Model-Free Subsampling Method Based on Uniform Designs

Mei Zhang, Yongdao Zhou, Zheng Zhou, and Aijun Zhang, Member, IEEE

Abstract—Subsampling or subdata selection is a useful approach in large-scale statistical learning. Most existing studies focus on model-based subsampling methods which significantly depend on the model assumption. In this article, we consider the model-free subsampling strategy for generating subdata from the original full data. In order to measure the goodness of representation of a subdata with respect to the original data, we propose a criterion, generalized empirical \(F\)-discrepancy (GEFD), and study its theoretical properties in connection with the classical generalized \(\ell_2\)-discrepancy in the theory of uniform designs. These properties allow us to develop a kind of low-GEFD data-driven subsampling method based on the existing uniform designs. By simulation examples and a real case study, we show that the proposed subsampling method is superior to the random sampling method. Moreover, our method keeps robust under diverse model specifications while other popular model-based subsampling methods are under-performing. In practice, such a model-free property is more appealing than the model-based subsampling methods, where the latter may have poor performance when the model is misspecified, as demonstrated in our simulation studies. In addition, our method is orders of magnitude faster than other model-free subsampling methods, which makes it more applicable for subsampling of Big Data.

Index Terms—Empirical \(F\)-discrepancy, generalized \(\ell_2\)-discrepancy, koksma-hlawka inequality, model-free subsampling, reproducing kernel.

I. INTRODUCTION

SUBSAMPLING methods are useful for saving computation and storage source for large sample datasets, see [15], [21], [22], [25]. Technological advances have enabled an extraordinary speed in data generation and collection in many scientific fields and practices, such as in astronomy, economics, and industrial problems. However, the growth rate of the storage memory and the computational power is still far from sufficiently handling the explosion of modern data sets. The demand for extracting a small sample from a large amount of data arises routinely.

Assume \(\mathcal{X} = \{x_i, i = 1, \ldots, N\}\) is an observed large-scale data in \(\mathbb{R}^s\). Here \(N\) may be very large, and \(s\) is relatively small. We consider the problem of selecting \(n\) points from \(\mathcal{X}\) to form a subdata \(\mathcal{P}\) that preserves the distribution of \(\mathcal{X}\). Most of existing works studied randomized algorithms based on weighted or importance sampling, see [1], [13], [15], [21], [25]. Among others, Wang et al. [21] proposed to perform subsampling for large-scale logistic regression, with the optimal subsampling probabilities determined by minimizing the asymptotic mean squared error of the target subsample-estimator (mMSE or mVc) given the full data. The mMSE and mVc are respectively based on A- and L-optimality criteria in the theory of optimal experimental design. Furthermore, Wang et al. [22] proposed an information-based optimal subdata selection (IBOSS) for the linear regression of the Big Data. The basic idea of IBOSS is to select the most informative data points in a deterministic manner according to the D-optimality criterion in optimal experimental design. Both methods in [21], [22] depend on explicit model assumptions in the simplest forms. As the underlying model becomes complex (instead of linear or logistic regression), these methods would no longer keep the optimal performance.

In statistical learning, we usually do not have prior knowledge about the underlying model, and it is often the case the working model can be misspecified. It is therefore meaningful to develop subsampling methods that are robust to model specification; and model-free subsampling methods emerge as an interesting area, see [4], [16], [17]. However, such model-free subsampling methods are mostly time-consuming, which makes them unsuitable for handling Big Data or stream data problems [32].

Our research aim is an efficient model-free subsampling method that is competitive no matter whether the underlying model is correctly specified. For statistical learning from large-scale data, the working model usually takes the complex form, either parametric or nonparametric. Such working model is usually misspecified and does not match the ground truth. In this case, a successful model-free subsampling method would be particularly useful. Among randomized subsampling techniques, the uniform random sampling (URS) is the simplest strategy and it is often regarded as the baseline method for developing other model-free subsampling methods. It is known that the
subdata selected by URS can preserve the distribution of the full data. Compared to the URS method, some quasi-Monte Carlo methods have better performances for representing uniform or general distribution. The uniform design [6], [8] is a popular method that possesses the model-robust property, as it is widely used in numerical integration, computer experiments, statistical simulations and other statistical areas. To measure the uniformity of a point set in an s-dimensional unit hypercube, there exist many forms of discrepancy criteria, such as the star discrepancy and the generalized $\ell_2$-discrepancy, wrap-around $\ell_2$-discrepancy and mixture discrepancy, see [9], [10], [20], [33]. More details can refer to Fang et al. [6]. Moreover, for a general distribution $F(x)$ in $\mathbb{R}^s$, Fang and Wang [8] proposed the concept of $F$-discrepancy for measuring the goodness of representation of a point set with respect to $F(x)$. Denote the empirical cumulative distribution function (ECDF) of the point set $P$ by $F_P(x)$, the $F$-discrepancy between $F(x)$ and $P$ is defined by $D_F(P) = \sup_{x \in \mathbb{R}^s} |F_P(x) - F(x)|$. The smaller the value of $F$-discrepancy, the better the point set represents the general distribution $F(x)$. However, such a criterion doesn’t have a closed-form expression and it is hard to be calculated by numerical methods when $s$ or the sample size of $P$ is large. Except for these discrepancies involved in the experimental design, there are many other measures used to examine the distribution distances, for example, the Kullback-Leibler (KL) divergence, $\alpha$-divergence, Hellinger divergence, and $f$-divergence, see [30], [31]. Here we aim to develop an efficient model-free subsampling method to obtain a subdata having good representation with respect to the full data, i.e., the ECDFs of the obtained subdata and the full data have low discrepancy according to some criteria.

