Statistical Papers (2024) 65:1411–1436
https://doi.org/10.1007/s00362-023-01449-6

REGULAR ARTICLE

# Deterministic sampling based on Kullback–Leibler divergence and its applications

**Sumin Wang**[1] · **Fasheng Sun**[2]

Received: 5 November 2022 / Revised: 6 April 2023 / Published online: 27 April 2023


## Abstract

This paper introduces a new way to extract a set of representative points from a continuous distribution, which focuses on a method where the selection of points is essentially deterministic, with an emphasis on achieving accurate approximation when the size of points is small. These points are generated by minimizing the Kullback–Leibler divergence, which is an information-based measure of the disparity between two probability distributions. We refer to these points as Kullback–Leibler points. Based on the link between the total variation and the Kullback–Leibler divergence, we prove that the empirical distribution of Kullback–Leibler points converges to the target distribution. Additionally, we illustrate that Kullback–Leibler points have advantages in simulations when compared with representative points generated by Monte Carlo or other representative points methods. In addition, to prevent the frequent evaluation of complex functions, a sequential version of Kullback–Leibler points is proposed, which adaptively updates the representative points by learning about the complex or unknown functions sequentially. Two potential applications of Kullback-Leibler points in simulation of complex probability densities and optimization of complex response surfaces are discussed and demonstrated with examples.

**Keywords** Bayesian computation · Computer experiments · Gaussian process model · Representative points · Space-filling design

**Mathematics Subject Classification** 62K15, 62K99

✉ Fasheng Sun
sunfs359@nenu.edu.cn

Sumin Wang
wangsm088@nankai.edu.cn

[1] Center for Combinatorics, LPMC & KLMDASR, Nankai University, Weijin Road No. 94, Tianjin 300071, China

[2] School of Mathematics and Statistics & KLAS, Northeast Normal University, Renmin Street No. 5268, Changchun 130024, Jilin Province, China

## 1 Introduction

Computational statistics and machine learning face an important issue, approximating a complex distribution $F$ with an empirical distribution supported on a set of representative points $\{\mathbf{x}_i\}_{i=1}^n$. Markov chain Monte Carlo (Brooks et al. 2011) methods are extensively used for this task. However, these methods suffer from 'clustering' and require a large number of samples to approximate a complex distribution, which can be costly when the distribution is expensive to evaluate. We illustrate this with the banana-shaped density function given by (Haario et al. 1999):

$$f(\mathbf{x}) \propto \exp\left\{-\frac{1}{2}\frac{x_1^2}{100} - \frac{1}{2}\left(x_2 - 0.03x_1^2 - 3\right)^2\right\}. \tag{1}$$

The left side of Fig. 1 shows 1000 Monte Carlo (MC) samples obtained by the R package `mcmc`. Cleary, although most of the samples are in high-density regions, many of them are repeated or very close. These repeated or very close samples provide little additional information when conducting expensive computer simulations, where the same input will lead to the same output, and then can be viewed as a waste of evaluations. Therefore, if we can spread the samples as far apart as possible, then more information about the distribution can be obtained with less effort. We can overcome this problem by reducing the repeated samples and those that are close to each other. This is the idea behind deterministic sampling methods such as Quasi-Monte Carlo (QMC) methods (Sobol' 1967). However, although QMC methods make the samples achieve a well-spaced configuration, very few fall into high-density regions, and thus, most of them are wasted. This is a major drawback of QMC methods since they were originally developed for generating samples from uniform distributions. One recommended strategy in the QMC literature is to map samples from a uniform distribution to a nonuniform distribution $F$ using the inverse of the distribution function. However, this can be performed only when the variables are independent, which is rarely observed for most types of distributions.

This paper focuses on exploring a good deterministic sampling method when $n$ is small, which is an important small-data application in expensive computer simulations (Worley 1987) and Bayesian calibration (Kennedy and O'Hagan 2001). A good deterministic sampling method should satisfy the following: (i) place more samples in high-density regions; (ii) ensure that the samples are spread out well; and (iii) avoid a large number of evaluations on the distribution. Such a 'space-filling' property can allow for improved integration performance over MC and QMC methods. We outline two classes of deterministic sampling methods.

The first class of good deterministic sampling methods is minimum energy design (MED, Joseph et al. 2015, 2019). The key ideas are the visualization of the sample points as charged particles inside a box with the same electrical properties and the minimization of the total potential energy of these particles when the charged particles reach equilibrium. An MED $\{\mathbf{x}_i\}_{i=1}^n$ of the density function $f$ can be obtained by minimizing the following potential energy function:

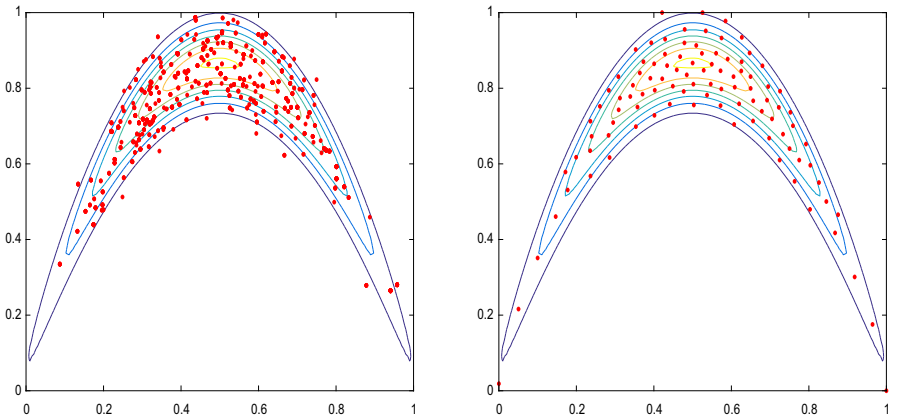

**Fig. 1** 1000 MC samples (left) vs. 100 KL points (right)

$$\sum_{i \neq j} \frac{f(\mathbf{x}_i)^{-k/2d} f(\mathbf{x}_j)^{-k/2d}}{\| \mathbf{x}_i - \mathbf{x}_j \|_2^k}, \quad k \in [1, \infty).$$

When $k \to \infty$, the MED can be constructed by minimizing

$$\max_{i \neq j} \frac{1}{f(\mathbf{x}_i)^{1/2d} f(\mathbf{x}_j)^{1/2d} \| \mathbf{x}_i - \mathbf{x}_j \|_2},$$

where $\mathbf{x}$ is a $d$-dimensional vextor in $R^d$, and $\|\cdot\|_2$ denotes the Euclidean norm. Constructions on an MED require tedious global optimizations and a number of evaluations of the distribution, which can be computationally expensive. Thus, Joseph et al. (2019) provided a fast algorithm for generating an MED, but the newly generated MED may perform worse in the space-filling property (see Fig. 2), the explanation can be seen in Sect. 4.1.

The second class of good deterministic sampling methods is support points (Mak and Joseph 2017, 2018), which can provide much better representative points than the MED. Support points aim to generate representative points by minimizing the energy distance, a statistical potential measure for testing goodness-of-fit. The support points $\{\mathbf{x}_i\}_{i=1}^n$ of $F$ can be obtained by minimizing

$$\frac{2}{n} \sum_{i=1}^{n} E \| \mathbf{x}_i - \mathbf{Y} \|_2 - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \| \mathbf{x}_i - \mathbf{x}_j \|_2 - E \| \mathbf{Y} - \mathbf{Y}' \|_2,$$

where $\mathbf{Y}, \mathbf{Y}' \overset{i.i.d.}{\sim} F$. However, the algorithm for generating support points needs to obtain a large number of MC samples first, and then compact these samples into a set of representative points. Hence, this algorithm will be computationally expensive when $F$ is complex and expensive to sample.

In this paper, we propose a new deterministic sampling method based on the Kullback–Leibler divergence to compact a continuous probability distribution $F$ into a set of representative points, hereafter referred to as *Kullback–Leibler (KL) points*. The Kullback–Leibler divergence was first introduced in Kullback and Leibler (1951); it measures the disparity between two probability distributions. This divergence is widely used in goodness-of-fit and is more computationally efficient than the classical Kolmogorov–Smirnov statistic (Dudewicz and Van 1981). Jourdan and Franco (2010) constructed a space-filling design by minimizing the Kullback–Leibler divergence between the empirical distribution of the design points and the uniform distribution. Inspired by this, we extend this method to a general probability distribution. The key idea is to optimize the divergence between the empirical distribution of the sample points and the goal distribution, and the generated points are concentrated in regions of high density and are separate from each other. The right side of Fig. 1 shows 100 KL points from a banana-shaped density function, and it is apparent that the 100 KL points seem to be no less informative than the 1000 samples obtained from the MC method.

The representative points of the mixed normal distribution generated by the four methods are shown in Fig. 2, the Monte Carlo points appear to be the worst, and the number of KL points escaping from the density contour is less than the number of MED and support points. To prevent the frequent evaluation of complex functions, this paper also proposes a sequential version of KL points and illustrates two important potential applications. One application is to obtain the representative points from complex distributions. Another important application is exploring complex surfaces, such as the objective is not only to find a global optimum, but also to find several good points that can serve as alternatives to the global optimum. The traditional method of exploring a complex surface mostly relies on a space-filling design (Fang et al. 2006; Lin and Tang 2015; Santner et al. 2019; Shi and Tang 2020); however, these design points may be placed in zero-yield regions that are not useful because they cannot provide any information about the surface. The MED can also be used for sampling from a complex distribution and exploring a complex surface. In later simulations, we demonstrate that the KL points perform better than the MED in two examples (see Figs. 4, 5, 6). Because the energy distance criterion used for generating the support points requires a normalized distribution, the support points fail to explore complex distributions and surfaces.

This article is organized as follows. In Sect. 2, we define KL points using Kullback–Leibler divergence and present several important theoretical properties of KL points. In Sect. 3, we propose a greedy algorithm for efficiently generating KL points. In Sect. 4, we outline several simulations to compare the space-filling property and integration performance of KL points with those of Monte Carlo points, MED points, and support points. In Sect. 5, to prevent the frequent evaluation of complex functions, we develop a sequential version of KL points and propose a generating algorithm. Two potential applications, one that simulates complex probability densities and another related to the optimization of nonnegative complex black-box functions, are discussed in this section. In Sect. 6, we present some concluding remarks and directions for future research. All the proofs of the theorems are presented in the Appendix.

