similar to the anchored Sobolev spaces employed in QMC, which requires integrands to have square-integrable mixed first derivatives.

2.2.3. *Error convergence rate.*   Next, we investigate the convergence rate of $I(g; F, F_n)$ under support points. Under eigenvalue decay conditions, the following theorem establishes an *existence* result, which demonstrates the existence of a point set sequence achieving a particular error rate. An additional theorem then clarifies when such decay conditions are satisfied in practice. The main purpose of these results is to demonstrate the quicker

theoretical convergence of support points over Monte Carlo. From the simulations in Section 4, the rate below does not appear to be tight, and a quicker convergence rate is conjectured in Appendix A.3 of supplementary materials.

THEOREM 5 (Error rate). *Let F_n be the e.d.f. for support points $\{\boldsymbol{\xi}\}_{i=1}^n$, and let $g \in \mathcal{G}_p$. Define the kernel $k(\mathbf{x}, \mathbf{y}) = \mathbb{E}\|\mathbf{x} - \mathbf{Y}\|_2 + \mathbb{E}\|\mathbf{y} - \mathbf{Y}\|_2 - \mathbb{E}\|\mathbf{Y} - \mathbf{Y}'\|_2 - \|\mathbf{x} - \mathbf{y}\|_2$, $\mathbf{Y}, \mathbf{Y}' \overset{i.i.d.}{\sim} F$. If (a) $\mathbb{E}[\|\mathbf{Y}\|_2^3] < \infty$, and (b) the weighted eigenvalues of k under F satisfy $\sum_{k=1}^\infty \lambda_k^{1/\alpha} < \infty$ for some $\alpha > 1$, then:*

$$(11) \qquad I(g; F, F_n) = \mathcal{O}\{\|g\|_{\mathcal{G}_p} n^{-1/2} (\log n)^{-(\alpha-1)/2}\},$$

with constant terms depending on α and p.

Here, the weighted eigenvalue sequence of k under F is the decreasing sequence $(\lambda_k)_{k=1}^\infty$ satisfying $\lambda_k \phi_k(\mathbf{x}) = \mathbb{E}[k(\mathbf{x}, \mathbf{Y})\phi_k(\mathbf{Y})]$, $\mathbb{E}[\phi_k^2(\mathbf{Y})] = 1$.

The proof of this theorem exploits the fact that $E(F, F_n)$ is a goodness-of-fit statistic. Specifically, writing $E(F, F_n)$ as a degenerate V-statistic V_n, we appeal to its limiting distribution and a uniform Barry-Esseen-like rate to derive an upper bound for the minimum of V_n. The full proof is outlined below, and relies on the following lemmas.

LEMMA 2. ([56]) *Let $(\mathbf{Y}_i)_{i=1}^\infty \overset{i.i.d.}{\sim} F$, and let k be a symmetric, positive definite (p.d.) kernel with $\mathbb{E}[k(\mathbf{x}, \mathbf{Y}_1)] = 0$, $\mathbb{E}[k^2(\mathbf{Y}_1, \mathbf{Y}_2)] < \infty$ and $\mathbb{E}|k(\mathbf{Y}_1, \mathbf{Y}_1)| < \infty$. Define the V-statistic $V_n \equiv n^{-2} \sum_{i=1}^n \sum_{j=1}^n k(\mathbf{Y}_i, \mathbf{Y}_j)$. Then $W_n \equiv nV_n \overset{d}{\to} \sum_{k=1}^\infty \lambda_k \chi_k^2 \equiv W_\infty$, where $(\chi_k^2)_{k=1}^\infty \overset{i.i.d.}{\sim} \chi^2(1)$, and $(\lambda_k)_{k=1}^\infty$ are the weighted eigenvalues of k under F.*

LEMMA 3. ([31]) *Adopt the same notation as in Lemma 2, and let F_{W_n} and F_{W_∞} denote the d.f.s for W_n and W_∞. If $\mathbb{E}[k(\mathbf{x}, \mathbf{Y}_1)] = 0$, $\mathbb{E}|k(\mathbf{Y}_1, \mathbf{Y}_2)|^3 < \infty$ and $\mathbb{E}|k(\mathbf{Y}_1, \mathbf{Y}_1)|^{3/2} < \infty$, then:*

$$(12) \qquad \sup_x |F_{W_n}(x) - F_{W_\infty}(x)| = \mathcal{O}(n^{-1/2}),$$

with constants depending on dimension p.

LEMMA 4 (Paley-Zygmund inequality; [50]). *Let $X \geq 0$, with constants $a_1 > 1$ and $a_2 > 0$ satisfying $\mathbb{E}(X^2) \leq a_1 \mathbb{E}^2(X)$ and $\mathbb{E}(X) \geq a_2$. Then, for any $\theta \in (0, 1)$, $\mathbb{P}(X \geq a_2\theta) \geq (1 - \theta)^2 / a_1$.*

The proof of Theorem 5 then follows:

PROOF. (Theorem 5) Following Section 7.4 of [63], the energy distance $E(F, F_n)$ can be written as the order-2 V-statistic:

$$(13) \qquad E(F, F_n) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) \leq \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(\mathbf{Y}_i, \mathbf{Y}_j) \equiv V_n,$$

where $k(\mathbf{x}, \mathbf{y})$ is defined in Theorem 5 and $(\mathbf{Y}_i)_{i=1}^{n} \overset{i.i.d.}{\sim} F$. The last inequality follows by the definition of support points.

By [69], the kernel k is symmetric and p.d., and the conditions for Lemma 2 can easily be shown to be satisfied. Invoking this lemma, we have:

$$(14) \qquad \inf\{x : F_{W_n}(x) > 0\} = nE(F, F_n),$$

The strategy is to lower bound the left-tail probability of W_∞, then use this to derive an upper bound for $\inf\{x : F_{W_n}(x) > 0\}$ using Lemma 3.

We first investigate the left-tail behavior of W_∞. Define $Z_t = \exp\{-tW_\infty\}$ for some $t > 0$ to be determined later. Since Z_t is bounded a.s., $\mathbb{E}(Z_t) = \prod_{k=1}^{\infty}(1 + 2\lambda_k t)^{-1/2}$ and $\mathbb{E}(Z_t^2) = \prod_{k=1}^{\infty}(1 + 4\lambda_k t)^{-1/2}$. From Lemma 4, it follows that, for fixed $x > 0$, if our choice of t satisfies:

$$(15) \quad [\mathbf{A1}]: \; \mathbb{E}(Z_t) \geq 2\exp\{-tx\} > \exp\{-tx\}, \; [\mathbf{A2}]: \; \mathbb{E}(Z_t^2) \leq a_1 \mathbb{E}^2(Z_t),$$

then, setting $\theta = 1/2$ and $a_2 = 2\exp\{-tx\}$, we have:

$$(16) \qquad F_{W_\infty}(x) = \mathbb{P}(Z_t \geq \exp\{-tx\}) \geq \mathbb{P}(Z_t \geq \mathbb{E}(Z_t)/2) \geq (4a_1)^{-1}.$$

Consider [$\mathbf{A1}$], or equivalently: $tx \geq \log 2 + (1/2)\sum_{k=1}^{\infty} \log(1 + 2\lambda_k t)$. Since $\log(1 + x) \leq x^q/q$ for $x > 0$ and $0 < q < 1$, and $\sum_{k=1}^{\infty} \lambda_k^{1/\alpha} < \infty$ by assumption, a sufficient condition for [$\mathbf{A1}$] is:

$$tx \geq \log 2 + (\alpha/2)\sum_{k=1}^{\infty}(2\lambda_k t)^{1/\alpha} \Leftrightarrow P_\alpha(s) \equiv s^\alpha - b_p s x^{-1} - (\log 2)x^{-1} \geq 0,$$

where $s = t^{1/\alpha}$ and $b_p = \alpha 2^{1/\alpha - 1}\sum_{k=1}^{\infty} \lambda_k^{1/\alpha} > 0$.