Motivated by the model-robust property of uniform designs and the idea of $F$-discrepancy, we propose a data-driven subsampling method based on a generalized empirical $F$-discrepancy (GEFD). The main idea is to utilize the uniform design on the unit hypercube and transform it to the observational data space. The proposed GEFD criterion is defined as the $\ell_2$-norm of the difference between empirical distributions of the small data and the Big Data in the observational space. Under the joint independence assumption, the GEFD criterion can be translated to the unit hypercube upon the suitable transformations, and then it can be calculated by an explicit expression easily. We study the asymptotic equivalence of such transformation, and then develop the subsampling method based on the existing uniform designs in the literature. Such a uniform design-based subsampling method only depends on the data, but not the model. Therefore, we call it data-driven subsampling (DDS). DDS is demonstrated through several numerical examples to enjoy the model-robust property even when the working model is misspecified and has the average time complexity $O(sN(s + \log N))$, where $N, s$ are the number of samples and dimensions of the original dataset, respectively. To illustrate the superiority of DDS, we compare the performance of DDS, URS, IBOSS and some popular model-free subsampling methods, such as kernel herding [4] and support points [16]. As expected, DDS keeps efficient and robust under diverse model specifications even when the working model is misspecified, while other popular subsampling methods are under-performing.

The remainder of this article is organized as follows. Section II proposes the new GEFD criterion in order to measure the goodness of presentation for a small data with respect to the full data. Section III establishes the asymptotic equivalence between the GEFD on the observational space and the generalized $\ell_2$-discrepancy on the unit hypercube. The corresponding empirical version of Koskma-Hlawka inequality is also derived. Section IV introduces the detailed algorithm of DDS. Section V compares the performance of DDS and other subsampling methods by the simulation and real case study to show the model-free property and the efficiency of DDS. All technical proofs and some additional numerical results are provided in the Supplementary Materials, available online.

II. GENERALIZED EMPIRICAL $F$-DISCREPANCY

For a given data $\mathcal{X} = \{x_i, i = 1, \ldots, N\} \subset \mathbb{R}^s$ with large $N$, denote its ECDF as $F_X$. To find a small data to represent $\mathcal{X}$, a natural idea is to find a subdata $P \subseteq \mathcal{X}$ that has low discrepancy with respect to the ECDF $F_X$. Following the $F$-discrepancy by Fang and Wang [8] and the generalized $\ell_2$-discrepancy by Hickernell [9], we define the generalized empirical $F$-discrepancy (GEFD) in this section. It will be shown that such a discrepancy has an analytic expression. Let $C^s = [0, 1]^s$ be the $s$-dimensional unit hypercube, and $\mathbb{K}(\cdot, \cdot)$ be a real-valued kernel function defined on $C^s \times C^s$, satisfying i) symmetric: $\mathbb{K}(u, v) = \mathbb{K}(v, u)$ for any $u, v \in C^s$; ii) non-negative definite: $\sum_{j=1}^n a_j \mathbb{K}(u_j, u_k) \geq 0$ for any $n > 0, a_i \in \mathbb{R}$ and $u_i \in C^s$. Denote the space of real-valued functions on $C^s$ with kernel function $\mathbb{K}$ by $W_\mathbb{K} = \{F : C^s \to \mathbb{R} : \mathbb{K}(u, v)dF(u)dF(v) < \infty\}$. Then the space $(W_\mathbb{K}, \langle \cdot, \cdot \rangle_\mathbb{K})$ is a Hilbert space, where $\langle \cdot, \cdot \rangle_\mathbb{K}$ denotes the inner product with formula $\langle F, G \rangle_\mathbb{K} = \int_{C^s} \mathbb{K}(u, v)dF(u)dG(v)$. Such a $\mathbb{K}$ is a reproducing kernel satisfying that $\mathbb{K}(\cdot, u) \in W_\mathbb{K}$ and $F(u) = \langle F, \mathbb{K}(\cdot, u) \rangle_\mathbb{K}$ for any $u \in C^s$ and $F \in W_\mathbb{K}$. For a point set $D = \{\zeta_1, \ldots, \zeta_n\} \subset C^s$, the reproducing kernel $\mathbb{K}$ induces the squared generalized $\ell_2$-discrepancy of $D$ with respect to the uniform distribution $F_u$ on $C^s$ as follows,

$$D^2(D; F_u, \mathbb{K}) = \int_{C^s} \mathbb{K}(u, v)d(F_D - F_u)(u)d(F_D - F_u)(v) = \int_{C^s} \mathbb{K}(u, v)dudv - \frac{2}{n} \sum_{i=1}^n \int_{C^s} \mathbb{K}(u, \zeta_k)dudv + \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \mathbb{K}(\zeta_i, \zeta_k).$$