## 2 Kullback–Leibler points

In this section, we first introduce the definition and some properties of Kullback–Leibler divergence and then define KL points. To this end, we address some theoretical results to demonstrate that KL points are appropriate for representing a target distribution.

**Definition 1** (Cover and Thomas 2006). Suppose that $f$ and $g$ are two continuous probability density functions supported by a compact set $\mathcal{X}$, where $\mathcal{X} \in R^d$. Then, the Kullback–Leibler divergence between them is defined as

$$D(g\|f) = \int_{\mathcal{X}} g(\mathbf{x}) \log \frac{g(\mathbf{x})}{f(\mathbf{x})} d\mathbf{x}. \tag{2}$$

**Proposition 1** (Theorem 8.6.1 of Cover and Thomas 2006) $D(g\|f) \geq 0$, *with* $D(g\|f) = 0$ *if and only if* $g = f$ *almost everywhere.*

To access deterministic samples from a given target density function, we use the Kullback–Leibler divergence between the kernel density estimator on the sample points and the target density to measure the representation of the sample points. We refer to $f_n$ as the kernel density estimator on $\{\mathbf{x}_i\}_{i=1}^n \subseteq \mathcal{X}$; then, the Kullback–Leibler divergence between $f$ and $f_n$ becomes:

$$D(f_n\|f) = \int_{\mathcal{X}} f_n(\mathbf{x}) \log \frac{f_n(\mathbf{x})}{f(\mathbf{x})} d\mathbf{x}. \tag{3}$$

Here, $f_n$ is defined as

$$f_n(\mathbf{x}) = \frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right), \tag{4}$$

where the kernel function $K$ and bandwidth $h_n$ satisfy the following:

(K1) $K$ is a continuous and compactly supported probability density function on $\mathcal{X}$;
(K2) $K(-\mathbf{t}) = K(\mathbf{t}), \forall \mathbf{t} \in \mathcal{X}$;
(K3) $\int_{\mathcal{X}} \mathbf{t}\mathbf{t}' K(\mathbf{t}) d\mathbf{t} = \mu_2(\mathbf{K}) I$, where $\mu_2(\mathbf{K}) = \int_{\mathcal{X}} t_i^2 K(\mathbf{t}) d\mathbf{t} < \infty$;
(K4) $\|\mathbf{K}\|_2^2 = \int_{\mathcal{X}} K^2(\mathbf{t}) d\mathbf{t} < \infty$;
(K5) $h_n \to 0, nh_n^d \to \infty$ as $n \to \infty$.

Under the conditions (K1)–(K5), $0 \leq D(f_n\|f) < \infty$, and it can be viewed as a metric. Hence, we can define the KL points as the point set that minimizes $D(f_n\|f)$.

**Definition 2** (*KL points*) Suppose that the support set of $f$ is a compact set $\mathcal{X}$, where $\mathcal{X} \subseteq R^d$. For a fixed point set size $n \in \mathbf{N}$, the KL points of a continuous probability density function $f$ are defined as:

$$\{\boldsymbol{\xi}_i\}_{i=1}^n = \arg\min_{\mathbf{x}_1,\ldots,\mathbf{x}_n} D(f_n\|f) = \arg\min_{\mathbf{x}_1,\ldots,\mathbf{x}_n} \int_{\mathcal{X}} f_n(\mathbf{x}) \log \frac{f_n(\mathbf{x})}{f(\mathbf{x})} d\mathbf{x}, \tag{5}$$

where $f_n$ is defined as in (4).

Let $F$ be the cumulative distribution function of the density $f$. Now, we present the theoretical justification for using the KL points defined in (5) as the representative points for the continuous distribution $F$.

Define $f_n^{KL}$ as the kernel density estimator on KL points $\{\boldsymbol{\xi}_i\}_{i=1}^n$. We first illustrate the total variation and the Kullback–Leibler divergence between $f_n^{KL}$ and the target density $f$ converge to 0 as in Theorem 1. This result is important to proof the theoretical justification for using the KL points as the representative points for the continuous distribution $F$, which are described in Theorems 2 and 3.

**Theorem 1** *Suppose that density $f$ satisfies the following:*

(A1) *The support set of $f$ is a compact set $\mathcal{X}$, $\mathcal{X} \subseteq R^d$;*
(A2) *and $f$ is Lipschitz continuous; i.e.,*

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{X},$$

*where $L$ is a positive number.*

*Suppose that the kernel $K$ and bandwidth $h_n$ satisfying (K1)–(K5). Then, we have*

$$\lim_{n\to\infty} D\left(f_n^{KL} \| f\right) = 0 \tag{6}$$

*and*

$$\lim_{n\to\infty} \int_{\mathcal{X}} \left| f_n^{KL}(\mathbf{x}) - f(\mathbf{x}) \right| d\mathbf{x} = 0. \tag{7}$$

Now, we introduce the first theoretical justification for using the KL points as the representative points. Let $F_n$ be the absolutely continuous empirical distribution corresponding to $f_n^{KL}$. According to Theorem 1, the next theorem states that $F_n$ converges to $F$. Moreover, this result is an important bridge to prove that the standard empirical distribution function of the KL points converges to $F$.

**Theorem 2** *Suppose probability density function $f$ satisfies conditions (A1)–(A2), kernel $K$ and bandwidth $h_n$ satisfy (K1)–(K5). Let $F_n(A) = \int_A f_n^{KL}(\mathbf{x})d\mathbf{x}$ and $A \in \mathcal{B}$, where $\mathcal{B}$ is the Borel $\sigma$-algebra of $\mathcal{X}$; then,*

$$\lim_{n\to\infty} \sup_{A\in\mathcal{B}} |F_n(A) - F(A)| = 0.$$

*In addition, if $A = (-\infty, \mathbf{x}] \in \mathcal{B}$, we have*

$$\lim_{n\to\infty} \sup_{\mathbf{x}} |F_n(\mathbf{x}) - F(\mathbf{x})| = 0.$$

Next, we introduce the second theoretical justification for using the KL points as the representative points. Let $F_n^{KL}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n I(\boldsymbol{\xi}_i \leq \mathbf{x})$ denote the standard empirical distribution of KL points, where $\boldsymbol{\xi}_i \leq \mathbf{x}$ means that each component of vector $\boldsymbol{\xi}_i$ is less than or equal to the corresponding component of vector $\mathbf{x}$. According to Theorems 1 and 2 we can show that $F_n^{KL}$ converges to $F$, which is presented in Theorem 3.

**Theorem 3** *Suppose probability density function $f$ satisfies conditions (A1)–(A2), kernel $K$ and bandwidth $h_n$ satisfy (K1)–(K5). Then, we have*

$$\lim_{n\to\infty} \sup_{\mathbf{x}} |F_n^{KL}(\mathbf{x}) - F(\mathbf{x})| = 0.$$

Theorem 2 and 3 convincingly demonstrate that KL points are indeed representative of the target distribution $F$ as the number of points $n$ becomes large.

## 3 Generating Kullback–Leibler points

Because (5) is an integral, so generating KL points by (5) is very difficult or infeasible in practice. According to the Pinsker's inequalities (Tsybakov 2009), we have

$$D(f_n \| f) \le \int_{\mathcal{X}} f_n(\mathbf{x}) \left| \log \frac{f_n(\mathbf{x})}{f(\mathbf{x})} \right| d\mathbf{x} \le D(f_n \| f) + \sqrt{D(f_n \| f)/2}.$$

Hence, choosing the points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ which minimize $\int_{\mathcal{X}} f_n(\mathbf{x}) \left| \log \frac{f_n(\mathbf{x})}{f(\mathbf{x})} \right| d\mathbf{x}$ will lead to minimized $D(f_n \| f)$. If we let $q(\mathbf{z}; \{\mathbf{x}_i\}_{i=1}^n) := f_n(\mathbf{z}) \left| \log \frac{f_n(\mathbf{z})}{f(\mathbf{z})} \right|$, then $\int_{\mathcal{X}} f_n(\mathbf{x}) \left| \log \frac{f_n(\mathbf{x})}{f(\mathbf{x})} \right| d\mathbf{x}$ can be considered as the expectation of $q(\mathbf{z}; \{\mathbf{x}_i\}_{i=1}^n)$ under the uniform distribution on $\mathcal{X}$. Thus, we can sample $N$ design points $\mathcal{Z} = \{\mathbf{z}_i\}_{i=1}^N$ in $\mathcal{X}$, and use

$$\begin{aligned} \hat{D}(f_n \| f) &= \frac{1}{N} \sum_{j=1}^{N} q\left(\mathbf{z}_j; \{\mathbf{x}_i\}_{i=1}^n\right) \\ &= \frac{1}{N} \sum_{j=1}^{N} \frac{1}{n h_n^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{z}_j - \mathbf{x}_i}{h_n}\right) \left| \log \frac{\frac{1}{n h_n^d} \sum_{i=1}^n K\left(\frac{\mathbf{z}_j - \mathbf{x}_i}{h_n}\right)}{f(\mathbf{z}_j)} \right| \end{aligned}$$

to estimate $\int_{\mathcal{X}} f_n(\mathbf{x}) \left| \log \frac{f_n(\mathbf{x})}{f(\mathbf{x})} \right| d\mathbf{x}$. We choose a large maximin Latin hypercube samples $\{\mathbf{z}_i\}_{i=1}^N$ using the R package `LHD` (Morris and Mitchell 1995) as $\mathcal{Z}$, which is a space-filling design and can achieve variance reduction than the Monte Carlo method (LHS, Mckay et al. 1979). Obviously, $\hat{D}(f_n \| f)$ is nonnegative. Due to the large LHS $\mathcal{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ is a space-filling design which can span the support set $\mathcal{X}$, so we can obtain the KL points by optimize $\hat{D}(f_n \| f)$ on $\mathcal{Z}$, i.e.