Since $\log 2 > 0$ and $b_p s x^{-1} > 0$, there exists exactly one (real) positive root for $P_\alpha(s)$. Call this root r, so the above inequality is satisfied for $s > r$. Define $\bar{P}_\alpha(s)$ as the linearization of $P_\alpha(s)$ for $s > \bar{s} = (b_p x^{-1})^{1/(\alpha-1)}$, i.e.:

$$\bar{P}_\alpha(s) = \begin{cases} P_\alpha(s), & 0 \leq s \leq \bar{s} \\ -x^{-1}\log 2 + P_\alpha'(\bar{s}) \cdot (s - \bar{s}), & s > \bar{s}. \end{cases}$$

From this, the unique root of $\bar{P}_\alpha(s)$ can be shown to be $\bar{r} = \bar{s} + x^{-1}(\log 2)[P'_\alpha(\bar{s})]^{-1}$. Since $P_\alpha(s) \geq \bar{P}_\alpha(s)$ for all $s \geq 0$, $\bar{r} \geq r$, the following upper bound for $\bar{r}$ can be obtained for sufficiently small x:

$$\bar{r} = (b_p x^{-1})^{1/(\alpha-1)} + (\log 2)(\alpha - 1)^{-1} b_p^{-1} \leq 2(b_p x^{-1})^{1/(\alpha-1)}.$$

Hence:

$$(17) \qquad \begin{aligned} t = s^\alpha \geq 2^\alpha (b_p x^{-1})^{\alpha/(\alpha-1)} &\Leftrightarrow s \geq 2(b_p x^{-1})^{1/(\alpha-1)} \geq \bar{r} \geq r \\ &\Rightarrow s^\alpha - b_p x^{-1} s - (\log 2) x^{-1} \geq 0, \end{aligned}$$

so setting $t = 2^\alpha (b_p x^{-1})^{\alpha/(\alpha-1)} \equiv c_p x^{-\alpha/(\alpha-1)}$ satisfies **[A1]** in (15).

The next step is to determine the smallest a_1 satisfying **[A2]** in (15), or equivalently, $\frac{1}{2} \sum_{k=1}^\infty \log(1 + 4\lambda_k t) \geq \sum_{k=1}^\infty \log(1 + 2\lambda_k t) - \log a_1$. Again, since $\log(1 + x) \leq x^q/q$ for $x > 0$ and $0 < q < 1$, a sufficient condition for **[A2]** is:

$$\log a_1 \geq \sum_{k=1}^\infty \log(1 + 2\lambda_k t) \Leftarrow \log a_1 \geq \alpha \sum_{k=1}^\infty (2\lambda_k t)^{1/\alpha}$$

Plugging in $t = c_p x^{-\alpha/(\alpha-1)}$ from (17) and letting $d_p \equiv \alpha(2c_p)^{1/\alpha} \left(\sum_{k=1}^\infty \lambda_k^{1/\alpha} \right)$, we get $\log a_1 \geq d_p x^{-1/(\alpha-1)} \Leftrightarrow a_1 \geq \exp\left\{ d_p x^{-1/(\alpha-1)} \right\}$.

The choice of $t = c_p x^{-\alpha/(\alpha-1)}$ and $a_1 = \exp\left\{ d_p x^{-1/(\alpha-1)} \right\}$ therefore **[A1]** and **[A2]** in (16). It follows from (16) that:

$$(18) \qquad F_{W_\infty}(x) \geq (4a_1)^{-1} = \exp\{-d_p x^{-1/(\alpha-1)}\}/4,$$

so $F_{W_\infty}(x)$ converges to 0 at a rate of $\mathcal{O}(\exp\left\{ -d_p x^{-1/(\alpha-1)} \right\})$ as $x \to 0^+$.

Consider now the behavior of $\inf\{x : F_{W_n}(x) > 0\}$ as $n \to \infty$. From the uniform bound in Lemma 3, there exists a sequence $(c_{n,p})_{n=1}^\infty$, $\lim_{n\to\infty} c_{n,p} = 0$ such that $|F_{W_n}(x) - F_{W_\infty}(x)| \leq c_{n,p} n^{-1/2}$ for all $x \geq 0$. Setting the right side of (18) equal to $2c_{n,p} n^{-1/2}$ and solving for x, we get:

(19)

$$x^* = \frac{d_p^{\alpha-1}}{[\frac{1}{2}\log n - \log(8c_{n,p})]^{\alpha-1}} \Rightarrow F_{W_\infty}(x^*) \geq \exp\left\{ -d_p(x^*)^{-1/(\alpha-1)} \right\} = 2c_{n,p} n^{-1/2}.$$

so Lemma 3 ensures the above choice of x^* satisfies $F_{W_n}(x^*) \geq c_{n,p} n^{-1/2} > 0$.

Using this with (14), it follows that:

$$E(F, F_n) = \mathcal{O}\left\{ n^{-1}(\log n)^{-(\alpha-1)} \right\},$$

with constants depending on p. Finally, by Theorem 4, we have:

$$I(g; F, F_n) = \mathcal{O}\{\|g\|_{\mathcal{G}_p} n^{-1/2}(\log n)^{-(\alpha-1)/2}\}$$

which is as desired. $\square$

The following theorem provides some insight on when the eigenvalue decay condition $\sum_{k=1}^{\infty} \lambda_k^{1/\alpha} < \infty$ in Theorem 5 is satisfied.

THEOREM 6 (Eigenvalue conditions).   *Let F_n and F be as in Theorem 5, and let $g \in \mathcal{G}_p$.*

   (a) If $\mathcal{X} \subseteq \mathbb{R}^p$ is a bounded Borel set with non-empty interior, then $I(g; F, F_n) = \mathcal{O}\{\|g\|_{\mathcal{G}_p} n^{-1/2} (\log n)^{-(1-\nu)/(2p)}\}$ for any $\nu \in (0,1)$,
   (b) If $\mathcal{X} \subseteq \mathbb{R}^p$ is measurable with positive Lebesgue measure, and there exists some $\beta > 0$ and $C \geq 0$ such that:

$$(20) \qquad \limsup_{r \to \infty} r^\beta \int_{\mathcal{X} \setminus B_r(\mathbf{y})} \mathbb{E}\|\mathbf{x} - \mathbf{Y}\|_2 \, dF(\mathbf{x}) \leq C \text{ for all } \mathbf{y} \in \mathcal{X},$$

   then $I(g; F, F_n) = \mathcal{O}\{\|g\|_{\mathcal{G}_p} n^{-1/2} (\log n)^{-(\gamma-\nu)/(2p)}\}$ for any $\nu \in (0, \gamma)$, where $\gamma = \beta/(\beta+1)$ and $B_r(\mathbf{y})$ denotes an r-ball around $\mathbf{y}$.

Here, constant terms may depend on ν, p or β.

PROOF. See Appendix A.2 of supplementary materials. $\square$

In words, Theorem 6 demonstrates the improvement of support points over MC under certain conditions on the sample space $\mathcal{X}$ or the desired distribution F. Specifically, part (a) requires the sample space $\mathcal{X}$ to be bounded with non-empty interior, whereas part (b) relaxes this boundedness restriction on $\mathcal{X}$ at the cost of the mild moment condition (20) on F. This condition holds for a large class of distributions which are not too heavy-tailed.

For illustration, consider the standard normal distribution for F, with sample space $\mathcal{X} = \mathbb{R}^p$. Note that, when $\|\mathbf{x}\|_2$ becomes large, $\mathbb{E}\|\mathbf{x} - \mathbf{Y}\|_2 \approx \|\mathbf{x}\|_2$. Hence, the condition in (20) becomes:

$$\limsup_{r \to \infty} r^\beta P(r), \quad P(r) \equiv (2\pi)^{-p/2} \int_{\mathbb{R}^p \setminus B_r(\mathbf{0})} \|\mathbf{x}\|_2 \exp\{-\|\mathbf{x}\|_2^2/2\} \, d\mathbf{x}.$$

Since $P'(r) \propto -r^p \exp\{-r^2/2\}$, it follows that $P(r) = \mathcal{O}(r^{p-1} \exp\{-r^2/2\})$, so $\limsup_{r \to \infty} r^\beta P(r) = 0$ for all $\beta > 0$. Applying part (b) of Theorem 6, support points enjoy a convergence rate of $\mathcal{O}\{n^{-1/2}(\log n)^{-(1-\nu)/(2p)}\}$ for any $\nu \in (0,1)$ in this case. An analogous argument shows a similar rate holds for any spherically symmetric distribution (see, e.g., [15]) with an exponentially decaying density in its radius.