By taking different kernel functions $\mathbb{K}$, we can obtain different kinds of generalized discrepancies such as the widely used centered $\ell_2$-discrepancy, wrap-around $\ell_2$-discrepancy and mixture discrepancy, whose kernel functions are defined as

$$\mathbb{K}^C(u, v) = \prod_{j=1}^s \left[1 + \frac{1}{2} |u_j - v_j| - \frac{1}{2} \frac{v_j - 1}{2} - \frac{1}{2} |u_j - v_j| \right],$$

$$\mathbb{K}^W(u, v) = \prod_{j=1}^s \left[\frac{3}{2} - |u_j - v_j| + |u_j - v_j|^2 \right].$$
\[ K^M(u, v) = \prod_{j=1}^{s} \left[ \frac{15}{8} - \frac{1}{4}u_j - \frac{1}{2} - \frac{1}{4}v_j - \frac{1}{2} - \frac{3}{4}|u_j - v_j| \right. \]
\[ + \left. \frac{1}{2}|u_j - v_j|^2 \right], \tag{2} \]

respectively. Zhou et al. [33] showed that the mixture discrepancy is a better choice for measuring the uniformity of point sets in \( C^s \). Let \( \overline{K} : \mathbb{R}^s \times \mathbb{R}^s \to \mathbb{R} \) be a kernel function defined on \( \mathbb{R}^s \times \mathbb{R}^s \). We are concerned with the representation of a small data \( P \) with respect to the full data \( \mathcal{X} \), which are associated with the ECDFs \( F_P \) and \( F_{\mathcal{X}} \), respectively. Similar to the generalized \( \ell_2 \)-discrepancy with respect to the uniform distribution \( F_u \) on \( C^s \), we consider a norm of the difference between \( F_P \) and \( F_{\mathcal{X}} \),

\[ \|F_P - F_{\mathcal{X}}\|_{\overline{K}} = \left[ \int_{\mathbb{R}^{2s}} \overline{K}(x, z) d(F_{\mathcal{X}}(x) - F_P(x))d(F_{\mathcal{X}}(z) - F_P(z)) \right]^{1/2}. \]

Moreover, we consider the transformation \( T_X : \mathbb{R}^s \to C^s \) of the form

\[ T_X(x) = (F_{X(1)}(x_1), \ldots, F_{X(s)}(x_s))^T, \tag{3} \]

where \( F_{X(j)} \) is the marginal ECDF of the \( j \)th component of \( \mathcal{X} \), \( j = 1, \ldots, s \). Clearly \( T_X \) in (3) can translate the data \( \mathcal{X} \) in \( \mathbb{R}^s \) to the unit hypercube \( C^s \) component by component. Upon such transformation, we consider the kernel function \( \overline{K}(x, z) = K(T_X(x), T_X(z)) \), where \( K(\cdot, \cdot) \) is a reproducing kernel function on \( C^s \times C^s \), and define the GEFD of \( P \) with respect to \( \mathcal{X} \) as follows.

**Definition 1**: Given a data \( \mathcal{X} = \{x_1, \ldots, x_N\} \subset \mathbb{R}^s \), the squared generalized empirical \( F \)-discrepancy for a point set \( P = \{x_1, \ldots, x_n\} \subset \mathbb{R}^s \) is defined by

\[ D^2(P; \mathcal{X}, K) = \int_{\mathbb{R}^{2s}} \overline{K}(T_X(x), T_X(z))d(F_{\mathcal{X}}(x) - F_P(x)) \]
\[ \times d(F_{\mathcal{X}}(z) - F_P(z)), \tag{4} \]

where \( \overline{K}(\cdot, \cdot) \) is a reproducing kernel function on \( C^s \times C^s \) and \( T_X(\cdot) \) is defined in (3).

By the definition of GEFD, the smaller the GEFD of the point set \( P \) with respect to \( \mathcal{X} \), the better it represents \( \mathcal{X} \). From (4), the GEFD can be equivalently expressed as

\[ D^2(P; \mathcal{X}, K) = \frac{1}{N^2} \sum_{i, k=1}^{n} \overline{K}(x_i, x_k) - \frac{2}{Nn} \sum_{i=1}^{N} \sum_{k=1}^{n} \overline{K}(x_i, x_k) \]
\[ + \frac{1}{n^2} \sum_{i, k=1}^{n} \overline{K}(x_i, x_k), \tag{5} \]

in which \( \overline{K}(x, z) = K(T_X(x), T_X(z)) \). Therefore, one can evaluate the GEFD criterion easily without calculating \( F_X \) or \( F_P \) because the calculation of \( T_X \) only depends on the marginal ECDFs of the \( s \) components of \( \mathcal{X} \).

Let us give a toy example to illustrate that the proposed GEFD is reasonable to measure the representation of a small data with respect to the original Big Data. Suppose the original dataset \( \mathcal{X}^{(1)} = \{x^{(1)}_i, i = 1, \ldots, N\} \) are generated from a binormal distribution with independent components, shown as the background grey points in Fig. 1 (upper panel). We obtain \( P^{(1)}_1 = \{x^{(1)}_k, k = 1, \ldots, n\} \subset \mathcal{X}^{(1)} \) by the URS method, and \( P^{(2)}_2 = \{x^{(1)}_k, k = 1, \ldots, n\} \subset \mathcal{X}^{(1)} \) by the subsampling method in Section IV. Let us choose the kernel function \( K^M \) of mixture discrepancy in (2). According to the analytical expression (5), it can be obtained that

\[ D^2(P^{(1)}_1; \mathcal{X}^{(1)}, K^M) = 1.1212 \times 10^{-2} \]
\[ > 4.0424 \times 10^{-4} = D^2(P^{(1)}_2; \mathcal{X}^{(1)}, K^M). \]