$$\begin{aligned} &\arg \min_{\mathbf{x}_1, \ldots, \mathbf{x}_n \subseteq \mathcal{Z}} \hat{D}(f_n \| f) \\ &= \arg \min_{\mathbf{x}_1, \ldots, \mathbf{x}_n \subseteq \mathcal{Z}} \frac{1}{N} \sum_{j=1}^{N} \frac{1}{n h_n^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{z}_j - \mathbf{x}_i}{h_n}\right) \log \left| \frac{\frac{1}{n h_n^d} \sum_{i=1}^n K\left(\frac{\mathbf{z}_j - \mathbf{x}_i}{h_n}\right)}{f(\mathbf{z}_j)} \right| \end{aligned}$$

$$= \arg \min_{\mathbf{x}_1,\ldots,\mathbf{x}_n \subseteq \mathcal{Z}} \frac{1}{N} \left\{ \sum_{j=1}^{n} \frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}_j - \mathbf{x}_i}{h_n}\right) \log \left| \frac{\frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}_j - \mathbf{x}_i}{h_n}\right)}{f(\mathbf{x}_j)} \right| \right.$$

$$\left. + \sum_{\mathbf{z}_j \in \mathcal{Z} \setminus \{\mathbf{x}_i\}_{i=1}^{n-1}} \frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{z}_j - \mathbf{x}_i}{h_n}\right) \log \left| \frac{\frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{z}_j - \mathbf{x}_i}{h_n}\right)}{f(\mathbf{z}_j)} \right| \right\}, \tag{8}$$

where the kernel function $K$ is a truncated multivariate exponential power distribution, i.e.

$$K(\mathbf{t}) \propto \exp(-\|\mathbf{t}\|_2) I(\mathbf{t} \in \mathcal{X}).$$

We have tried many types of kernel functions include the multivariate Gaussian kernel, but we observe that the multivariate exponential power kernel performs much better than other types of kernels, which is also used in Wu and Ghosal (2008). The bandwidth $h_n$ is chosen by the minimized approximate mean integrated squared error criterion (Härdle et al. 2004); that is,

$$h_n = \left( \frac{d\|\mathbf{K}\|_2^2}{n\mu_2^2(\mathbf{K}) \int_{\mathcal{X}} [tr(H_f(\mathbf{x}))]^2 d\mathbf{x}} \right)^{1/(d+4)} = O(n^{-1/(d+4)}),$$

where $H_f(\mathbf{x})$ is the Hessian matrix of second partial derivatives of $f(\mathbf{x})$. Based on condition (K5), we recommend the empirical bandwidth $h_n = d^{-1/(d+4)} d^{-1} \sum_{i=1}^{d} \sigma_i^2 n^{-1/(d+4)}$ in practice, where $\sigma_i^2$ is the true marginal variance for the $i$-th dimension of $F$. From (8), we can see that optimizing $\hat{D}(f_n \| f)$ may place more KL points in high-density regions and ensure them spread out well.

Finding the KL points is a computationally difficult problem. In theory, we should adopt the optimal design algorithms such as the exchange algorithm to do a global optimization directly on the $n$ points by minimizing the criterion (8). But it is difficult in practice and time-consuming. Consequently, in this section, we present a one-point-at-a-time greedy algorithm to generate KL points. Suppose we have already generated $m-1$ points by (8); then, the $m$-th point is generated by

$$\begin{aligned}
\mathbf{x}_m &= \arg \min_{\mathbf{x} \in \mathcal{X}} \hat{D}(f_m \| f) \\
&= \arg \min_{\mathbf{x} \in \mathcal{Z} \setminus \{\mathbf{x}_i\}_{i=1}^{m-1}} \frac{1}{N} \sum_{j=1}^{N} \frac{1}{mh_m^d} \left( \sum_{i=1}^{m-1} K\left(\frac{\mathbf{x}_i - \mathbf{z}_j}{h_m}\right) + K\left(\frac{\mathbf{x} - \mathbf{z}_j}{h_m}\right) \right) \\
&\quad \times \log \left| \frac{\sum_{i=1}^{m-1} K\left(\frac{\mathbf{x}_i - \mathbf{z}_j}{h_m}\right) + K\left(\frac{\mathbf{x} - \mathbf{z}_j}{h_m}\right)}{mh_m^d f(\mathbf{z}_j)} \right| \\
&= \arg \min_{\mathbf{x} \in \mathcal{Z} \setminus \{\mathbf{x}_i\}_{i=1}^{m-1}} M(\mathbf{x} | \{\mathbf{x}_i\}_{i=1}^{m-1}; \{\mathbf{z}_j\}_{j=1}^{N}),
\end{aligned}$$

where, $M(\mathbf{x}|\{\mathbf{x}_i\}_{i=1}^{m-1};\{\mathbf{z}_j\}_{j=1}^{N}) = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{mh_m^d}\left(\sum_{i=1}^{m-1}K\left(\frac{\mathbf{x}_i-\mathbf{z}_j}{h_m}\right)+K\left(\frac{\mathbf{x}-\mathbf{z}_j}{h_m}\right)\right)\times \log\left|\frac{\sum_{i=1}^{m-1}K\left(\frac{\mathbf{x}_i-\mathbf{z}_j}{h_m}\right)+K\left(\frac{\mathbf{x}-\mathbf{z}_j}{h_m}\right)}{mh_m^d f(\mathbf{z}_j)}\right|$.

We choose the first point by

$$\mathbf{x}_1 = \arg\max_{\mathbf{x}\in\mathcal{Z}} f(\mathbf{x}).$$

We traverse $\mathcal{Z}$ to generate the $m$-th KL point $\mathbf{x}_m$. Algorithm 1 below outlines the detailed steps for generating the $n$ KL points of $f$.

**Algorithm 1** Generating $n$ KL points via the greedy algorithm

**Require:** $n, \{\mathbf{z}_j\}_{j=1}^{N}$

**Ensure:** $\mathbf{x}_1, \ldots, \mathbf{x}_n$

1: **if** m=1 **then**
2: $x_1 = \arg\max_{\mathbf{x}\in\mathcal{Z}} f(\mathbf{x})$
3: **else**
4: **for** $m = 2, \ldots, n$ **do**
5: $\mathbf{x}_m = \arg\min_{\mathbf{x}\in\mathcal{Z}\setminus\{\mathbf{x}_i\}_{i=1}^{m-1}} M(\mathbf{x}|\{\mathbf{x}_i\}_{i=1}^{m-1};\{\mathbf{z}_j\}_{j=1}^{N})$
6: **end for**
7: **end if**

## 4 Simulation study

The KL points can be viewed as the optimal sampling points of $F$ (in the sense of minimum Kullback–Leibler divergence) for any desired sample size $n$. These points are concentrated in regions with high densities and are sufficiently spread out to maximize the representative power. Thus, KL points perform well as representative points. This property allows improved integration performance, and these points tend to perform well when used in other applications. In this section, we provide several simulations to illustrate the distinct advantages of KL points compared with MC points, MED points and support points. The MC points are generated by the R package `mcmc`, and the MED points, support points are generated by the R package `mined` and `support`, respectively.

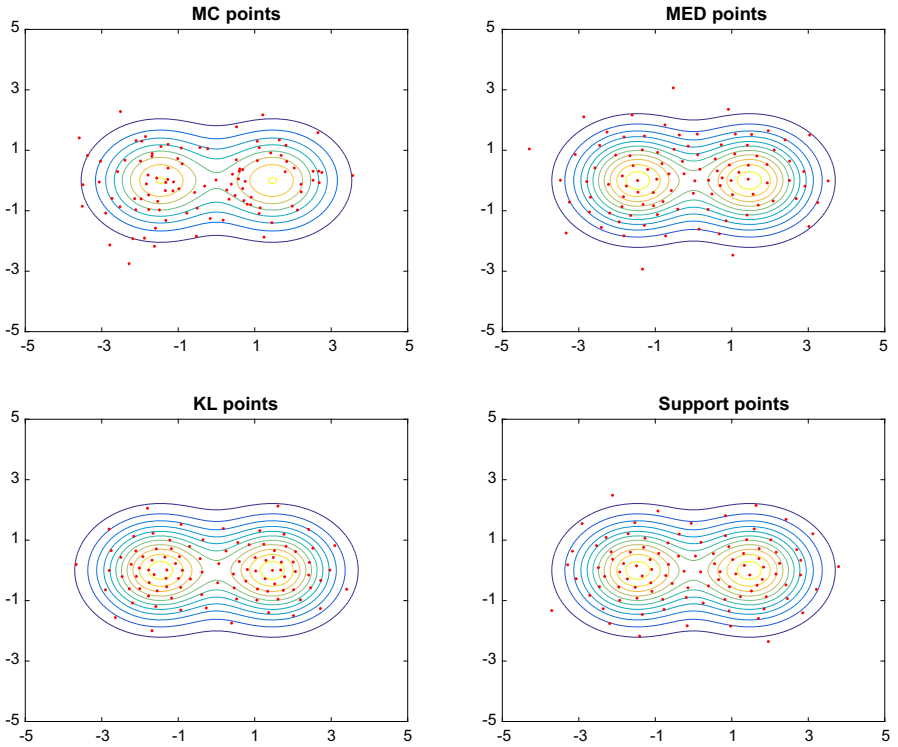


**Fig. 2** Space-filling for a mixed normal distribution with $n = 100$, solid lines represent density contours

### 4.1 Space-filling property

For visualization, Fig. 2 shows the $n = 100$ points for the mixed normal distribution

$$P = \frac{1}{2}N(\mu_1, \Sigma_1) + \frac{1}{2}N(\mu_2, \Sigma_2)$$

defined on $[-5, 5]^2$, where $\mu_1 = (-1.5, 0)'$, $\mu_2 = (1.5, 0)'$, and $\Sigma_1 = \Sigma_2 = I$. From this figure, the KL points appear to be slightly more representative than the MED points and support points, and much more representative than the MC points. For MED points, the worse space-filling property may be caused by the criteria for generating MED points, I think the energy function may not be a good criterion to measure the performance of the representative points for the target density $f$. For support points, the worse space-filling property may be caused by the algorithm which needs to obtain a large number of MC samples first, and then compress the large MC samples into a set of representative points. Hence, the space-filling property of support points rely on the large MC samples, and the worse space-filling property of large MC samples may be result in worse space-filling property of the support points.

### 4.2 Numerical integration

We now investigate the integration performance of KL points in comparison with MC points, MED points, and support points. The absolute error is used to measure the precision of the integral, i.e. $\left| \int_{\mathcal{X}} g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} - n^{-1} \sum_{i=1}^{n} g(\mathbf{x}_i) \right|$, where $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ are the representative points of $f$, which are obtained by these four methods. The simulation setup is as follows. We generate MC points, MED points, support points, and KL points with point set sizes ranging from $n = 50$ to 300. Since the MC points are randomized, we repeat it 1000 times to evaluate the error by its average. We use the distributions and integrand functions based on the following cases.