2.3. *Comparison with MC and existing QMC methods.* We first discuss the implications of Theorems 5 and 6 in comparison to Monte Carlo. Using the law of iterated logarithms [29], one can show that the error convergence rate for MC is bounded a.s. by $\mathcal{O}(n^{-1/2}\sqrt{\log\log n})$ for any distribution F. Comparing this with (11), the error rate of support points is asymptotically quicker than MC by at least some log-factor when dimension p is fixed. This improvement is reflected in the simulations in Section 4, where support points enjoy a considerable improvement over MC for all point set sizes n. When dimension p is allowed to vary (and assuming $\|g\|_{\mathcal{G}_p}$ and $\mathrm{Var}\{g(\mathbf{X})\}$, $\mathbf{X} \sim F$, do not depend on p), note that the MC rate is independent of p, while the rate in (11) can have constants which depend on p. From a theoretical perspective, this suggests support points may be inferior to MC for high-dimensional integration problems. Such a *curse-of-dimensionality*, however, is not observed in our numerical experiments, where support points enjoy a sizable error reduction over MC for p as large as 500.

Compared to existing QMC techniques, the existence rate in Theorem 5 falls short in the uniform setting of $F = U[0,1]^p$. For fixed dimension p, [15] showed that for any integrand g with bounded variation (in the sense of Hardy and Krause), the error rate for classical QMC point sets is $\mathcal{O}\{n^{-1}(\log n)^p\}$, which is faster than (11). Moreover, when p is allowed to vary, it can be shown (see [32, 9]) that certain randomized QMC (RQMC) methods, such as the randomly-shifted lattice rules in [58], enjoy a root-mean-squared error rate of $\mathcal{O}(n^{-1+\delta})$ with $\delta \in (0, 1/2)$, where constant terms do not depend on dimension p. On the other hand, support points provide *optimal* integration points (in the sense of minimum energy) for *non-uniform* distributions at fixed sample size n. Because of this optimality, support points can enjoy reduced errors to existing QMC methods in practice, which we demonstrate later for a specific RQMC method called randomly-scrambled Sobol' sequences [59, 47]. This suggests the rate in Theorem 5 may not be tight, and further theoretical work is needed (we outline one possible proof approach in Appendix A.3 of supplementary materials).

3. Generating support points. The primary appeal of support points is the efficiency by which these point sets can be optimized, made possible by exploiting the d.c. structure of the energy distance. Here, we present two algorithms, `sp.ccp` and `sp.sccp`, which employ a combination of the convex-concave procedure (CCP) with resampling to quickly optimize support points. `sp.ccp` should be used when sample batches are computationally expensive to obtain from F, whereas `sp.sccp` should be used when samples can be easily obtained. We prove the convergence of both algorithms to a stationary

point set, and briefly discuss their running times.

3.1. *Algorithm statements.*   We first present the steps for sp.ccp, then introduce sp.sccp as an improvement on sp.ccp when multiple sample batches from F can be efficiently obtained. Suppose a single sample batch $\{\mathbf{y}_m\}_{m=1}^N$ is obtained from F. Using this, sp.ccp optimizes the following Monte Carlo approximation of the support points formulation (O):

$$\text{(MC)}\quad \underset{\mathbf{x}_1,\cdots,\mathbf{x}_n}{\operatorname{argmin}} \hat{E}(\{\mathbf{x}_i\}; \{\mathbf{y}_m\}) \equiv \frac{2}{nN}\sum_{i=1}^n\sum_{m=1}^N \|\mathbf{y}_m - \mathbf{x}_i\|_2 - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|\mathbf{x}_i - \mathbf{x}_j\|_2.$$

The approximated objective $\hat{E}$ was originally proposed by [62] as a two-sample GOF statistic for testing whether $\{\mathbf{y}_m\}_{m=1}^N$ and $\{\mathbf{x}_i\}_{i=1}^n$ are generated from the same distribution. Posed as an optimization problem, however, the goal in (MC) is to recover the point set which best represents the random sample $\{\mathbf{y}_m\}_{m=1}^N$ from F in terms of goodness-of-fit.

The key observation here is that the objective function $\hat{E}$ can be written as a difference of convex functions in $\mathbf{x} = (\mathbf{x}_1,\cdots,\mathbf{x}_n)$, namely, the two terms in (MC). This structure allows for efficient optimization using d.c. programming methods, which enjoy a well-established theoretical and numerical framework [66, 64]. While global optimization algorithms have been proposed for d.c. programs (e.g., [65]), such methods are typically quite slow in practice [35], and may not be appropriate for the large-scale problem at hand. Instead, we employ a d.c. algorithm called the convex-concave procedure (CCP, see [70]) which, in conjunction with the distance-based property of the energy distance, allows for efficient optimization of (MC).

The main idea in CCP is to first replace the concave term in the d.c. objective with a convex upper bound, then solve the resulting "surrogate" formulation (which is convex) using convex programming techniques. This procedure is then repeated until the solution iterates converge. CCP can be seen as a specific case of majorization-minimization (MM, see [33]), a popular optimization technique in statistics. The key to computational efficiency lies in finding a convex surrogate formulation which can be minimized in closed-form. Here, such a formulation can be obtained by exploiting the distance-based structure of (MC), with its closed-form minimizer given by the iterative map $\mathbf{x}_i^{[l+1]} \leftarrow M_i(\{\mathbf{x}_j^{[l]}\}_{j=1}^n; \{\mathbf{y}_m\}_{m=1}^N)$, $i = 1,\cdots,n$, where M_i is given in (22). The appeal of CCP here is two-fold. First, the evaluation of the iterative maps $M_i, i = 1,\cdots,n$ requires $\mathcal{O}(n^2 p)$ work, thereby allowing for the efficient generation of moderately-sized point sets in moderately-high dimensions. Second, the computation of these maps can be greatly sped up using parallel computing, a point further discussed in Section 3.3.

Algorithm 1 `sp.ccp`: Support points using one sample batch

- Sample $\mathcal{D}^{[0]} = \{\mathbf{x}_i^{[0]}\}_{i=1}^n$ i.i.d. from $\{\mathbf{y}_m\}_{m=1}^N$.
- Set $l = 0$, and **repeat** until convergence of $\mathcal{D}^{[l]}$:
    - **For** $i = 1, \cdots, n$ **do parallel**:
        - Set $\mathbf{x}_i^{[l+1]} \leftarrow M_i(\mathcal{D}^{[l]}; \{\mathbf{y}_m\}_{m=1}^N)$, with M_i defined in (22).
    - Update $\mathcal{D}^{[l+1]} \leftarrow \{\mathbf{x}_i^{[l+1]}\}_{i=1}^n$, and set $l \leftarrow l + 1$.
- Return the converged point set $\mathcal{D}^{[\infty]}$.

Algorithm 2 `sp.sccp`: Support points using multiple sample batches

- Sample $\mathcal{D}^{[0]} = \{\mathbf{x}_i^{[0]}\}_{i=1}^n \overset{i.i.d.}{\sim} F$.
- Set $l = 0$, and **repeat** until convergence of $\mathcal{D}^{[l]}$:
    - Resample $\{\mathbf{y}_m^{[l]}\}_{m=1}^N \overset{i.i.d.}{\sim} F$.
    - **For** $i = 1, \cdots, n$ **do parallel**:
        - Set $\mathbf{x}_i^{[l+1]} \leftarrow M_i(\mathcal{D}^{[l]}; \{\mathbf{y}_m^{[l]}\}_{m=1}^N)$, with M_i defined in (22).
    - Update $\mathcal{D}^{[l+1]} \leftarrow \{\mathbf{x}_i^{[l+1]}\}_{i=1}^n$, and set $l \leftarrow l + 1$.
- Return the converged point set $\mathcal{D}^{[\infty]}$.