Thus \( P^{(2)}_2 \) is better than \( P^{(1)}_1 \) by the criterion of GEFD. This comparison is in accordance with the observed fact in Fig. 1 (upper panel), where \( P^{(2)}_2 \) represents the full data \( \mathcal{X} \) better than \( P^{(1)}_1 \) does. Meanwhile, it is worthy mentioning that the GEFD by Definition 1 is also well-defined when the components of \( \mathcal{X} \) are correlated, although the transformation \( T_X \) in (3) translates each coordinate of \( \mathcal{X} \) into \([0,1]\) independently. The right hand side of (4) is a norm of the function \( F_X - F_P \), where the ECDF \( F_X \) is general enough to cover any joint distributions of \( \mathcal{X} \). For illustration, we give another example where the two components of \( \mathcal{X}^{(2)} \) are correlated, see Fig. 1 (lower panel). We obtain two different subsamples \( P^{(2)}_1 \) and \( P^{(2)}_2 \) respectively by the URS method and the subsampling method in Section IV, and calculate their GEFD criteria,

\[ D^2(P^{(2)}_1; \mathcal{X}^{(2)}, K^M) = 7.7165 \times 10^{-3} \]
\[ > 1.2544 \times 10^{-3} = D^2(P^{(2)}_2; \mathcal{X}^{(2)}, K^M). \]

It is implied that \( P^{(2)}_2 \) is better than \( P^{(2)}_1 \) for representing the full data, which agrees with Fig. 1 (lower panel). Therefore, the GEFD is a reasonable goodness measure for a small subdata in representing the original full data.

Note that the criterion-based optimization approach always leads to an NP-hard problem, such as the method of support
points [16], which is time-consuming to solve and avoidable to run into local optimum. Solving the optimization problem directly by (5) is also NP-hard. Alternatively, we establish the connections between GEFD and the uniformity criterion for constructing uniform designs [6], [7], [8], then convert the original GEFD-optimization problem to direct utilization of existing uniform design tables. This eventually leads to the proposed DDS method discussed in detail in the following sections.

III. PROPERTIES OF THE GEFD

When \( N \to \infty \), the ECDF \( F_X(x) \) of \( X \) converges to the CDF \( F(x) \) for \( x \in \mathbb{R}^d \), and the marginal ECDF \( F_{X(j)}(x_j) \) of the \( j \)-th component converges to the \( j \)-th marginal CDF of \( F(x) \) for \( j = 1, \ldots, s \). Under the joint independence condition, we can derive the asymptotic equivalence between the GEFD defined in (4) and the generalized \( \ell_2 \)-discrepancy in (1) upon the transformation \( T_X \) in (3).

Theorem 1: Given a reference data \( \mathcal{X} \subset \mathbb{R}^s \) satisfying the joint independence

\[
F_X(x) = \prod_{j=1}^s F_{X(j)}(x_j), \forall x = (x_1, \ldots, x_s)^T \in \mathbb{R}^s,
\]

is asymptotically optimal (as \( N \to \infty \)) with respect to the GEFD using the kernel function \( \mathbb{K}^C \), \( \mathbb{K}^W \) or \( \mathbb{K}^M \) defined in (2). Here \( F_X \) is the ECDF of \( X \).

Hickernell [9] pointed out an important property of the generalized \( \ell_2 \)-discrepancy, known as the famous Koksma-Hlawka Inequality. Let \( D = \{ \xi_k, k = 1, \ldots, n \} \) be a set of points on \( C^n \), and \( f \) be a function on \( C^n \), then the upper bound of the difference between the integral of \( f \) over \( C^n \) and the averaged value of \( f \) among \( D \) is given by

\[
\int_{C^n} f(u) du - \frac{1}{n} \sum_{k=1}^n f(\xi_k) \leq D(D; F_u, \mathbb{K}) V_2(f, \mathbb{K}), \tag{7}
\]

where \( V_2(f, \mathbb{K}) \) is the generalized \( \ell_2 \)-variation defined in [9]. Such Koksma-Hlawka inequality is an important result in numerical integration and quasi-Monte Carlo methods. Given the function \( f \) and a reproducing kernel \( \mathbb{K}, V_2(f, \mathbb{K}) \) is a fixed quantity. When \( V_2(f, \mathbb{K}) \) is bounded in \( C^n \), the lower the generalized \( \ell_2 \)-discrepancy \( D(D; F_u, \mathbb{K}) \), the smaller the upper bound in (7).

Now consider the difference between the averaged values of \( f \) under two data sets. Given an \( N \)-size data set \( \mathcal{E} = \{ u_i, i = 1, \ldots, N \} \subset C^n \), an \( n \)-size data set \( \mathcal{D} = \{ \xi_k, k = 1, \ldots, n \} \subset C^n \), and a reproducing kernel \( \mathbb{K} \) on \( C^n \times C^n \), write

\[
D_{\mathbb{K}}^2(\mathcal{E}, \mathcal{D}) = \int_{C^n} \mathbb{K}(u, v) d(F_{\mathcal{E}}(u) - F_{\mathcal{D}}(u))d(F_{\mathcal{E}}(v) - F_{\mathcal{D}}(v)) \tag{8}
\]

where \( D_{\mathbb{K}}^2(\mathcal{E}, \mathcal{D}) \) is a indeed a squared norm \( ||F_{\mathcal{E}} - F_{\mathcal{D}}||^2 \) induced by \( \mathbb{K} \). When the data set \( \mathcal{E} \) contains a large sample of points generated from the uniform distribution on \( C^n \), \( D_{\mathbb{K}}^2(\mathcal{E}, \mathcal{D}) \) becomes the empirical form of the squared generalized \( \ell_2 \)-discrepancy. In this view, \( D_{\mathbb{K}}^2(\mathcal{E}, \mathcal{D}) \) assesses the difference between \( \mathcal{E} \) and \( \mathcal{D} \) in terms of their empirical CDFs. Then we obtain the following lemma of the empirical form of (7) on the unit hypercube \( C^n \).