(M1) a two-dimensional function $g(\mathbf{x}) = \exp(-4(x_1^2+x_2^2))+6\sin(\pi(x_1+x_2))/(2-\sin(\pi(x_1+x_2))$, where the distribution of $\mathbf{x} = (x_1, x_2)$ is a truncated multivariate Gaussian $N(0, \sigma^2 R)$, whose support is $[-5, 5]^2$, $\sigma = 1$, $R_{ij} = 0.5^{|i-j|}$ for $i, j = 1, 2$;

(M2) a five-dimensional function $g(\mathbf{x}) = \exp(-x_1^2 + x_2 + x_3) + \sin(x_4^2 + x_5)$, where the distribution of $\mathbf{x} = (x_1, \ldots, x_5)$ is a truncated multivariate Gaussian $N(0, \sigma^2 R)$, whose support is $[-5, 5]^5$, $\sigma = 1$, $R_{ij} = 0.5^{|i-j|}$ for $i, j = 1, 2$;

(M3) a four-dimensional function $g(\mathbf{x}) = \exp(-\sum_{l=1}^{4} \alpha_l (x_l - \mu_l)) - \sin(x_3 + x_4)$, where the distribution of $\mathbf{x} = (x_1, \ldots, x_4)$ is a truncated multivariate Gaussian $\sim N(0, I)$, whose support is $[-5, 5]^4$, $\alpha_l = 1/45$, $\mu_l$ is the marginal means of $N(0, I)$;

(M4) a six-dimensional function $g(\mathbf{x}) = \exp(-\sum_{l=1}^{6} \alpha_l (x_l - \mu_l))$,where the distribution of $\mathbf{x} = (x_1, \ldots, x_4)$ is a truncated multivariate Gaussian $\sim N(0, I)$, whose support is $[-5, 5]^6$, $\alpha_l = 1/45$, $\mu_l$ is the marginal means of $N(0, I)$.

Figure 3 shows the absolute errors of numerical integration for the cases (M1)–(M4), where the true values for (M1)–(M3) are estimated by a large number of MC points, and the true values for (M4) is 0.9362. From these figures, we observe that KL points enjoy considerably reduced errors compared to MC points and MED points, and even have smaller errors than support points in the case (M1)–(M3). Due to the algorithm for generating support points need to obtain a large number of MC samples first. Hence, it will be computationally expensive when $F$ is expensive to sample. Consequently, KL points may be a good alternative sampling method for expensive distributions.

## 5 Sequential Kullback–Leibler points for complex functions

The direct application of Algorithm 1 for generating KL points cannot be performed easily when $f$ is complex and expensive to compute, because it requires to evaluate $N$ times $f(\mathbf{z})$, i.e. $f(\mathbf{z}_1), \ldots, f(\mathbf{z}_N)$, where $N \gg n$ is usually large. This section proposes an approximate version of KL points to reduce the frequent evaluation of $f$. By analogy with the method of sequential minimum energy design (SMED, Joseph et al. 2015), we replace $f$ with an easy-to-evaluate approximation $\hat{f}$. Here, a sequential strategy is used to implement KL points in such situations. The key idea is to learn about $f$ sequentially and implement the KL points accordingly. For this purpose, we

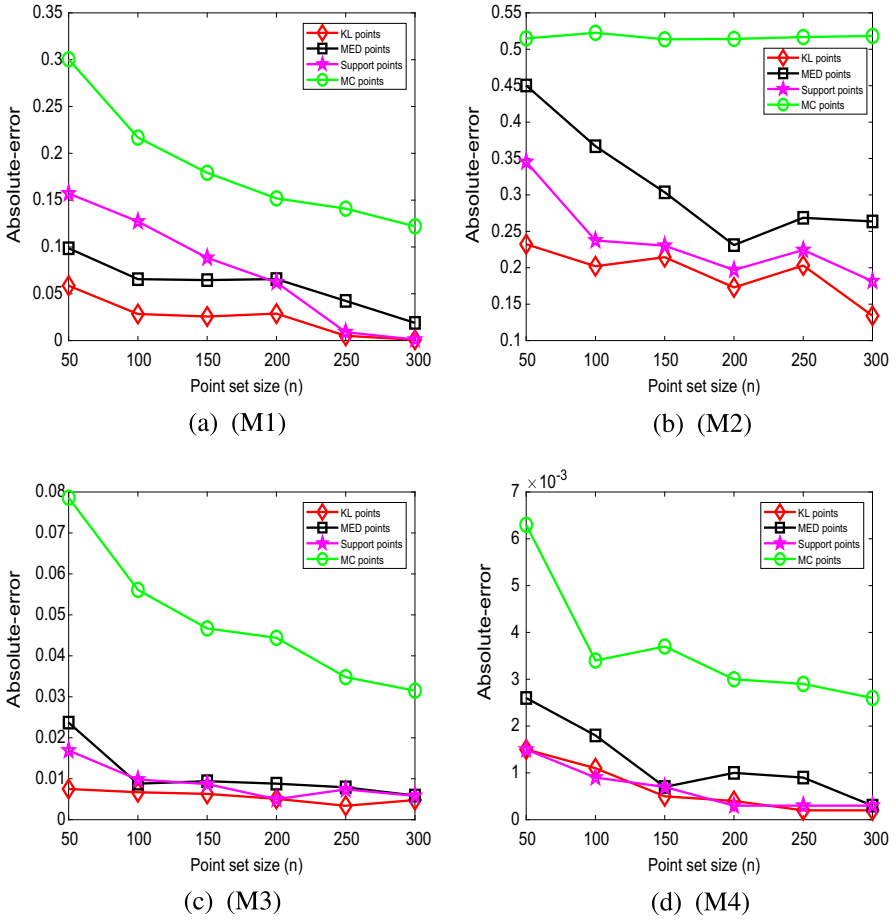


**Fig. 3** The absolute errors for the cases (M1)–(M4)

still adopt the one-point-at-a-time greedy algorithm described in the previous section. To obtain a good estimator $\hat{f}$, we first generate a maximin Latin hypercube design $D_{n_0} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_0}\}$ using the R package `LHD` (Morris and Mitchell 1995) as the initial point set. Then, the corresponding outputs $\{y_i\}_{i=1}^{n_0}$ are obtained by $y_i = f(\mathbf{x}_i)$. Finally, we can use statistical methods such as kriging to estimate $f$, denoting the estimator as $\hat{f}^{(n_0)}$. Based on the initial set $D_{n_0}$, we can generate the next $n - n_0$ KL points sequentially. At each $m = n_0 + 1, \ldots, n$, the estimator $\hat{f}^{(m-1)}$ can be updated

via $\{\mathbf{x}_j, y_j\}_{j=1}^{m-1}$. The $m$-th point $\mathbf{x}_m$ can be generated by traversing on $\mathcal{Z} \setminus \{\mathbf{x}_i\}_{i=1}^{m-1}$

$$\begin{aligned}\mathbf{x}_m &= \arg \min_{\mathbf{x}\in\mathcal{Z}\setminus\{\mathbf{x}_i\}_{i=1}^{m-1}} \hat{D}\left(f_m \| \hat{f}^{(m-1)}\right) \\ &= \arg \min_{\mathbf{x}\in\mathcal{Z}\setminus\{\mathbf{x}_i\}_{i=1}^{m-1}} \frac{1}{N}\sum_{j=1}^{N} \frac{1}{mh_m^d}\left(\sum_{i=1}^{m-1} K\left(\frac{\mathbf{x}_i-\mathbf{z}_j}{h_m}\right) + K\left(\frac{\mathbf{x}-\mathbf{z}_j}{h_m}\right)\right) \\ &\quad \times \log\left|\frac{\sum_{i=1}^{m-1} K\left(\frac{\mathbf{x}_i-\mathbf{z}_j}{h_m}\right) + K\left(\frac{\mathbf{x}-\mathbf{z}_j}{h_m}\right)}{mh_m^d \hat{f}^{(m-1)}(\mathbf{z}_j)}\right| \\ &= \arg \min_{\mathbf{x}\in\mathcal{Z}\setminus\{\mathbf{x}_i\}_{i=1}^{m-1}} M'\left(\mathbf{x}|\{\mathbf{x}_i, y_i\}_{i=1}^{m-1}, \{\mathbf{z}_j\}_{j=1}^{N}\right),\end{aligned}$$

where, $M'(\mathbf{x}|\{\mathbf{x}_i, y_i\}_{i=1}^{m-1}, \{\mathbf{z}_j\}_{j=1}^{N}) = \frac{1}{N}\sum_{j=1}^{N} \frac{1}{mh_m^d}\left(\sum_{i=1}^{m-1} K\left(\frac{\mathbf{x}_i-\mathbf{z}_j}{h_m}\right) + K\left(\frac{\mathbf{x}-\mathbf{z}_j}{h_m}\right)\right)$ $\times \log\left|\frac{\sum_{i=1}^{m-1} K\left(\frac{\mathbf{x}_i-\mathbf{z}_j}{h_m}\right)+K\left(\frac{\mathbf{x}-\mathbf{z}_j}{h_m}\right)}{mh_m^d \hat{f}^{(m-1)}(\mathbf{z}_j)}\right|$.

We also choose the multivariate exponential power distribution as the kernel function $K$, and $h_m$ is chosen as similar in Sect. 3, i.e. $h_m = d^{-1/(d+4)} d^{-1} \sum_{i=1}^{d} \hat{\sigma}_i^2 m^{-1/(d+4)}$, where $\hat{\sigma}_i^2$ is the estimation variance for the $i$-th dimension of $f$ with the sampling points. We call the KL points generated by the above algorithm as *sequential Kullback-Leibler (SKL) points*.

Algorithm 2 outlines the detailed process for generating $n - n_0$ SKL points based on the initial set $D_{n_0}$.