Algorithm 1 outlines the detailed steps for `sp.ccp` following the above discussion. One caveat for `sp.ccp` is that it uses only one sample batch from F, even when multiple sample batches can be generated efficiently. This motivates the second algorithm, `sp.sccp`, whose steps are outlined in Algorithm 2. The main difference for `sp.sccp` is that $\{\mathbf{y}_m\}_{m=1}^N$ is *resampled* within each CCP iteration (a procedure known as stochastic MM). This resampling scheme allows `sp.sccp` to converge to a stationary point set for the desired problem (O), which we demonstrate next.

3.2. *Algorithmic convergence.* For completeness, a brief overview of MM is provided, following [33].

DEFINITION 6 (Majorization function). *Let $f : \mathbb{R}^s \to \mathbb{R}$ be the objective function to be minimized. A function $h(\mathbf{z}|\mathbf{z}')$ majorizes $f(\mathbf{z})$ at a point $\mathbf{z}' \in \mathbb{R}^s$ if $h(\mathbf{z}|\mathbf{z}') \geq f(\mathbf{z})$, with equality holding when $\mathbf{z} = \mathbf{z}'$.*

Starting at an initial point $\mathbf{z}^{[0]}$, the goal in MM is to minimize the majorizing function h as a surrogate for the true objective f, and iterate the updates $\mathbf{z}^{[l+1]} \leftarrow \arg\min_{\mathbf{z}} h(\mathbf{z}|\mathbf{z}^{[l]})$ until convergence. This iterative procedure has the so-called descent property $f(\mathbf{x}^{[l+1]}) \leq f(\mathbf{x}^{[l]})$, which ensures solution iterates

are always decreasing in f. The key for efficiency is to find a majorizing function g with a closed-form minimizer which is easy to compute.

Consider now the Monte Carlo approximation in (MC), which has a d.c. formulation in $\{\mathbf{x}_i\}_{i=1}^n$, with concave term $-n^{-2}\sum_{i=1}^n\sum_{j=1}^n\|\mathbf{x}_i - \mathbf{x}_j\|_2$. Following CCP, we first majorize this term using a first-order Taylor expansion at the current iterate $\{\mathbf{x}_j'\}_{j=1}^n$, yielding the surrogate convex program:

(21)
$$\underset{\mathbf{x}_1,\cdots,\mathbf{x}_n}{\operatorname{argmin}} h(\{\mathbf{x}_i\}_{i=1}^n; \{\mathbf{x}_j'\}_{j=1}^n)$$

$$\equiv \frac{2}{nN}\sum_{i=1}^n\sum_{m=1}^N\|\mathbf{y}_m - \mathbf{x}_i\|_2 - \frac{1}{n^2}\left[\sum_{i=1}^n\sum_{j=1}^n\left(\|\mathbf{x}_i' - \mathbf{x}_j'\|_2 + \frac{2(\mathbf{x}_i - \mathbf{x}_i')^T(\mathbf{x}_i' - \mathbf{x}_j')}{\|\mathbf{x}_i' - \mathbf{x}_j'\|_2}\right)\right].$$

Implicit here is the assumption that the current point set is pairwise distinct, i.e., $\mathbf{x}_i' \neq \mathbf{x}_j'$ for all $i, j = 1, \cdots, n$. From simulations, this appears to be always satisfied by initializing the algorithm with a pairwise distinct point set, because the random sampling of $\{\mathbf{y}_m\}$ and the "almost-random" round-off errors [1] in the evaluation of M_i force subsequent point sets to be pairwise distinct. Such an assumption can also be easily checked after each iteration.

While (21) can be solved using gradient-based convex programming techniques, this can be computationally burdensome when n or p becomes large, because such methods may require many evaluations of h and its subgradient. Instead, the following lemma allows us to perform a slight "convexification" of the convex term in (21), which then yields a efficient closed-form minimizer.

LEMMA 5 (Convexification). $Q(\mathbf{x}|\mathbf{x}') = \frac{\|\mathbf{x}\|_2^2}{2\|\mathbf{x}'\|_2} + \frac{\|\mathbf{x}'\|_2}{2}$ *majorizes* $\|\mathbf{x}\|_2$ *at* $\mathbf{x}'$ *for any* $\mathbf{x}' \in \mathbb{R}^p$.

PROOF. See Appendix A.4 of supplementary materials.   □

Lemma 5 has an appealing geometric interpretation. Viewing $\|\mathbf{x}\|_2$ as a second-order cone centered at $\mathbf{0}$, $Q(\mathbf{x}|\mathbf{x}')$ can be interpreted as the tightest convex paraboloid intersecting this cone at $\mathbf{x}'$. Note that the quadratic nature of the majorizer Q, which is crucial for deriving a closed-form minimizer, is made possible by the distance-based structure of the energy distance.

From this, the following lemma provides a quadratic majorizer for (21), along with its corresponding closed-form minimizer:

LEMMA 6 (Closed-form iterations).   *Define the function* h^Q *as:*

$$h^Q(\{\mathbf{x}_i\}_{i=1}^n; \{\mathbf{x}_j'\}_{j=1}^n) \equiv \frac{2}{nN}\sum_{i=1}^n\sum_{m=1}^N\left\{\frac{\|\mathbf{y}_m - \mathbf{x}_i\|_2^2}{2\|\mathbf{y}_m - \mathbf{x}_i'\|_2} + \frac{\|\mathbf{y}_m - \mathbf{x}_i'\|_2}{2}\right\}$$

$$- \frac{1}{n^2} \left[\sum_{i=1}^{n} \sum_{j=1}^{n} \left(\|\mathbf{x}'_i - \mathbf{x}'_j\|_2 + \frac{2(\mathbf{x}_i - \mathbf{x}'_i)^T (\mathbf{x}'_i - \mathbf{x}'_j)}{\|\mathbf{x}'_i - \mathbf{x}'_j\|_2} \right) \right],$$

Then $h^Q(\cdot; \{\mathbf{x}'_j\}_{j=1}^{n})$ majorizes $\hat{E}$ at $\{\mathbf{x}'_j\}_{j=1}^{n}$. Moreover, the global minimizer of $h^Q(\cdot; \{\mathbf{x}'_j\}_{j=1}^{n})$ is given by:

(22)
$$\mathbf{x}_i = M_i(\{\mathbf{x}'_j\}_{j=1}^{n}; \{\mathbf{y}_m\}_{m=1}^{N})$$
$$\equiv \left(\sum_{m=1}^{N} \|\mathbf{x}'_i - \mathbf{y}_m\|_2^{-1} \right)^{-1} \left(\frac{N}{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\mathbf{x}'_i - \mathbf{x}'_j}{\|\mathbf{x}'_i - \mathbf{x}'_j\|_2} + \sum_{m=1}^{N} \frac{\mathbf{y}_m}{\|\mathbf{x}'_i - \mathbf{y}_m\|_2} \right), \ i = 1, \cdots, n.$$

PROOF. See Appendix A.5 of supplementary materials. □

One can now prove the convergence of `sp.ccp` and `sp.sccp`.