Lemma 1 (Empirical Koksma-Hlawka Inequality on \( C^n \)): Given two point sets \( \mathcal{E}, \mathcal{D} \subset C^n \), a reproducing kernel \( \mathbb{K} \) on \( C^n \times C^n \), and the function \( f \) with the bounded \( \ell_2 \)-variation \( V_2(f, \mathbb{K}) \) on \( C^n \), we have

\[
\left| \frac{1}{N} \sum_{u \in \mathcal{E}} f(u) - \frac{1}{n} \sum_{\xi \in \mathcal{D}} f(\xi) \right| \leq D_{\mathbb{K}}(\mathcal{E}, \mathcal{D}) V_2(f, \mathbb{K}),
\]

where \( D_{\mathbb{K}}(\mathcal{E}, \mathcal{D}) \) is given by (8).

Note that for any two data sets \( \mathcal{X} = \{ x_i, i = 1, \ldots, N \} \subset \mathbb{R}^s \), and \( \mathcal{P} = \{ \xi_k, k = 1, \ldots, n \} \subset \mathbb{R}^s \), through the transformation \( T_X \) in (3), it is easy to see that \( T_X(\mathcal{X}) \subset C^n \) and \( T_X(\mathcal{P}) \subset C^n \). According to the analytical expression of GEFD in (5), we have that \( D^2(\mathcal{P}; \mathcal{X}, \mathbb{K}) = D_{\mathbb{K}}^2(T_X(\mathcal{X}), T_X(\mathcal{P})) \). By combining Lemma 1 and the equivalence between \( D(\mathcal{P}; \mathcal{X}, \mathbb{K}) \)
and $D_k(T_X(X), T_X(P))$, we can deduce the empirical Koksma-Hlawka inequality in terms of GEFD defined on $\mathbb{R}^s$.

**Theorem 2 (Empirical Koksma-Hlawka Inequality on $\mathbb{R}^s$):**

Given a reference data $X \subset \mathbb{R}^s$, $\mathbb{K}(u, v) = \prod_{j=1}^s K(u_j, v_j)$ is a reproducing kernel function on $C^s \times C^s$ and $f$ has a bounded $\ell_2$-variation $V_2(f, \mathbb{K})$ on $C^s$, then for any point set $P \subset \mathbb{R}^s$,

$$
\left| \frac{1}{N} \sum_{x \in X} f(T_X(x)) - \frac{1}{n} \sum_{z \in P} f(T_X(z)) \right| \leq D(P; X, \mathbb{K}) V_2(f, \mathbb{K}),
$$

where $T_X$ takes the form of (3), and $D(P; X, \mathbb{K})$ is given by Definition 1.

It is worth noting that in Theorem 2 the reference data $X$ on $\mathbb{R}^s$ is not required to satisfy the joint independence condition as in Theorem 1.

Theorem 2 provides another rationale for the proposed GEFD criterion used for measuring the closeness of $P$ to $X$.

**IV. DATA-DRIVEN SUBSAMPLING**

Given a reference data $X \subset \mathbb{R}^s$, we want to find a small data $P \subset \mathbb{R}^s$ such that $P$ has a good representation of $X$. As discussed in the previous section, the goodness of representation can be measured by the GEFD criterion. Theorem 1 translates the GEFD of $P \subset \mathbb{R}^s$ to the generalized $\ell_2$-discrepancy of $T_X(P)$ in $C^s$, subject to an approximation error of order $1/N^s$. Therefore, if $X$ satisfies the joint independence assumption (6), an $n$-point design $D$ in $C^s$ with low generalized $\ell_2$-discrepancy could lead to a point set, $P$, with low GEFD by the inverse transformation of $T_X$, i.e., $P = T_X^{-1}(D)$. Here the low-discrepancy design $D$ can be obtained by the following methods. The first method is finding designs straightforwardly from the library of uniform designs given by [7], which does not require any calculations. If there is no $n \times s$ design in the library, the other two methods can be used to construct such a design. Let $k(n) = \phi(n)/2 + 1$, where $\phi(\cdot)$ is the Euler function. If $s < k(n + 1)$, the leave-one-out good lattice point method [6] can be used to construct a nearly uniform design. For the other cases, we can use the R package, UniDOE proposed by [29], to search for a nearly uniform design. Actually, the UniDOE package can be used to search for a nearly uniform design for arbitrary $n, s$. However, the UniDOE package is a little slower than the leave-one-out good lattice point method in general, then we recommend to use the former only when $s \geq k(n + 1)$.

The joint independence assumption in (6) may be not satisfied for a real data set $X \subset \mathbb{R}^s$. As a common practice, we can apply the statistical procedures such as PCA to transform the reference data to a latent space, $Z$. PCA is based on the singular value decomposition (SVD), to convert the data to have linearly uncorrelated coordinates. Based on the principal scores on the latent space, one can use the inversion method to find a small representation $Q_P$ with low GEFD. Finally, such $Q_P$ can be converted back to the original space as the desired small data representation of $X$ on $\mathbb{R}^s$.

The above procedure can be called the rotation-inversion construction, and it can be described more precisely in the following three steps: i) performing SVD for $X$ to obtain the rotation $V$, the singular-valued matrix $\Lambda$, and the rotated data $Z$; ii) constructing the data-driven space-filling design $Q_P = \{\eta_k, k = 1, \ldots, n\}$ in the space of $Z$ by performing the $T_Z^{-1}$ transformation on each $\zeta_k \in D$ as follows,

$$
\eta_k = T_Z^{-1}(\zeta_k) = \left(\int_{Z_1}^{-1}(\zeta_{k1}), \ldots, \int_{Z_s}^{-1}(\zeta_{ks})\right)^T;
$$

(iii) generating the point set $P$ by $\xi_k = V \Lambda \eta_k$ for each $k = 1, \ldots, n$. Such a procedure is computationally efficient for a large-scale data in a low dimensional space. For an illustration, we apply the rotation-inversion construction to a two-dimensional reference data $X$ and obtain the new sampled points $P$, shown in Fig. 2(a) and (b). The original data $X$ is simulated from a truncated binormal distribution with correlation, same as that in Fig. 1. It can be found that such a small data $P$ (plotted as circles) represents the original data $X$ (plotted as dots) quite well.