**Algorithm 2** Generating the $n - n_0$ SKL points via the greedy algorithm

**Require:** $n$, $D_{n_0}$, $\mathcal{Z} = \{\mathbf{z}_j\}_{j=1}^{N}$

**Ensure:** $\mathbf{x}_{n_0+1}, \ldots, \mathbf{x}_n$

1: Compute $\{y_i\}_{i=1}^{n_0}$ by $y_i = f(\mathbf{x}_i)$
2: **for** $m = (n_0 + 1), \ldots, n$ **do**
3: Estimate $\hat{f}^{(m-1)}$ via $\{\mathbf{x}_j, y_j\}_{j=1}^{m-1}$
4: Compute $\hat{f}^{(m-1)}(\mathbf{z}_1), \ldots, \hat{f}^{(m-1)}(\mathbf{z}_N)$
5: Update $M'(\mathbf{x}|\{\mathbf{x}_i, y_i\}_{i=1}^{m-1}, \{\mathbf{z}_j\}_{j=1}^{N})$
6: $\mathbf{x}_m = \arg\min_{\mathbf{x}\in\mathcal{Z}\setminus\{\mathbf{x}_i\}_{i=1}^{m-1}} M'(\mathbf{x}|\{\mathbf{x}_i, y_i\}_{i=1}^{m-1}, \{\mathbf{z}_j\}_{j=1}^{N})$
7: $y_m = f(\mathbf{x}_m)$
8: **end for**

The rest of this section demonstrates two important applications of SKL points.

### 5.1 Simulation from complex probability densities

When density is complex and expensive to evaluate, Algorithm 1 may be a time-consuming implementation, and we will not be able to find the KL points efficiently by directly minimizing the Kullback-Leibler divergence. Hence, we adopt Algorithm 2 to generate the SKL points of the expensive density. Here, we choose a Gaussian process model (or kriging) to estimate the expensive density. Because the approximated density should be nonnegative, we fit the following stationary Gaussian process model (or ordinary kriging) after taking a logarithmic transformation of the density (which can be unnormalized ):

$$\log f(\cdot) \sim GP(\mu, \sigma^2 R(\cdot)),$$

where the correlation function is defined as $\text{cor}(\log f(\mathbf{x}_i), \log f(\mathbf{x}_j)) = R(\mathbf{x}_i - \mathbf{x}_j)$. If we generate an initial set of $n_0$-points using a maximin Latin hypercube design $D_{n_0} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_0}\}$ and obtain the outputs $y^{(n_0)} = (y_1, \ldots, y_{n_0})'$, where $y_i = f(\mathbf{x}_i)$, then the meta-model is given by (Sacks et al. 1989)

$$\hat{f}^{(n_0)}(\mathbf{x}) = \exp\left\{\hat{\mu}^{(n_0)} + r^{(n_0)}(\mathbf{x})' R_{(n_0)}^{-1}\left(y^{(n_0)} - \hat{\mu}^{(n_0)}\mathbf{1}_{n_0}\right)\right\}, \tag{9}$$

where $r^{(n_0)}(\mathbf{x})$ is a vector of length $n_0$ with the $i$-th element $R(\mathbf{x} - \mathbf{x}_i)$, $R_{(n_0)}$ is an $n_0 \times n_0$ correlation matrix with the $ij$-th element $R(\mathbf{x}_i - \mathbf{x}_j)$, $\mathbf{1}_{n_0}$ is a vector of $1's$, and $\hat{\mu}^{(n_0)} = (\mathbf{1}'_{n_0} R_{(n_0)}^{-1} \mathbf{1}_{n_0})^{-1} \mathbf{1}'_{n_0} R_{(n_0)}^{-1} y^{(n_0)}$. For the correlation function, we choose the popular Gaussian correlation function given by

$$R(\mathbf{t}) = \exp\left\{-\sum_{i=1}^{d} \theta_i t_i^2\right\}.$$

Finally, conducting Algorithm 2, we can obtain the SKL points of the density function $f$.

For example, consider the two-dimensional probability density with banana-shaped contours given by equation (1). We generate $n$=50 points; the first $n_0 = 20$ points are a maximin Latin hypercube design, and the remaining 30 points are generated using the SKL points in Algorithm 2 marked as $*$'s in Fig. 4. For comparison, via the same initial points, we also generate 30 points using the SMED marked as $*$'s in Fig. 4. Clearly, the SKL points are a better set of representative points than those of the SMED in this case. In addition, we find that the SMED is sensitive to different initial points and that the SKL points are robust to different initial points and produce better representative points.

### 5.2 Exploration and optimization of complex black-box functions

Global optimization is an important but difficult problem. Many algorithms, such as simulated annealing and genetic algorithms, can be used to search for global optima.

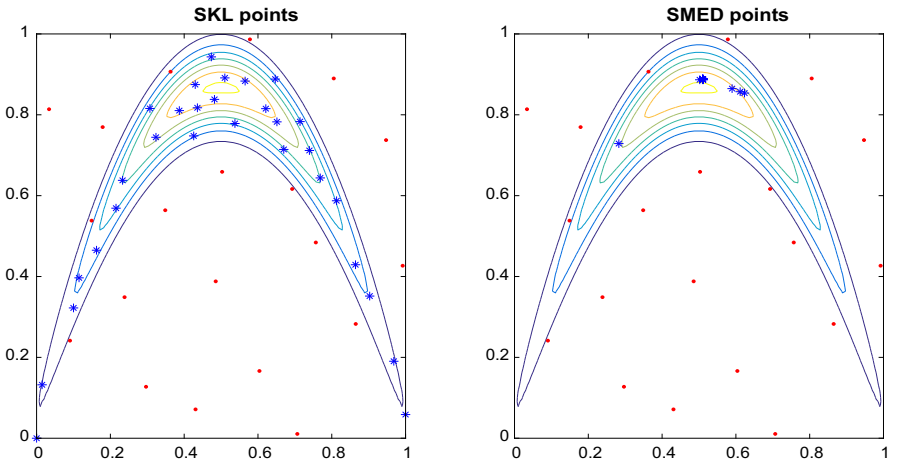


**Fig. 4** Comparison of SKL points and SMED points with the same initial points, which are shown as •'s

However, these algorithms require numerous evaluations of the function, which can become costly if the objective function is complex and expensive to evaluate. By a closer look at Algorithm 2, we can see that this algorithm can generate the SKL points for any nonnegative function (can be regarded as a non-standard density) with high-density and ensure them spread out well. Hence, suppose that the complex black-box function is nonnegative, i.e. $f(\mathbf{x}) \geq 0$, then we can treat any nonnegative complex black-box function as a non-standard density, and then applying Algorithm 2 to obtain its SKL points. These SKL points may contain more points with large function values.

Instead of tuning the SKL points for the purpose of global optimization, we focus on a slightly different objective which is not only to find a global optimum, but also to find several good points that can serve as alternatives to the global optimum ( SKL points have a greater probability of choosing to a value around the maximum). This situation arises quite often in multi-objective optimization (Miettinen 2012). Obviously, SKL points may offer a good solution to this problem. Without loss of generality, let the optimization problem be to maximize $f$ in some bounded region $\mathcal{X}$. Similar to Sect. 5.1, we again choose a Gaussian process model to estimate any nonnegative complex black-box function. In other words, based on an initial space-filling point set containing $n_0$ points, we still choose $\hat{f}^{(n_0)}$ as in (9) to estimate $f$. Applying Algorithm 2, we can obtain the SKL points of $f$.

For illustration, we consider Franke's two-dimensional function (Fasshauer 2007)

$$
\begin{aligned}
f(\mathbf{x}) = \frac{3}{4}\exp\left\{-\frac{1}{4}(9x_1-2)^2 - \frac{1}{4}(9x_2-2)^2\right\} \\
+\frac{3}{4}\exp\left\{-\frac{1}{49}(9x_1+1)^2 - \frac{1}{10}(9x_2+1)^2\right\} \\
+\frac{1}{2}\exp\left\{-\frac{1}{4}(9x_1-7)^2 - \frac{1}{4}(9x_2-3)^2\right\}
\end{aligned}
$$

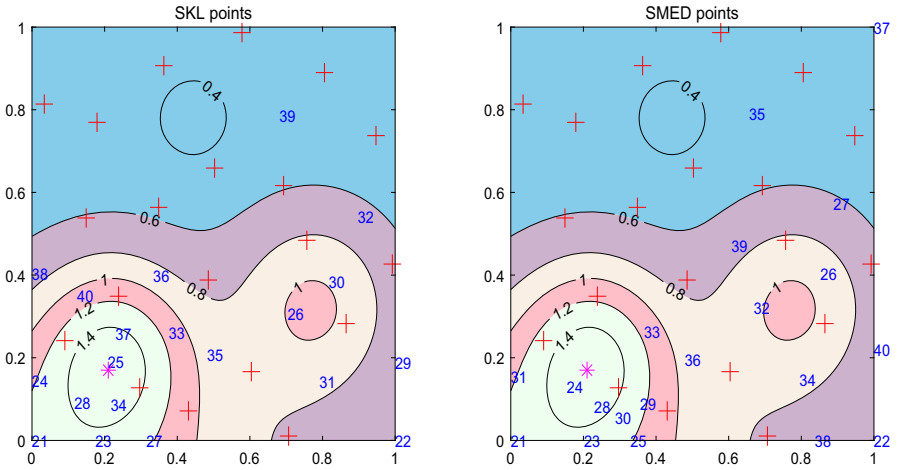


**Fig. 5** Comparison of SKL points and SMED points: the +'s represent the initial point set, and the next points appear in order sequentially

$$-\frac{1}{5}\exp\left\{-(9x_1-4)^2-(9x_2-2)^2\right\}+0.5.$$

The initial space-filling point set is also a maximin Latin hypercube design with $n_0 = 20$ points. Applying Algorithm 2, we can obtain 20 SKL points on the left of Fig. 5, where the new points are numbered 21–40. The global optimum of $f$ is $\mathbf{x}^* = (0.21, 0.17)$, which is denoted as $*$ in Fig. 5. The approximate global optimum finding by SKL points is the 25th points $\mathbf{x}_{\text{SKL}} = (0.209, 0.192)$.

For comparison, we also generate points by SMED via the same 20 initial points. The next 20 SMED points are shown on the right side of Fig. 5 numbered 21–40. The approximate global optimum of $f$ finding by SMED is $\mathbf{x}_{\text{SMED}} = (0.155, 0.130)$ in the 24th point. From Fig. 5, we can see that: 1) the approximate global optimum finding by SKL points is more close to the true global optimum than SMED points; 2) SKL points can provide more alternative points around the true global optimum than SMED points.