THEOREM 7. *(Convergence - `sp.ccp`) Assume $\mathcal{X}$ is closed and convex. For any pairwise distinct $\mathcal{D}^{[0]} \subseteq \mathcal{X}$ and fixed sample batch $\{\mathbf{y}_m\}_{m=1}^{N} \subseteq \mathcal{X}$, the sequence $(\mathcal{D}^{[l]})_{l=1}^{\infty}$ in Algorithm 1 converges to a limiting point set $\mathcal{D}^{[\infty]}$ which is stationary for $\hat{E}$.*

PROOF. See Appendix A.6 of supplementary materials. □

THEOREM 8. *(Convergence - `sp.sccp`) Assume $\mathcal{X}$ is compact and convex. For any pairwise distinct $\mathcal{D}^{[0]} \subseteq \mathcal{X}$, the sequence $(\mathcal{D}^{[l]})_{l=1}^{\infty}$ in Algorithm 2 converges a.s. to a limiting point set $\mathcal{D}^{[\infty]}$ which is stationary for E.*

PROOF. See Appendix A.7 of supplementary materials. □

(Recall that $\mathbf{z} \in D$ is a stationary solution for a function $f : D \subseteq \mathbb{R}^s \to \mathbb{R}$ if:

$$f'(\mathbf{z}, \mathbf{d}) \geq 0 \quad \text{for all } \mathbf{d} \in \mathbb{R}^s \text{ s.t. } \mathbf{z} + \mathbf{d} \in D,$$

where $f'(\mathbf{z}, \mathbf{d})$ is the directional derivative of f at $\mathbf{z}$ in direction $\mathbf{d}$.) Note that the compactness condition on $\mathcal{X}$ in Theorem 8 is needed to prove the convergence of stochastic MM algorithms, since it allows for an application of the law of large numbers (see [37] for details).

3.3. *Running time and parallelization.* Regarding the running time of `sp.ccp`, it is well known that MM algorithms enjoy a linear error convergence rate [46]. This means $L = \mathcal{O}(\log \delta^{-1})$ iterations of (22) are sufficient for achieving an objective gap of $\delta > 0$ from the stationary solution. Since the maps in (22) require $\mathcal{O}\{n(n + N)p\}$ work to compute, the running time of `sp.ccp` is $\mathcal{O}\{n(n + N)p \log \delta^{-1}\}$. Assuming the batch sample size N does not increase with n or p, this time reduces to $\mathcal{O}(n^2 p \log \delta^{-1})$, which suggests the proposed algorithm can efficiently generate moderately-sized point sets in moderately-high dimensions, but may be computationally burdensome for large point sets. While a similar linear error convergence is difficult to establish for `sp.sccp` due to its stochastic nature (see [4, 20]), its running time is quite similar to `sp.ccp` from simulations.

The separable form of (22) also allows for further computational speed ups using parallel processing. As outlined in Algorithms 1 and 2, the iterative map for each point $\mathbf{x}_i$ can be computed in parallel using separate processing cores. Letting P be the total number of computation cores available, such a parallelization scheme reduces the running time of `sp.ccp` and `sp.sccp` to $\mathcal{O}(\lceil n/P \rceil np \log \delta^{-1})$, thereby allowing for quicker optimization of large point sets. This feature is particularly valuable given the increasing availability of multi-core processors in personal laptops and computing clusters.

4. Simulations. Several simulations are presented here which demonstrate the effectiveness of support points in practice. We first discuss the space-filling property of support points, then comment on its computation time using `sp.sccp`. Finally, we compare the integration performance of support points with MC and a RQMC method called IT-RSS (defined later).

4.1. *Visualization and timing.* For visualization, Figure 2 shows the $n = 128$-point point sets for the i.i.d. $N(0, 1)$ and $Exp(1)$ distributions in $p = 2$ dimensions, with lines outlining density contours (additional visualizations provided in Appendix B of supplementary materials). Support points are plotted on the left, Monte Carlo samples in the middle and inverse Sobol' points on the right. The latter is generated by choosing the Sobol' points on $U[0, 1]^2$ which maximize the minimum interpoint distance over 10,000 random scramblings (see next section for details), then performing an inverse-transform of F on such a point set. From this figure, support points appear to be slightly more visually representative of the underlying distribution F than the inverse Sobol' points, and much more representative than MC. Specifically, the proposed point set is concentrated in regions with high density, but each point is sufficiently spaced out from one another to maximize their representative power. Borrowing a term from design-of-experiments
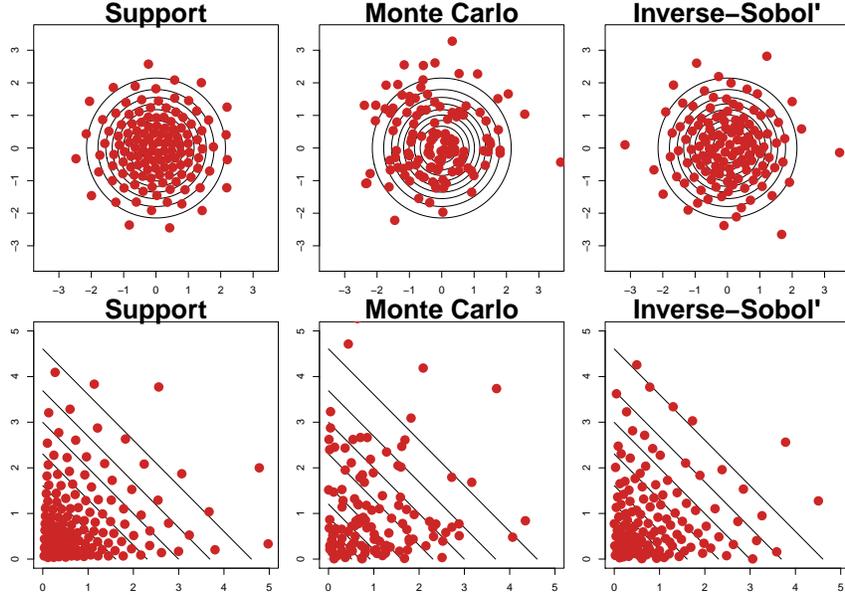
Figure 2: $n = 128$ *support points, MC points and inverse Sobol' points for i.i.d.* $N(0,1)$ *and* $Exp(1)$ *in* $p = 2$ *dimensions. Lines represent density contours.*

literature [55], we call point sets with these two properties to be *space-filling* on F. A key reason for this space-fillingness is the distance-based property of the energy distance: the two terms for $E(F, F_n)$ in (4) force support points to not only mimic the desired distribution F, but also ensure no two points are too close together. This allows for a more appealing visual representation of F, and can provide more robust integration performance.

Regarding computation time, Figure 3 shows the times (in seconds) needed for `sp.sccp` to generate support points for the i.i.d. $Beta(2, 4)$ distribution, first as a function of point set size n with fixed dimension p, then as a function of p with fixed n. The resampling size is fixed at $N = 10,000$ for all choices of n and p. Similar times are reported for other distributions, and are not reported for brevity. All computations are performed on a 12-core Intel Xeon 3.50 Ghz processor. From this figure, two interesting observations can be made. First, for fixed n, these plots show that the empirical running times grow quite linearly in p, whereas for fixed p, these running times exhibit a slow quadratic (but almost linear) growth in n. This provides evidence for the $\mathcal{O}(n^2 p)$ running time asserted in Section 3.3. Second, as a result of this running time, support points can be generated efficiently for moderate-sized point sets in moderately-high dimensions. For $p = 2$, the required times for generating $n = 50 - 10,000$ points range from 3 seconds to 2 minutes; for
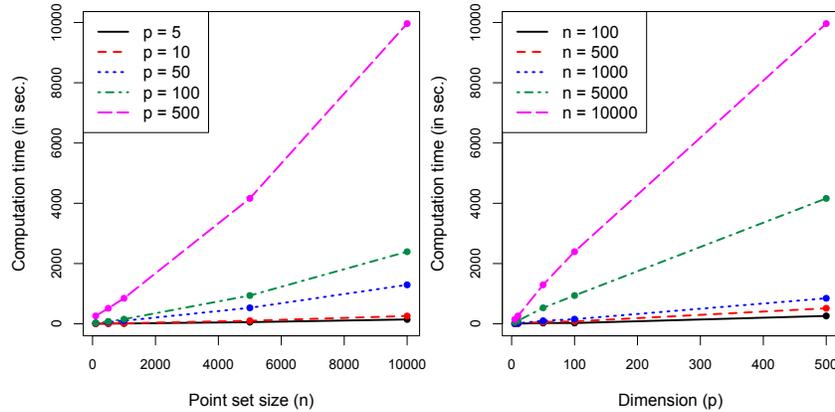
Figure 3: *Computation time (in seconds) of* `sp.sccp` *as a function of point set size (n) and dimension (p) for the i.i.d. Beta(2, 4) distribution.*

$p = 50$, 27 seconds to 20 minutes; and for $p = 500$, 4 minutes to 2.5 hours. While these times are quite fast from an optimization perspective, they are still slower than number-theoretic QMC methods, which can generate, say, $n = 10^6$ points in $p = 10^3$ dimensions in a matter of seconds. The appeal for support points is that, by exploiting the d.c. structure of the energy distance in [62], one obtains for any distribution (locally) minimum energy sampling points which can outperform number-theoretic QMC methods.