For the subsampling purpose, it is required that the obtained small data $P$ should be a subset of the original data $X$. Since the rotated data $Z$ is only nearly independent, there exist some points in $Q_P$ not belonging to $Z$. For each of such data point, we suggest to find its nearest neighbor in $Z$ as a replacement. In Fig. 2(c), the points of $Q_P$ are shown as the circle points, and their nearest points from $Z$ are shown as the star points. To find the nearest neighbor of each point in $Q_P \backslash Z$, KD-tree is the conventional methods.

We summarize the above procedure by Algorithm 1. As an example, it can output the subsampled points shown in Fig. 2(c). For each design point (circle points) in the rotated space, a corresponding sample (star annotation) is found from $Z$, which is closest the ideal design point. It is anticipated that such kind of subdata $P^1$ is similar to the designed $P$ and has low GEFD value.

DDS is a deterministic method, and its time complexity is analyzed below. In Step 1, the complexity of SVD $O(s^2 N)$. In Step 2, given $i, k$, calculating $F_{Z(i)}^{-1}(\zeta_{ki})$ is equivalent to calculating the $\zeta_{ki}$-quantile of $Z(i)$. Quick sort [24] and BFPRT [2] can be used to obtain the quantile of a large number of samples efficiently. The worst case and average case complexity of computing $n$ quantiles of $N$ numbers by quick sorting are $O(N^2)$ and $O(N \log N)$, respectively. The complexity of BFPRT for the
same task is $O(nN)$. To avoid the worst case of quick sorting, we can implement quick sorting and BFPT in parallel and end them when either of them is completed. Then the worst case and the average case complexity of Step 2 are $O(snN)$ and $O(sN \log N)$, respectively. In Step 3, the complexity of building the KD-tree is $O(sN \log N)$. In Step 4, the worst case and average case complexity of searching a nearest neighbor from KD-tree is $O(N^{1−1/s})$ and $O(\log N)$, respectively. Then, the overall complexity of DDS in the worst case is $O(s^2 N) + O(snN) + O(sN \log N) + O(nN^{1−1/s}) = O(snN)$.

and the average case complexity is

$$O(s^2 N) + O(snN) + O(n \log N) = O\{sn(s + \log N)\}.$$ 

Note that the worst case of quick sort rarely happens when the sample size $N$ is relatively large. As a result, we recommend using quick sort without parallel computing in Step 2 to save memory in application.

To further certify the performance of DDS, we consider the case of $n = 2000$, $s = 10$, where the full data set $X$ is randomly generated from the $s$-dimensional uniform distribution with independent components each on $[0, 1]$. Let the data size be $N = 10^4, 10^5, 10^6$ and $10^7$ increasingly. We compared DDS with three subsampling methods, where IBOSS is model-based and the other two methods are model-free. Table I shows the computing time for each $N$ carried out on a server Intel(R) Xeon(R) with CPU E5-2650 v4 and 2.20 GHz. It can be seen that DDS is several orders of magnitude faster than the other two model-free subsampling methods, which makes it more applicable.

V. NUMERICAL EXAMPLES

In this section, we show the model-free property of the subsampling method through both simulation data and a real case study. When using Algorithm 1, we only select the leading components that give no less than 85% of the variance explained among all possible components. The operation yields a lower dimensional rotated pace, and reduces the computational burden due to non-important components.

For the given subsample size $n$, and the determined dimension of the rotated pace $s$, the $n$-run $s$-factor uniform design in Algorithm 1 is constructed by the leave-one-out good lattice point method with a power generator. In this construction method, the generator vector is given by a positive integer $\alpha$ which satisfies that the great common divisor of $n + 1$ and $\alpha$ is one and $\alpha, \alpha^2, \ldots, \alpha^s$ are distinct. Then for each $j = 1, \ldots, s − 1$, the remainders after dividing $\alpha^j, 2\alpha^j, \ldots, s\alpha^j$ by $n + 1$ are $n$ distinct integers. Denote this generator vector by $\gamma_0$, and it has a power form $(\alpha^0, \alpha^1, \ldots, \alpha^{s−1})$. Then the corresponding design generated by $\gamma_0$ is $D^{(\alpha)} = \{\zeta^{(\alpha)}_i, i = 1, \ldots, n\}$, where $\zeta_i^{(\alpha)} = \text{mod}(\gamma_0, n + 1)/n − 1/2n)$ with mod(·, ·) denoted as the modulus operation. Different values of $\alpha$ lead to different designs. We use the mixture discrepancy as the uniformity criterion. Finally, the uniform design $D$ constructed by the leave-one-out good lattice point method with a power generator is the one which owes the smallest mixture discrepancy value among all possible $D^{(\alpha)}$. This construction procedure provides a fast and effective method for selecting the $n$ design points from $n^s$ lattice points. The designs constructed by the leave-one-out good lattice point method with a power generator are indeed approximate uniform designs. For comparison, another simplest model-free subsampling method URS is also implemented. To make the different subsampling methods comparable, for each given subsample size $n$, the implementation of URS is executed for $R$ repetitions to obtain the mean, median and bounds of the results of the URS’s subsample. Moreover, we use a random shift of $\varepsilon \in C^s$ for all design points in $D = \{\zeta_k, k = 1, \ldots, n\}$, to obtain another approximate uniform design $D_\varepsilon = \{\text{mod}(\zeta_k + \varepsilon, 1), k = 1, \ldots, n\}$. Performing this random shift for $R$ repetitions results in $R$ approximate uniform designs. Then we also obtain the mean, median and bounds of the results corresponding to DDS.