Given that the choice of the initial design may affect both the SKL points and SMED in finding the global optimum, we utilize 50 randomly generated Latin hypercube designs as the initial design to compare the performance of SKL points and SMED in terms of finding the global optimum. The boxplots in Fig. 6 show the Euclidean distance of the optimum found by the SKL points and SMED points from the true global optimum $\mathbf{x}^*$ with both methods repeated 50 times. As expected, the SKL points show better performance in finding the global optimum.

## 6 Conclusion

This article proposed a new sampling method by minimizing the Kullback-Leibler divergence, which causes the sample points to be concentrated in regions with high

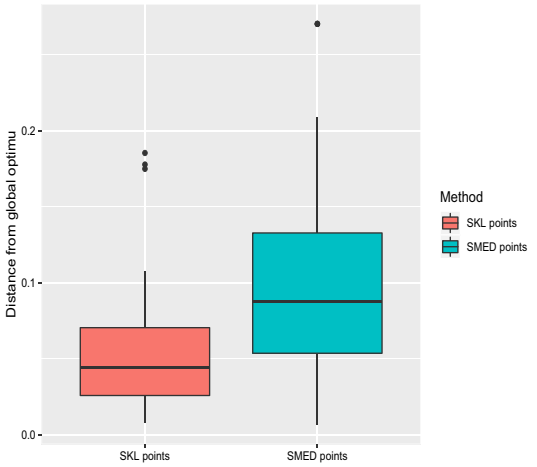


**Fig. 6** Boxplots for the Euclidean distance of the optimum found by the SKL points and SMED points with both methods repeated 50 times

densities and sufficiently spreads the points away from each other. The proposed KL points are based on ideas from information theory, which is used to measure the distance between two distributions. We derived some theoretical results to show that the limiting distribution of KL points converge to the target distribution. The simulations show that the KL points are a better set of representative points of $F$ when $n$ is small. Moreover, to address complex functions, we developed a sequential algorithm for adaptive implementation of KL points, called SKL points, in Sect. 5. Compared with the existing methods, the SKL points may perform better in simulation and optimization.

The MC method is flexible and easy to implement in practice as long as the distribution is fairly inexpensive to evaluate. The clear advantage of KL points over MC is that the distribution information provided by KL points is roughly the same as that provided by MC with multiple sample sizes of KL points. This advantage is important in the small-data application for the expensive distributions in Bayesian problems and expensive computer simulations when the dimension is not very large.

While this paper establishes some interesting results for KL points, there are still many exciting avenues for future research. First, the advantage of KL points generated via the kernel density estimate is not apparent for high-dimensional distributions. At this point, we suggest generating KL points based on estimating the density via the nearest neighbour distance (Wang et al. 2006), i.e., by using $f_n^{\text{Neib}}(\mathbf{x}) = \frac{1}{nV_1(d)\rho_1^d(\mathbf{x}_i)} I(\mathbf{x} \in S(\mathbf{x}_i, \rho_1(\mathbf{x}_i)))$ instead of $f_n(\mathbf{x}) = \frac{1}{nh_n^d}\sum_{i=1}^n K(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})$, where $V_1(d) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ denotes the volume of the unit ball in $R^d$, $\rho_1(\mathbf{x}_i) = \min_{j\neq i} \|\mathbf{x}_i - \mathbf{x}_i\|_2$, and $S(\mathbf{x}_i, \rho_1(\mathbf{x}_i))$ denotes the ball with centre $\mathbf{x}_i$ and radius $\rho_1(\mathbf{x}_i)$. In this case, an unbiased estimator of the Kullback–Leibler divergence between $f_n^{\text{Neib}}$ and $f$ is

$$D(f_n^{\text{Neib}} \| f) = -\left(\frac{d}{n}\sum_{i=1}^n \log \rho_1(\mathbf{x}_i) + \frac{1}{n}\sum_{i=1}^n \log f(\mathbf{x}_i) + \log V_1(d) + \log n + \gamma\right),$$

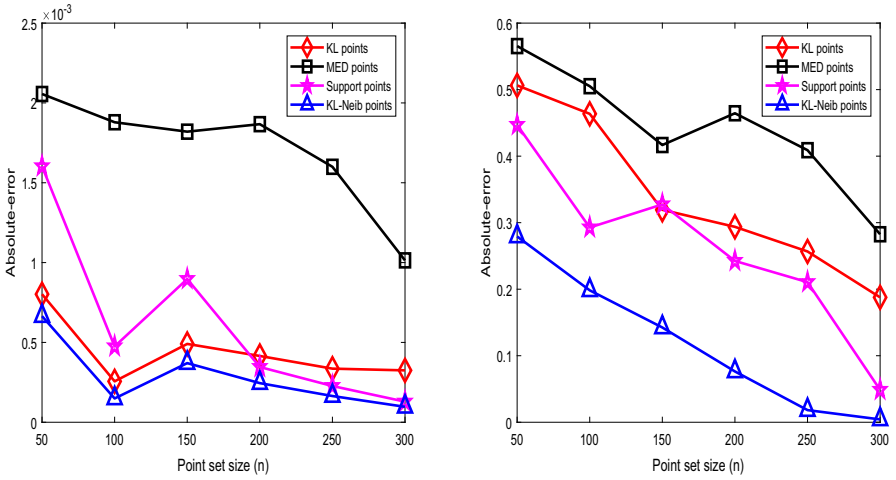


**Fig. 7** The absolute errors of numerical integration for $g(\mathbf{x}) = \exp(-\sum_{l=1}^{d} \alpha_l (x_l - \mu_l)^2)$ under a truncated $N(0, I)$ on $[-5, 5]^d$ (**a**), and $g(\mathbf{x}) = \exp(-x_1^2 - x_2 - x_3)\sin(x_6 x_{10})$ under a truncated $N(0, \sigma^2 R)$ on $[-5, 5]^d$ (**b**), where $\alpha_l = 20/d$, $\mu_l = 0$, and $\sigma = 1$, $R_{ij} = 0.5^{|i-j|}$ for $i, j = 1, \ldots, d$

where $\gamma$ is the Euler constant. The KL points obtained by minimizing $D(f_n^{\text{Neib}} \| f)$ are referred to as KL-Neib points, which may enjoy smaller errors for high-dimensional numerical integration (see Fig. 7). However, these new methods of obtaining KL points can be costly in terms of time; an efficient algorithm for obtaining KL-Neib points in high dimensions will be one direction for future work. Next, motivated by Hickernell (1998), the KL points in high dimensions should provide a good representation of not only the full distribution $F$ but also of marginal distributions of $F$. Such a projective property is enjoyed by most QMC point sets (Dick et al. 2013). It would also be interesting to incorporate this within the KL point framework.

**Acknowledgements** This study was supported by the National Natural Science Foundation of China (Grant Nos. 11971098, 11471069 and 12131001) and National Key Research and Development Program of China (Grant Nos. 2020YFA0714102 and 2022YFA1003701).

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## Appendix A: Proofs

This appendix material provides proofs of Theorems 1–3. Define $f_n^*$ as the kernel density estimator on $\{\mathbf{x}_j^*\}_{j=1}^n$, where $\{\mathbf{x}_j^*\}_{j=1}^n \overset{i.i.d.}{\sim} f(\mathbf{x})$. The proof of Theorem 1 relies on Lemma 1 below, which indicates that the expectation of $D(f_n^* \| f)$ converges to 0 as $n \to \infty$.

**Lemma 1** *Suppose probability density function $f$ satisfies conditions (A1)–(A2) and that $\{\mathbf{x}_j^*\}_{j=1}^n \overset{i.i.d.}{\sim} f$. $f_n^*$ is the kernel density estimator on $\{\mathbf{x}_j^*\}_{j=1}^n$, and the kernel function $K$ and the bandwidth $h_n$ satisfy (K1)–(K5). Then,*

$$\lim_{n\to\infty} E[D(f_n^* \| f)] = 0.$$

***Proof*** For simplicity, we take $d = 1$ for example, the proof for $d > 1$ is similar. Let $\{x_j^*\}_{j=1}^n \overset{i.i.d.}{\sim} f$, $f_n^*(x) = \frac{1}{nh_n}\sum_{i=1}^n K(\frac{x-x_i^*}{h_n})$ be the kernel density estimator based on $\{x_j^*\}_{j=1}^n$, and kernel function $K$ and the bandwidth $h_n$ satisfy (K1)–(K5).

Since $f$ satisfies (A1)-(A2), then $f$ is continuous over $\mathcal{X}$. And note that $f$ is a density function then we know that there exists constant $f_{\max}$ such $f \le f_{\max} < \infty$ holds. The variance of $f_n^*(x)$ satisfies:

$$\begin{aligned}
\text{var}\,(f_n^*(x)) &= \text{var}\left(\frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{x-x_i^*}{h_n}\right)\right) \\
&\le \frac{1}{nh_n^2} E\left[K^2\left(\frac{x-x_1^*}{h_n}\right)\right] \\
&= \frac{1}{nh_n^2}\int_{\mathcal{X}} K^2\left(\frac{z-x}{h_n}\right) f(z)dz \\
&= \frac{1}{nh_n}\int_{\mathcal{X}} K^2(t) f(th_n + x)dt \\
&\le \frac{f_{\max}}{nh_n}\int_{\mathcal{X}} K^2(t)dt = \frac{C_1}{nh_n}, \qquad \text{(A1)}
\end{aligned}$$

where $C_1 = f_{\max}\|K\|_2^2$, $\|K\|_2^2 = \int_{\mathcal{X}} K^2(t)dt$.

The bias of $f_n^*(x)$ have the following results.

$$\begin{aligned}
E[f_n^*(x)] - f(x) &= \frac{1}{nh_n}\sum_{i=1}^n E\left[K\left(\frac{x-x_i^*}{h_n}\right)\right] - f(x) \\
&= \frac{1}{h_n}\int_{\mathcal{X}} K\left(\frac{z-x}{h_n}\right) f(z)dz - f(x) \\
&= \int_{\mathcal{X}} K(t) f(th_n + x)dt - f(x) \\
&= \int_{\mathcal{X}} K(t)[f(th_n + x) - f(x)]dt. \qquad \text{(A2)}
\end{aligned}$$

In term of $f$ satisfies $\forall\, x, y \in \mathcal{X}$, $|f(x) - f(y)| \le L|x - y|$. Hence, we obtain

$$\begin{aligned}
|E[f_n^*(x)] - f(x)| &\le \int_{\mathcal{X}} K(t)|Lth_n|dt \\
&= Lh_n\int_{\mathcal{X}} |t|K(t)dt
\end{aligned}$$

$$\leq Lh_n \left\{ \int_{\mathcal{X}} t^2 K(t) dt \right\}^{1/2} = C_2 h_n, \tag{A3}$$

where $C_2 = L \left\{ \int_{\mathcal{X}} t^2 K(t) dt \right\}^{1/2}$.