4.2. *Numerical integration.* We now investigate the integration performance of support points in comparison with Monte Carlo and an RQMC method called the inverse-transformed randomized Sobol' sequences (IT-RSS). The former is implemented using the Mersenne twister [42], the default pseudo-random number generator in the software R [51]. The latter is obtained by (a) generating a randomized Sobol' sequence using the R package `randtoolbox` [12] (which employs Owen-style scrambling [47] with Sobol' sequences generated in the implementation of [26]), and (b) performing the inverse-transform of F on the resulting point set. As mentioned in Section 2, IT-RSS performs well in the uniform setting $F = U[0, 1]^p$, and provides a good benchmark for comparing support points with existing QMC methods.

The simulation set-up is as follows. Support points are generated using `sp.sccp`, with point set sizes ranging from $n = 50$ to 10,000 and resampling size N fixed at 10,000. Since MC and IT-RSS are randomized methods, we replicate both for 100 trials to provide an estimate of error variability, with replications seeded for reproducibility. Three distributions are considered for F: the i.i.d. $N(0, 1)$, the i.i.d. $Exp(1)$ and the i.i.d. $Beta(2, 4)$ distributions,
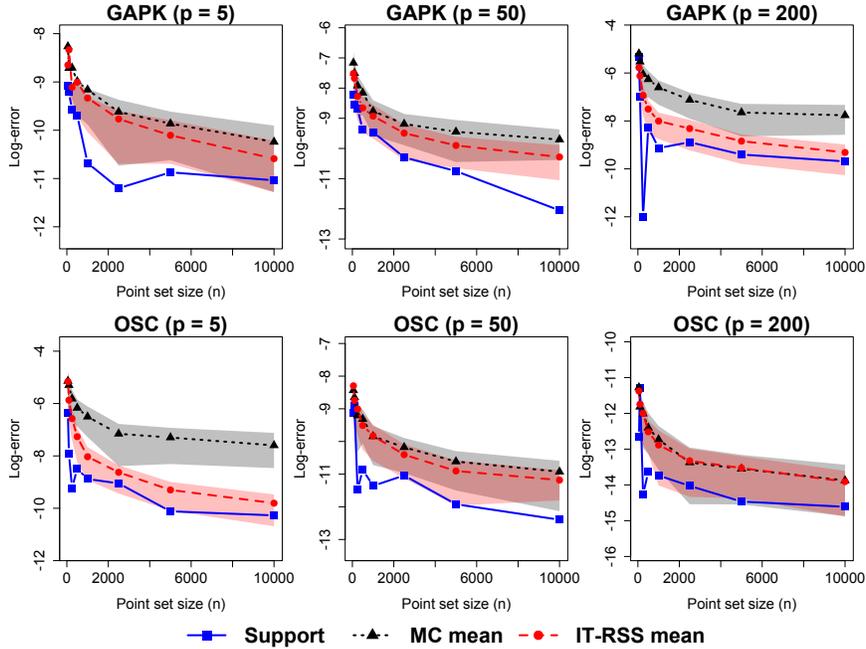
Figure 4: *Log-absolute errors for GAPK under the i.i.d. Exp(1) distribution (top)
and for OSC under the i.i.d. N(0,1) distribution (bottom). Lines denote log average-
errors, and shaded bands mark the 25-th and 75-th quantiles.*

with p ranging from 5 to 500. For the integrand g, two (modified) test
functions are taken from [18]: the Gaussian peak function (GAPK): $g(\mathbf{x}) = \exp\left\{-\sum_{l=1}^{p} \alpha_l^2 (x_l - u_l)^2\right\}$ and the (modified) oscillatory function (OSC):
$g(\mathbf{x}) = \exp\{-\sum_{l=1}^{p} \beta_l x_l^2\} \cos\left(2\pi u_1 + \sum_{l=1}^{p} \beta_l x_l\right)$. Here, $\mathbf{x} = (x_l)_{l=1}^{p}$, u_l is
the marginal mean for the l-th dimension of F, and the scale parameters α_l
and β_l are set as $20/p$ and $5/p$, respectively.

Figure 4 shows the resulting log-absolute errors in p = 5, 50 and 200
dimensions for GAPK under the i.i.d. $Exp(1)$ distribution, and for OSC
under the i.i.d. $N(0,1)$ distribution (results are similar for other settings,
and are omitted for brevity). For MC and IT-RSS, the dotted lines indicate
average error decay, and the shaded bands mark the area between the 25-th
and 75-th error quantiles. Two observations can be made here. First, for all
choices of n, support points enjoy considerably reduced errors compared to
the averages of both MC and IT-RSS, with the proposed method providing
an improvement to the 25-th quantiles of IT-RSS for most settings. Second,
this advantage over MC and IT-RSS persists in both low and moderate
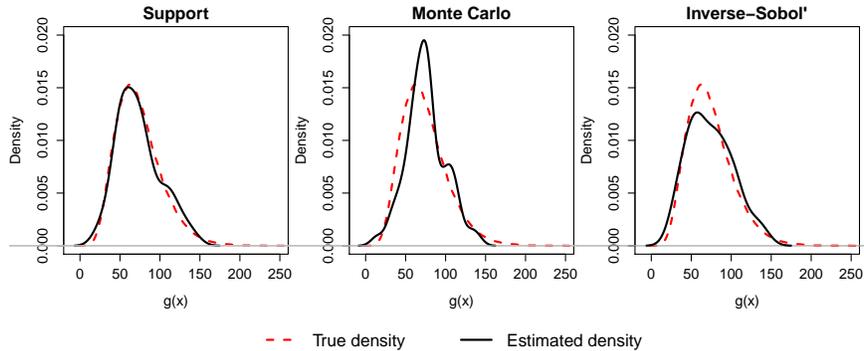dimensions. In view of the relief from dimensionality enjoyed by IT-RSS,

Figure 5: *True and estimated density functions for $g(\mathbf{X})$ using $n = 60$ points.*

this gives some evidence that support points *may* enjoy a similar property as well, a stronger assertion than is provided in Theorem 5 or 6. Exploring the theoretical performance of support points in high dimensions will be an interesting direction for future work.

In summary, for point set sizes as large as $10,000$ points in dimensions as large as 500, simulations show that support points can be efficiently generated and enjoy improved performance over MC and IT-RSS. This opens up a wide range of important applications for support points in both small-data and big-data problems, two of which we describe next.

5. Applications of support points.

5.1. *Uncertainty propagation in expensive simulations.* We first highlight an important small-data application of support points in simulation. With the development of powerful computational tools, computer simulations are becoming the de-facto method for conducting engineering experiments. For such simulations, a key point of interest is *uncertainty propagation*, or how uncertainty in input variables (resulting from, say, manufacturing tolerances) propagate and affect output variability. Mathematically, let $g(\mathbf{x})$ be the observed output at input setting $\mathbf{x}$, and let $\mathbf{X} \sim F$ denote input uncertainties. The distribution $g(\mathbf{X})$ can then be seen as the resulting uncertainty on system output. For engineers, the estimation of $g(\mathbf{X})$ using as few simulation runs as possible is of great importance, because each run can be computationally and monetarily expensive.