A. Simulation Studies

In this subsection, we utilize the data-driven subsampling method for both classification and regression problems. For each kind of model, we compare the prediction property of the trained models based on the subdata sets by different subsampling strategies. It will be shown that the proposed DDS method possesses the model-free property, and outperforms URS method in various settings.
In each simulation, we generate two \( N \)-size data sets respectively as the full data \( \mathcal{X}_{\text{full}} \) and the test data \( \mathcal{X}_{\text{test}} \) through the same generation way as well as the corresponding binary class label (and response in regression problem) \( Y_{\text{full}} \) and \( Y_{\text{test}} \). To measure the prediction performance of a specified type of model trained upon the \( n \)-size subdata \( \mathcal{P} \) from \( \mathcal{X}_{\text{full}} \), we fit a corresponding model using \( \mathcal{P} \) to predict the class label (or response) for each point in \( \mathcal{X}_{\text{test}} \). Denote the fitted model by \( \hat{h}(\cdot) \). For each \( x \in \mathcal{X}_{\text{test}} \), the corresponding predicted class label (or response) is \( \hat{h}(x) \). Then a prediction error could be computed by \( \epsilon = \frac{1}{2} N \| \hat{h}(\mathcal{X}_{\text{test}}) - Y_{\text{test}} \|_2^2 \) which represents the misclassification rate for the classification problem with class labels being 0 or 1, and mean squared prediction error in the regression case, upon the test data \( \mathcal{X}_{\text{test}} \). Denote this criterion by \( \text{Err} \) and \( \text{MSPE} \) respectively for the classification and regression problem. The lower value of this criterion, the better the prediction performance.

(A) Working model is true.

For the classification simulation, let \( N = 10^4 \) and both the data sets \( \mathcal{X}_{\text{full}} \) and \( \mathcal{X}_{\text{test}} \) be generated from the multinormal distribution \( \mathcal{N}(0, \Sigma) \), where \( \Sigma_{ij} = 0.5^{(i \times j)} \), \( i, j = 1, \ldots, 7 \). The kind of data set is also used in Wang et al. [21].

The logistic regression model is a classical and widely used classification method for its simplicity and effectiveness. Consider the logistic regression as a classification model. Fortunately, if the underlying model is the logistic model, for example, the probability of the class label being 1 for a point \( x \) is \( h(x, \beta) = 1 / (1 + \exp(-x^T \beta)) \), where \( \beta \) is a \( 7 \times 1 \) vector of 0.5, the optimal subsampling methods (mMSE and mMVc) in Wang et al. [21] could be adopted. We consider the four different kinds of subsampling strategies: URS, mMVc, mMSE and DDS, to obtain the subdata with \( n = 400, 550, 750, 950, 1200 \) and 1500, respectively. The implementations of mMVc and mMSE are represented by a two-step algorithm. The first step obtains \( r_0 \) points randomly, and the other \( n - r_0 \) points are obtained in the second step based on their optimal subsampling probabilities. For this type of datasets, Wang et al. [21] suggested \( r_0 = 200 \) to well present the good performance of mMVc and mMSE. For each \( n \), every subsample strategy is executed 1000 times because of the randomness. Recall that, the randomness of DDS is reflected from the random shift for an initial constructed approximate uniform design.

Fig. 3 contains the misclassification rates of the fitted logistic regression models based on the subdata sets obtained by the four methods. For comparison, the fitted logistic regression model based on \( \mathcal{X}_{\text{full}} \) is also considered. For each subsample size \( n \), Fig. 3(a) presents the mean values of the 1000 results for the four subsampling methods which are denoted by URS\( \text{mean} \), mMVc\( \text{mean} \), mMSE\( \text{mean} \), and DDS\( \text{mean} \). Fig. 3(b) presents the median values and lower and upper bounds of the 1000 results corresponding to the four methods which are denoted by URS\( \text{median} \), mMVc\( \text{median} \), mMSE\( \text{median} \), DDS\( \text{median} \), and URS\( \text{bounds} \), mMVc\( \text{bounds} \), mMSE\( \text{bounds} \), DDS\( \text{bounds} \), respectively. In the version of mean and median, Fig. 3 illustrates that the mMVc, mMSE and DDS all outperform URS. The subsampling methods mMVc and mMSE are based on the logistic model, they have better performance than URS when the model is correct, as shown in Wang et al. [21]. The subdata sets obtained by DDS have better representation with respect to the original data than that by URS. Therefore, the better similarity of the original data helps to better fit the model. Fig. 3(b) also shows a narrower interval of DDS than URS. It means that the subdata from DDS is more robust than other methods. Moreover, the performance of DDS is as similar as that of mMVc and mMSE.

In this simulation, the fitted model is indeed the true model. We also calculate the mean squared prediction error of parameter estimations to compare the fitting effects of different methods. We could obtain that between the two model-free subsampling methods, DDS outperforms URS, especially when \( n \) is relatively small. Table II shows the mean values and median values of 1000 results for each method for the subsample size \( n = 400 \).

(B) Working model is misspecified.