The Kullback–Leibler divergence between $f$ and $f_n^*$ is

$$D(f_n^* \| f) = \int_{\mathcal{X}} f_n^*(x) \log \frac{f_n^*(x)}{f(x)} dx.$$

By the inequality $\log x \leq (x-1)$, the expectation of $D(f_n^* \| f)$ satisfies:

$$\begin{aligned}
E[D(f_n^* \| f)] &= E\left[ \int_{\mathcal{X}} f_n^*(x) \log \frac{f_n^*(x)}{f(x)} dx \right] \\
&\leq E\left[ \int_{\mathcal{X}} f_n^*(x) \left( \frac{f_n^*(x)}{f(x)} - 1 \right) dx \right] \\
&= E\left[ \int_{\mathcal{X}} f_n^*(x) \frac{f_n^*(x)}{f(x)} dx \right] - 1 \\
&= \int_{\mathcal{X}} E\left[ f_n^*(x) \frac{f_n^*(x)}{f(x)} \right] dx - 1 \\
&= \int_{\mathcal{X}} \frac{E[f_n^*(x)]^2 - f^2(x)}{f(x)} dx \\
&\leq \int_{\mathcal{X}} \frac{|E[f_n^*(x)]^2 - f^2(x)|}{f(x)} dx.
\end{aligned} \tag{A4}$$

The penultimate "=" sign holds following from Fubini theorem.

According to (A1) and (A3), we can obtain

$$\begin{aligned}
|E[f_n^*(x)]^2 - f^2(x)| &= \left| E[f_n^*(x) - f(x) + f(x)]^2 - f^2(x) \right| \\
&= \left| E[f_n^*(x) - f(x)]^2 + 2f(x) E[f_n^*(x) - f(x)] \right| \\
&\leq E[f_n^*(x) - f(x)]^2 + 2f(x) \left| E[f_n^*(x)] - f(x) \right| \\
&= E\left\{ f_n^*(x) - E[f_n^*(x)] + E(f_n^*(x)) - f(x) \right\}^2 \\
&\quad + 2f(x) \left| E[f_n^*(x)] - f(x) \right| \\
&= \text{var}\,(f_n^*(x)) + \left| E[f_n^*(x)] - f(x) \right|^2 + 2f(x) \left| E[f_n^*(x)] - f(x) \right| \\
&\leq \text{var}\,(f_n^*(x)) + C_2^2 h_n^2 + 2f(x) \left| E[f_n^*(x)] - f(x) \right| \\
&\leq \frac{C_1}{nh_n} + C_2^2 h_n^2 + 2C_2 h_n f(x) \triangleq G_n(x).
\end{aligned} \tag{A5}$$

Obviously, $\frac{G_n(x)}{f(x)}$ is monotonically decreasing in $n$, and $\lim_{n\to\infty} \frac{G_n(x)}{f(x)} = 0$. So, $\forall n$, $\frac{G_n(x)}{f(x)} \leq \frac{G_1(x)}{f(x)}$. Due to $\int_{\mathcal{X}} \frac{G_1(x)}{f(x)} dx < \infty$, then, by the Lebesgue's convergence theo-

rem (Billingsley 2008), we obtain

$$\lim_{n\to\infty}\int_{\mathcal{X}}\frac{G_n(x)}{f(x)}dx=\int_{\mathcal{X}}\lim_{n\to\infty}\frac{G_n(x)}{f(x)}dx=0. \tag{A6}$$

Note that

$$\int_{\mathcal{X}}\frac{|E[f_n^*(x)^2]-f^2(x)|}{f(x)}dx\le\int_{\mathcal{X}}\frac{G_n(x)}{f(x)}dx$$

and we have

$$\lim_{n\to\infty}\int_{\mathcal{X}}\frac{|E[f_n^*(x)^2]-f^2(x)|}{f(x)}dx=0. \tag{A7}$$

Then the conclusion $\lim_{n\to\infty}E[D(f_n^*\|f)]=0$ is established.

To prove Theorems 1 and 2, we also need the following lemma. □

**Lemma 2** *Let $f$ and $g$ be two density functions supported on $\mathcal{X}\subseteq R^d$, then*

(a)

$$V(f,g)\overset{\text{def}}{=}\sup_{A\in\mathcal{B}}\left|\int_A(f(\mathbf{x})-g(\mathbf{x}))d\mathbf{x}\right|=\frac{1}{2}\int_{\mathcal{X}}|f(\mathbf{x})-g(\mathbf{x})|d\mathbf{x},$$

*where $\mathcal{B}$ is the Borel $\sigma$-algebra of $\mathcal{X}$.*

(b)

$$2V^2(f,g)\le D(g\|f).$$

This Lemma can be obtained by Scheffé's theorem (refer to Tsybakov 2009 p.84) and Pinsker's inequality [refer to Tsybakov (2009, p. 88)], respectively.

***Proof of Theorem 1.*** Define the sequence of random variables $\{\mathbf{x}_j^*\}_{j=1}^{\infty}\overset{i.i.d.}{\sim}f$, and let $f_n^*$ denote the kernel density estimator on $\{\mathbf{x}_j^*\}_{j=1}^n$. According the Lemma 1, $\lim_{n\to\infty}E[D(f_n^*\|f)]=0$.

Consider now the kernel density estimator $f_n^{KL}$ on the KL points $\{\boldsymbol{\xi}_i\}_{i=1}^n$. By the definition of KL points,

$$D(f_n^{KL}\|f)\le E[D(f_n^*\|f)],$$

so $\lim_{n\to\infty}D(f_n^{KL}\|f)=0$.

Using Pinsker's inequality in Lemma 2 (b),

$$\frac{1}{2}\left(\int_{\mathcal{X}}|f_n^{KL}(\mathbf{x})-f(\mathbf{x})|d\mathbf{x}\right)^2\le D(f_n^{KL}\|f),$$

then conclusion $\lim_{n\to\infty}\int_{\mathcal{X}}|f_n^{KL}(\mathbf{x})-f(\mathbf{x})|d\mathbf{x}=0$ follows.

***Proof of Theorem 2*** Due to $f_n^{KL}$ is the kernel density estimator on KL points $\{\boldsymbol{\xi}_i\}_{i=1}^n$ and satisfies (K1)–(K5), then

$$\sup_{A\in\mathcal{B}}\left|\int_A [f_n^{KL}(\mathbf{x}) - f(x)]d\mathbf{x}\right| = \frac{1}{2}\int_{\mathcal{X}} |f_n^{KL}(\mathbf{x}) - f(\mathbf{x})|d\mathbf{x} \tag{A8}$$

based on Lemma 2 (a). Combing Theorem 1 and (A8), we have

$$\lim_{n\to\infty}\sup_{A\in\mathcal{B}}\left|\int_A [f_n^{KL}(\mathbf{x}) - f(\mathbf{x})]d\mathbf{x}\right| = 0. \tag{A9}$$

Set $A = (-\infty, \mathbf{x}] \in \mathcal{B}$, we have

$$\lim_{n\to\infty}\sup_{\mathbf{x}}|F_n(\mathbf{x}) - F(\mathbf{x})| = 0.$$

where $F_n$ is the cumulative distribution function of density function $f_n^{KL}$, and $F$ is the cumulative distribution function of $f$. □

Two lemmas will be needed for the proof of Theorem 3.

**Lemma 3** *Suppose kernel $K$ satisfys (K1)–(K5). Then,*

(a) $\forall\, \epsilon > 0$, *there exist* $M > 0$, *such that* $\int_{[-M,M]^d} K(\mathbf{t})d\mathbf{t} \geq 1 - \epsilon$.
(b) *For the above* $M > 0$, $\exists\, \mathbf{y}_0$, *such that* $\varphi(\mathbf{x}, \mathbf{y}_0) \geq 1/3$, *where*

$$\varphi(\mathbf{x}, \mathbf{y}) = \int_{\prod_{i=1}^d [x_i - y_i, x_i + y_i]} K(\mathbf{t})d\mathbf{t},$$

$\mathbf{x} = (x_1, \ldots, x_d) \in [-M, M]^d$ *and* $\mathbf{y} = (y_1, \ldots, y_d)$.

***Proof*** (a) follows from is a density function.

(b) Note that, $\lim_{\mathbf{y}\to+\infty}\varphi(\mathbf{x}, \mathbf{y}) \geq 1/2$ for any $\mathbf{x} = (x_1, \ldots, x_d) \in [-M, M]^d$, then conclusion follows. □

**Lemma 4** *Let $\{\boldsymbol{\xi}_i\}_{i=1}^n$ be the KL points of $f$. Then*

$$\lim_{n\to\infty}\sup_{\mathbf{x}\in\mathcal{X}}\frac{N_{\mathbf{x}}}{n} = 0,$$

*where*

$$N_{\mathbf{x}} = \sum_{i=1}^n I\left(\frac{\mathbf{x} - \boldsymbol{\xi}_i}{h_n} \in [-M, M]^d\right),$$

*$h_n$ is the bandwidth used to generate KL points and $M$ is defined in Lemma 3.*

***Proof*** For simplicity, we take $d = 1$ for example. Due to $\forall\ \delta > 0$, $\lim_{n\to\infty} \frac{\delta}{h_n} = +\infty$, so $\exists\ n_0$, such that when $n > n_0$, $\frac{\delta}{h_n} \geq y_0$ holds, where $y_0$ satisfies $I(|x| \leq M)\varphi(x, y_0) \geq 1/3$ defined in Lemma 3.