To demonstrate the effectiveness of support points for this problem, we use the *borehole physical model* [68], which simulates water flow rate through a borehole. The 8 input variables for this model, along with their corresponding uncertainty distributions (assumed to be mutually independent),

Parameter	Prior	$R_\mu(375)$	$R_\mu(750)$	$R_{\sigma^2}(375)$	$R_{\sigma^2}(750)$
ϕ_{i1}	$\log \phi_{i1} \overset{indep.}{\sim} N(\mu_1, \sigma_1^2)$	2.27	2.75	15.89	6.37
ϕ_{i2}	$\log(\phi_{i2} + 1) \overset{indep.}{\sim} N(\mu_2, \sigma_2^2)$	2.10	3.58	18.01	2.47
ϕ_{i3}	$\log(-\phi_{i3}) \overset{indep.}{\sim} N(\mu_3, \sigma_3^2)$	1.59	2.23	11.90	102.49
σ_C^2	$\sigma_C^2 \sim \text{Inv-Gamma}(0.001, 0.001)$	0.98	2.80	6.15	7.69
$r(1600)$		1.95	3.17	-	-
$r(1625)$	$r(t) = \frac{1}{5}\sum_{i=1}^{5} \frac{\partial}{\partial s}\eta_i(s)\big|_{s=t}$	2.30	3.28	-	-
$r(1650)$		2.51	3.04	-	-
μ_j	$\mu_j \overset{i.i.d.}{\sim} N(0, 100)$	-	-	-	-
σ_j^2	$\sigma_j^2 \overset{i.i.d.}{\sim} \text{Inv-Gamma}(0.01, 0.01)$	-	-	-	-

Table 1: *Prior specification for the tree growth model (left), and the ratios of thinning over support point error for posterior quantities (right). $R_\mu(n)$ and $R_{\sigma^2}(n)$ denote the error ratios for posterior means and variances using n points, respectively.*

are summarized in Appendix C of supplementary materials. To reflect the expensive cost of simulations, we test only small point set sizes ranging from $n = 20$ to $n = 100$ runs. Support points are generated using `sp.sccp` with the same settings as before, with the randomized MC and IT-RSS methods replicated for 100 trials.

Consider now the estimation of the output distribution $g(\mathbf{X})$, which quantifies the uncertainty in water flow rate. Figure 5 compares the estimated density function of $g(\mathbf{X})$ using $n = 60$ points with its true density, where the latter estimated using a large Monte Carlo sample. Visually, support points provide the best density approximation for $g(\mathbf{X})$, capturing well both the peak and tails of the desired output distribution. This suggests support points are not only asymptotically consistent for density estimation, but may also be optimal in some sense. A similar conclusion holds in the estimation of the expected flow rate $\mathbb{E}[g(\mathbf{X})]$ (see Appendix C of supplementary materials).

5.2. *Optimal MCMC reduction.* The second application of support points is as an improved alternative to MCMC thinning for Bayesian computation. Thinning here refers to the discarding of all but every k-th sample for an MCMC sample chain obtained from the posterior distribution. This is performed for several reasons (see [34]): it reduces high autocorrelations in the MCMC chain, saves computer storage space, and reduces processing time for computing derived posterior quantities. However, by carelessly throwing away samples, a glaring fault of thinning is that samples from thinned chains are inherently less accurate than that from the full chain. To this end, the proposed algorithm `sp.ccp` can provide considerable improvements to thinning by optimizing for a point set which best captures the distribution

of the full MCMC chain.

We illustrate this improvement using the orange tree growth model in [11]. The data here consists of trunk circumference measurements $\{Y_i(t_j)\}_{i=1}^{5}{}_{j=1}^{7}$, where $Y_i(t_j)$ denotes the measurement taken on day t_j from tree i. To model these measurements, the growth model $Y_i(t_j) \overset{indep.}{\sim} N(\eta_i(t_j), \sigma_C^2), \eta_i(t_j) = \phi_{i1}/(1 + \phi_{i2} \exp\{\phi_{i3} t_j\})$ was assumed in [11], where ϕ_{i1}, ϕ_{i2} and ϕ_{i3} control the growth behavior of tree i. There are 16 parameters in total, which we denote by the set $\Theta = (\phi_{11}, \phi_{12}, \cdots, \phi_{53}, \sigma^2)$. Since no prior information is available on Θ, vague priors are assigned, with the full specification provided in the left part of Table 1. MCMC sampling is then performed for the posterior distribution using the R package STAN [5], with the chain run for 150,000 iterations and the first 75,000 of these discarded as burn-in. The remaining $N = 75,000$ samples are then thinned at a rate of 200 and 100, giving $n = 375$ and $n = 750$ thinned samples, respectively. Support points are generated using sp.ccp for the same choices of n, using the full MCMC chain as the approximating sample $\{\mathbf{y}_m\}_{m=1}^{N}$. Since posterior variances vary greatly between parameters, we first rescale each parameter in the MCMC chain to unit variance before performing sp.ccp, then scale back the resulting support points after.

These two methods are then compared on how well they estimate two quantities: (a) marginal posterior means and standard deviations of each parameter, and (b) the averaged instantaneous growth rate $r(t)$ (see Table 1) at three future times. True posterior quantities are estimated by running a longer MCMC chain with 600,000 iterations. This comparison is summarized in the right part of Table 1, which reports the ratios of thinning over support point error for each parameter. Keeping in mind that a ratio exceeding 1 indicates lower errors for support points, one can see that sp.ccp provides a sizable improvement over thinning for nearly all posterior quantities. Such a result should not be surprising, because sp.ccp compacts the *full* MCMC chain into a set of optimal representative points, whereas thinning wastes valuable information by discarding a majority of this chain.

6. Conclusion and future work. In this paper, a new method is proposed for compacting a continuous distribution F into a set of representative points called support points, which are defined as the minimizer of the energy distance in [63]. Three theorems are proven here which justify the use of these point sets for integration. First, we showed that support points are indeed representative of the desired distribution, in that these point sets converge in distribution to F. Second, we provided a Koksma-Hlawka-like bound which connects integration error with the energy distance for a large

class of integrands. Lastly, using an existence result, we demonstrated the theoretical error improvement of support points over Monte Carlo. A key appeal of support points is its formulation as a difference-of-convex optimization problem. The two proposed algorithms, `sp.ccp` and `sp.sccp`, exploit this structure to efficiently generate moderate-sized point sets ($n \leq 10,000$) in moderately-high dimensions ($p \leq 500$). Simulations confirm the improved performance of support points to MC and a specific QMC method, and the practical applicability of the proposed point set is illustrated using two real-world applications, one for small-data and the other for big-data. An efficient C++ implementation of `sp.ccp` and `sp.sccp` is made available in the R package `support` [38].

While the current paper establishes some interesting results for support points, there are still many exciting avenues for future research. First, we are interested in exploring a tighter convergence rate for support points which reflects its empirical performance from simulations, particularly for high-dimensional problems. Next, the d.c. formulation of the energy distance can potentially be further exploited for the global optimization of support points. Moreover, by minimizing the *distance-based* energy distance, support points also have an inherent link to the *distance-based* designs used in computer experiments [55, 28, 39], and exploring this connection may reveal interesting insights between the two fields, and open up new approaches for uncertainty quantification in engineering [40] and machine-learning [41] problems. Lastly, motivated by [23] and [28], rep-points in high-dimensions should not only provide a good representation of the full distribution F, but also for *marginal* distributions of F. Such a projective property is enjoyed by most QMC point sets in the literature [9], and new methodology is needed to incorporate this within the support points framework.

References.

[1] Ascher, U. M. and Greif, C. (2011). *A First Course on Numerical Methods*. SIAM.
[2] Bahouri, H., Chemin, J.-Y., and Danchin, R. (2011). *Fourier Analysis and Nonlinear Partial Differential Equations*, volume 343. Springer Science & Business Media.
[3] Borodachov, S. V., Hardin, D. P., and Saff, E. B. (2014). Low complexity methods for discretizing manifolds via Riesz energy minimization. *Foundations of Computational Mathematics*, 14(6):1173–1208.
[4] Bousquet, O. and Bottou, L. (2008). The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168.

[5] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32.