We use reduction to absurdity to prove this result. If Lemma 4 doesn't hold, then $\exists\ x^*$, for $\forall\ N > n_0, \exists\ n_k > N$, such that $\frac{N_{x^*}}{n_k} \geq c_0$, which means

$$\frac{\sum_{i=1}^{n_k} I(|\frac{x^*-\xi_i}{h_{n_k}}| \leq M)}{n_k} \geq c_0,$$

where $c_0$ is a positive constant.

Then, we have

$$\begin{aligned}
F_{n_k}(x^*+\delta) - F_{n_k}(x^*-\delta) &= \frac{1}{n_k}\sum_{i=1}^{n_k}\int_{x^*-\delta}^{x^*+\delta} \frac{1}{h_{n_k}} K\left(\frac{t-\xi_i}{h_{n_k}}\right) dt \\
&= \frac{1}{n_k}\sum_{i=1}^{n_k}\int_{\frac{x^*-\xi_i}{h_{n_k}}-\frac{\delta}{h_{n_k}}}^{\frac{x^*-\xi_i}{h_{n_k}}+\frac{\delta}{h_{n_k}}} K(z)dz \\
&\geq \frac{1}{n_k}\sum_{i=1}^{n_k} I\left(\left|\frac{x^*-\xi_i}{h_{n_k}}\right| \leq M\right)\int_{\frac{x^*-\xi_i}{h_{n_k}}-y_0}^{\frac{x^*-\xi_i}{h_{n_k}}+y_0} K(z)dz \\
&\geq \frac{\sum_{i=1}^{n_k} I(|\frac{x^*-\xi_i}{h_{n_k}}| \leq M)}{3n_k} \\
&\geq \frac{c_0}{3},
\end{aligned}$$

where $F_{n_k}$ is the cumulative distribution function correspond the kernel density estimator $f_{n_k}$, which used to generated KL points. The penultimate "$\geq$" holds by Lemma 3. This contradicts to $F_{n_k}$ is a continuous distribution. We conclude the proof. □

***Proof of Theorem 3*** Let $F_n^{KL}$ denote the standard empirical distribution of KL ponits $\{\boldsymbol{\xi}_i\}_{i=1}^n$, $F_n$ denote the cumulative distribution function of $f_n^{KL}$. We first prove that

$$\lim_{n\to\infty} \sup_{\mathbf{x}} |F_n(\mathbf{x}) - F_n^{KL}(\mathbf{x})| = 0.$$

For simplicity, we take $d = 1, \forall\ x \in \mathcal{X}$,

$$\begin{aligned}
|F_n(x) - F_n^{KL}(x)| &= \frac{1}{n}\sum_{i=1}^{n}\left|\int_{-\infty}^{x} \frac{1}{h_n} K\left(\frac{t-\xi_i}{h_n}\right) dt - I(\xi_i \leq x)\right| \\
&= \frac{1}{n}\sum_{i=1}^{n}\left|\int_{-\infty}^{\frac{x-\xi_i}{h_n}} K(z)dz - I(\xi_i \leq x)\right|.
\end{aligned}$$

If $\xi_i \le x$, then $\frac{x-\xi_i}{h_n} \ge 0$, and

$$\begin{aligned}\left|\int_{-\infty}^{\frac{x-\xi_i}{h_n}} K(z)dz - I(\xi_i \le x)\right| &= \left|\int_{-\infty}^{\frac{x-\xi_i}{h_n}} K(z)dz - 1\right| \\ &= \int_{\frac{x-\xi_i}{h_n}}^{+\infty} K(z)dz \\ &\le I\left(0 \le \frac{x-\xi_i}{h_n} \le M\right)\int_{\frac{x-\xi_i}{h_n}}^{M} K(z)dz + \frac{\epsilon}{2},\end{aligned}$$

where $M$ and $\epsilon$ are defined as in Lemma 3, i.e. $\forall\, \epsilon > 0$, there exist $M > 0$, such that $\int_{-M}^{M} K(t)dt \ge 1-\epsilon$, and $\int_{-\infty}^{-M} K(t)dt = \int_{M}^{+\infty} K(t)dt < \frac{\epsilon}{2}$.

If $\xi_i > x$, then $\frac{x-\xi_i}{h_n} < 0$, and

$$\begin{aligned}\left|\int_{-\infty}^{\frac{x-\xi_i}{h_n}} K(z)dz - I(\xi_i \le x)\right| &= \left|\int_{-\infty}^{\frac{x-\xi_i}{h_n}} K(z)dz - 0\right| \\ &\le I\left(-M \le \frac{x-\xi_i}{h_n} < 0\right)\int_{-M}^{\frac{x-\xi_i}{h_n}} K(z)dz + \frac{\epsilon}{2}.\end{aligned}$$

In summary, for $\forall\, x \in \mathcal{X}$ the absolute error between $F_n(x)$ and $F_n^{KL}(x)$ is

$$\begin{aligned}0 \le |F_n(x) - F_n^{KL}(x)| &\le \frac{1}{n}\sum_{i=1}^{n} I\left(\left|\frac{x-\xi_i}{h_n}\right| \le M\right)\int_{|\frac{x-\xi_i}{h_n}|}^{M} K(z)dz + \frac{\epsilon}{2} \\ &\le \frac{1}{2n}\sum_{i=1}^{n} I\left(\left|\frac{x-\xi_i}{h_n}\right| \le M\right) + \frac{\epsilon}{2}. \qquad \text{(A10)}\end{aligned}$$

In term of Lemma 4,

$$\lim_{n\to\infty} \sup_x \frac{1}{n}\sum_{i=1}^{n} I\left(\left|\frac{x-\xi_i}{h_n}\right| \le M\right) = 0.$$

Hence, $\forall\, \epsilon > 0$, there exist $N$, such that when $n \ge N$, for $\forall\, x \in \mathcal{X}$, we have

$$0 \le |F_n(x) - F_n^{KL}(x)| \le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Consequently, we obtain $\lim_{n\to\infty} \sup_{\mathbf{x}} |F_n(x) - F_n^{KL}(x)| = 0$. The proof for $d > 1$ is similar, hence

$$\lim_{n\to\infty} \sup_{\mathbf{x}} |F_n(\mathbf{x}) - F_n^{KL}(\mathbf{x})| = 0. \qquad \text{(A11)}$$

Due to

$$\sup_{\mathbf{x}}|F_n^{KL}(\mathbf{x}) - F(\mathbf{x})| \leq \sup_{\mathbf{x}}|F_n^{KL}(\mathbf{x}) - F_n(\mathbf{x})| + \sup_{\mathbf{x}}|F_n(\mathbf{x}) - F(\mathbf{x})|,$$

and combing with Theorem 2, we have

$$\lim_{n\to\infty} \sup_{\mathbf{x}}|F_n^{KL}(\mathbf{x}) - F(\mathbf{x})| = 0.$$

□

## References


Billingsley P (2008) Probability and measure. Wiley, New York

Brooks S, Gelman A, Jones G, Meng XL (2011) Handbook of Markov chain Monte Carlo. Chapman and Hall/CRC, Boca Raton

Chen WY, Mackey L, Gorham J, Briol FX, Oates CJ (2018) Stein points. In Proceedings of the 35th international conference on machine learning, vol 80, pp 844–853

Cover TM, Thomas JA (2006) Elements of information theory. Wiley, New York

Dick J, Kuo FY, Sloan IH (2013) High-dimensional integration: the quasi-Monte Carlo way. Acta Numer 22:133–573

Dudewicz EJ, Van DMEC (1981) Entropy-based tests of uniformity. J Am Stat Assoc 76:967–974

Fang KT, Li RZ, Sudijanto A (2006) Designs and modeling for computer experiments. Chapman and Hall/CRCl, Boca Raton

Fasshauer G (2007) Meshfree approximation methods with MATLAB. World Scientific, Singapore

Haario H, Saksman E, Tamminen J (1999) Adaptive proposal distribution for random walk Metropolis algorithm. Comput Stat 14:375–395

Härdle WG, Werwatz A, Müller M, Sperlich S (2004) Nonparametric and semiparametric models. Springer, New York

Hickernell F (1998) A generalized discrepancy and quadrature error bound. Math Comput 67:299–322

Lin CD, Tang BX (2015) Latin hypercubes and space-filling designs. Handb Des Anal Exp Chap 17:593–626

Joseph VR, Dasgupta T, Tuo R, Wu CFJ (2015) Sequential exploration of complex surfaces using minimum energy designs. Technometrics 57:64–74

Joseph VR, Wang DP, Gu L, Lv SJ, Tuo R (2019) Deterministic sampling of expensive posteriors using minimum energy designs. Technometrics 61:297–308

Jourdan A, Franco J (2010) Optimal Latin hypercube designs for the Kullback–Leibler criterion. AStA Adv Stat Anal 94:341–351

Kennedy MC, O'Hagan A (2001) Bayesian calibration of computer models (with discussion). J R Stat Soc B 63:425–464

Kullback S, Leibler R (1951) On information and sufficiency. Ann Math Stat 22:79–86

Mak S, Joseph VR (2017) Projected support points: a new method for high-dimensional data reduction. arXiv: 1708.06897

Mak S, Joseph VR (2018) Support points. Ann Stat 46:2562–2592

McKay MD, Beckman RJ, Conover WJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21:239–245

Morris MD, Mitchell TJ (2018) Exploratory designs for computer experiments. J Stat Plan Inference 43:381–402

Miettinen K (2012) Nonlinear multiobjective optimization. Springer, New York

Sacks J, Schiller SB, Welch WJ (1989) Designs for computational experiments. Technometrics 31:41–47

Santner TJ, Williams BJ, Notz WI (2019) The design and analysis of computer experiments. Springer, New York

Shi CL, Tang BX (2020) Construction results for strong orthogonal arrays of strength three. Bernoulli 26:418–431

Sobol' IM (1967) On the distribution of points in a cube and the approximate evaluation of integrals. Zh Vychisl Mat Mat Fiz 7:784–802

Tsybakov AB (2009) Introduction to nonparametric estimation. Springer, New York

Wang Q, Kulkarni SR, Verdú S (2006) A nearest-neighbor approach to estimating divergence between continuous random vectors. In: IEEE international symposium on information theory

Worley BA (1987) Deterministic uncertainty analysis. Technical Report ORNL-6428. Oak Ridge National Laboratories

Wu Y, Ghosal S (2008) Kullback Leibler property of kernel mixture priors in Bayesian density estimation. Electron J Stat 2:298–331