[6] Cox, D. R. (1957). Note on grouping. *Journal of the American Statistical Association*, 52(280):543–547.

[7] Dalenius, T. (1950). The problem of optimum stratification. *Scandinavian Actuarial Journal*, 1950(3-4):203–213.

[8] Di Nezza, E., Palatucci, G., and Valdinoci, E. (2012). Hitchhiker's guide to the fractional Sobolev spaces. *Bulletin des Sciences Mathématiques*, 136(5):521–573.

[9] Dick, J., Kuo, F. Y., and Sloan, I. H. (2013). High-dimensional integration: the quasi-Monte Carlo way. *Acta Numerica*, 22:133–288.

[10] Dick, J. and Pillichshammer, F. (2010). *Digital Nets and Sequences: Discrepancy Theory and Quasi–Monte Carlo Integration*. Cambridge University Press.

[11] Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*. John Wiley & Sons.

[12] Dutang, C. and Savicky, P. (2013). `randtoolbox`: Generating and testing random numbers. *R package*.

[13] Fang, K.-T. (1980). The uniform design: application of number-theoretic methods in experimental design. *Acta Math. Appl. Sinica*, 3(4):363–372.

[14] Fang, K.-T., Lu, X., Tang, Y., and Yin, J. (2004). Constructions of uniform designs by using resolvable packings and coverings. *Discrete Mathematics*, 274(1):25–40.

[15] Fang, K.-T. and Wang, Y. (1994). *Number-Theoretic Methods in Statistics*, volume 51. CRC Press.

[16] Flury, B. A. (1990). Principal points. *Biometrika*, 77(1):33–41.

[17] Gelfand, I. and Shilov, G. (1964). *Generalized Functions, Vol. I: Properties and Operations*. Academic Press, New York.

[18] Genz, A. (1984). Testing multidimensional integration routines. In *Proc. of International Conference on Tools, Methods and Languages for Scientific and Engineering Computation*, pages 81–94. Elsevier North-Holland, Inc.

[19] Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 7(4):473–483.

[20] Ghadimi, S. and Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368.

[21] Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 73(2):123–214.

[22] Graf, S. and Luschgy, H. (2000). *Foundations of Quantization for Probability Distributions*. Springer-Verlag Berlin Heidelberg.

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

[24] Hickernell, F. J. (1999). Goodness-of-fit statistics, discrepancies and robust designs. *Statistics & Probability Letters*, 44(1):73–78.

[25] Hunter, J. K. and Nachtergaele, B. (2001). *Applied Analysis*. World Scientific Publishing.

[26] Joe, S. and Kuo, F. Y. (2003). Remark on algorithm 659: Implementing Sobol's quasirandom sequence generator. *ACM Transactions on Mathematical Software*, 29(1):49–57.

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

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

[29] Kiefer, J. (1961). On large deviations of the empiric df of vector chance variables and a law of the iterated logarithm. *Pacific Journal of Mathematics*, 11(2):649–660.

[30] Kolmogorov, A. (1933). Sulla determinazione empirica delle leggi di probabilita. *Giorn. Ist. Ital. Attuari*, 4:1–11.

[31] Korolyuk, V. and Borovskikh, Y. V. (1989). Convergence rate for degenerate von Mises functionals. *Theory of Probability & Its Applications*, 33(1):125–135.

[32] Kuo, F. Y. and Sloan, I. H. (2005). Lifting the curse of dimensionality. *Notices of the AMS*, 52(11):1320–1328.

[33] Lange, K. (2016). *MM Optimization Algorithms*. SIAM.

[34] Link, W. A. and Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution*, 3(1):112–115.

[35] Lipp, T. and Boyd, S. (2016). Variations and extension of the convex–concave procedure. *Optimization and Engineering*, 17(2):1–25.

[36] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

[37] Mairal, J. (2013). Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems*, pages 2283–2291.

[38] Mak, S. (2017). `support`: *Support Points*. R package version 0.1.0.

[39] Mak, S. and Joseph, V. R. (2017). Minimax and minimax projection designs using clustering. *Journal of Computational and Graphical Statistics*. In press.

[40] Mak, S., Sung, C.-L., Wang, X., Yeh, S.-T., Chang, Y.-H., Joseph, V. R., Yang, V., and Wu, C. F. J. (2017). An efficient surrogate model for emulation and physics extraction of large eddy simulations. *arXiv preprint arXiv:1611.07911*.

[41] Mak, S. and Xie, Y. (2017). Uncertainty quantification and design for noisy matrix completion-a unified framework. *arXiv preprint arXiv:1706.08037*.

[42] Matsumoto, M. and Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, 8(1):3–30.

[43] Nichols, J. A. and Kuo, F. Y. (2014). Fast CBC construction of randomly shifted lattice rules achieving $\mathcal{O}(n^{-1+\delta})$ convergence for unbounded integrands over $\mathbb{R}^s$ in weighted spaces with POD weights. *Journal of Complexity*, 30(4):444–468.

[44] Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM.

[45] Nuyens, D. and Cools, R. (2006). Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces. *Mathematics of Computation*, 75(254):903–920.

[46] Ortega, J. M. and Rheinboldt, W. C. (2000). *Iterative Solution of Nonlinear Equations in Several Variables*. SIAM.

[47] Owen, A. B. (1998). Scrambling Sobol' and Niederreiter–Xing points. *Journal of Complexity*, 14(4):466–489.

[48] Owen, A. B. and Tribble, S. D. (2005). A quasi-Monte Carlo Metropolis algorithm. *Proceedings of the National Academy of Sciences*, 102(25):8844–8849.

[49] Pagès, G., Pham, H., and Printems, J. (2004). Optimal quantization methods and applications to numerical problems in finance. In *S. T. Rachev (ed.), Handbook of Computational and Numerical Methods in Finance*, pages 253–297. Birkhäuser, Boston.

[50] Paley, R. and Zygmund, A. (1930). On some series of functions. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 26, pages 337–357. Cambridge Univ Press.

[51] R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[52] Resnick, S. I. (1999). *A Probability Path*. Springer Science & Business Media.

[53] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472.

[54] Royden, H. L. and Fitzpatrick, P. (2010). *Real Analysis*. Macmillan New York.

[55] Santner, T. J., Williams, B. J., and Notz, W. I. (2013). *The Design and Analysis of Computer Experiments*. Springer Science & Business Media.

[56] Serfling, R. J. (2009). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons.

[57] Shorack, G. R. (2000). *Probability for Statisticians*. Springer Science & Business Media.

[58] Sloan, I. H., Kuo, F. Y., and Joe, S. (2002). Constructing randomly shifted lattice rules in weighted Sobolev spaces. *SIAM Journal on Numerical Analysis*, 40(5):1650–1665.

[59] Sobol', I. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 7(4):784–802.

[60] Su, Y. (2000). Asymptotically optimal representative points of bivariate random vectors. *Statistica Sinica*, 10(2):559–576.

[61] Székely, G. J. (2003). E-statistics: The energy of statistical samples. Technical Report 03-05, Bowling Green State University, Department of Mathematics and Statistics.

[62] Székely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, 5:1–6.

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

[64] Tao, P. D. and An, L. T. H. (1997). Convex analysis approach to DC programming: Theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355.

[65] Tuy, H. (1986). A general deterministic approach to global optimization via dc programming. In *J. B. Hiriart-Urruty (ed.), Fermat Days 1985: Mathematics for Optimization*, pages 137–162. North-Holland, Amsterdam.

[66] Tuy, H. (1995). DC optimization: theory, methods and algorithms. In *Handbook of Global Optimization*, pages 149–216. Springer.

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

[68] Worley, B. A. (1987). Deterministic uncertainty analysis. Technical Report ORNL-6428, Oak Ridge National Laboratories.

[69] Yang, G. (2012). *The energy goodness-of-fit test for univariate stable distributions*. PhD thesis, Bowling Green State University.

[70] Yuille, A. L. and Rangarajan, A. (2003). The concave-convex procedure. *Neural Computation*, 15(4):915–936.

[71] Zador, P. (1982). Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28(2):139–149.