We evaluate the performance of different subsampling strategies while the working model is misspecified. We simulate the real life borehole example of the flow rate of water through a borehole from an upper aquifer to a lower aquifer separated by an impermeable rock layer. This example was investigated by many authors such as Worley [23], Morris et al. [18], Ho and
Xu [11], and Fang et al. [7]. The response variable \( \mathcal{Y} \), the flow rate through the borehole in \( m^3/yr \), is determined by a complex nonlinear function as follows,

\[
\mathcal{Y} = \frac{2\pi T_u (H_u - H_l)}{\ln(r/r_w)} \left[ 1 + \frac{2LT_u}{\ln(r/r_w)} r_w^2 K_w^2 + \frac{T_u}{T_l} \right],
\]

(10)

where the 8 input variables with their usual input ranges are listed as follows: \( r_w \in [0.05, 0.15] \) means the radius of borehole (\( m \)); \( r \in [100, 50000] \) means the radius of influence (\( m \)); \( T_u \in [63070, 115600] \) means the transmissivity of upper aquifer (\( m^2/yr \)); \( T_l \in [63.1, 116] \) means the transmissivity of lower aquifer (\( m^2/yr \)); \( H_u \in [990, 1110] \) means the potentiometric head of upper aquifer (\( m \)); \( H_l \in [700, 820] \) means the potentiometric head of lower aquifer (\( m \)); \( L \in [1120, 1680] \) means the length of the borehole (\( m \)); \( K_w \in [9855, 12045] \) means the hydraulic conductivity of the borehole (\( m/yr \)). The distribution of \( r_w \) is the normal distribution \( N(0.10, 0.0161812^2) \), the distribution of \( r \) is the lognormal distribution \( \text{Lognormal}(7.71, 1.0056^2) \), and the distributions of other variables are all continuous uniform distribution on their corresponding domains.

To compare the performances of different subsampling methods for the large-scale data sets, let the size of the full data \( N = 10^8 \), and the size of subdata \( n = 50, 80, 150, 250 \) and 400, respectively. We generate the \( N \)-size 8-factor \( X_{\text{Full}} \) and \( X_{\text{Test}} \) for training and testing, respectively, and the corresponding responses \( \mathcal{Y}_{\text{Full}} \) and \( \mathcal{Y}_{\text{Test}} \) through (10). For the data (\( X_{\text{Full}}, \mathcal{Y}_{\text{Full}} \)), we use different subsampling methods to obtain the subdata sets and use different models to fit each subdata set. The test data (\( X_{\text{Test}}, \mathcal{Y}_{\text{Test}} \)) can be used to compare the performance of the different subsampling methods.

First, we compare URS, DDS, IBOSS, kernel herding (KH) and support points (SP) under the simple linear model. IBOSS proposed by Wang et al. [22] is a kind of optimal subsampling for linear regression model. KH proposed by Chen et al. [4] and SP proposed by Mak and Joseph [16] are two popular data-based subsampling methods. According to the generation method of \( X_{\text{Full}} \), the components of the data are mutually independent, so there do not need the rotation step in the process of DDS any more. For each subsample size \( n \), URS, DDS and SP are all executed 1000 times because of the randomness. IBOSS and KH are both executed one time. For comparison, we also fit the same model for the full data.

Fig. 4(a) and (b) show the MSPE values of the fitted linear regression model based on the full data and the subdata sets with different subsample sizes obtained by the five subsampling methods. To present the results more explicitly, the vertical axes in 4 are logarithmically transformed. For all subsampling methods except KH, the MSPE values of the fitted models based on the subdata sets are closer to that based on full data sets as the subsampling size increases. As shown in Fig. 4, the MSPE values of the DDS are much lower than that of URS especially when the subsample size \( n \) is relatively small, and are close to that of the full data no matter from the version of mean or

| Value type | Method |
|------------|--------|
|            | Full   | mVc | mMSE | URS | DDS |
| Mean       | 0.00076 | 0.02939 | 0.02967 | 0.04105 | 0.03903 |
| Median     | 0.00076 | 0.02643 | 0.02627 | 0.03675 | 0.03507 |

Table II: The Mean Values and Median Values of 1000 Results for Each Method for the Subsample Size \( n = 400 \)
median. Moreover, DDS performs better than KH and IBOSS. In addition, Table 4 provide in Appendix A, available online, shows for most subsample size $n$, the upper bounds of the MSPE values of DDS are even lower than the MSPE values of IBOSS, KH, the mean and median MSPE values of URS. Thus the prediction ability of the fitted linear models based on the subsamples by DDS significantly outperforms that of URS, KH and IBOSS. Moreover, under the MSPE criterion, the model-based IBOSS performs worst among the three subsampling methods, which may be caused by that the simple linear model is not enough to capture the relationship between the output $\mathcal{Y}$ and the 8 input variables. The MSPE values of SP is the lowest in this case because our original variables are generated by some simple distribution. SP is adept in mining such distribution from the data and using them to make prediction. If some nonlinear transformation are applied to the parameters, it will be difficult for SP to handle the data, which can be demonstrated by the following model.

To compare the robustness of the sampling methods, we consider generalized regression models. A careful study by Ho and Xu [11] suggested fitting $\log(\mathcal{Y})$ with 10 terms: $\log(r_w)$, $\log(r)$, $H_u$, $H_1$, $L$, $K_w$, $H_{10}$, $H_{11}$, $H_2^2$, $H_3^2$ and $L^2$. In the generalized linear regression model, there are only 6 significant variables: $r_w$, $r$, $H_u$, $H_1$, $L$ and $K_w$. Then the DDS is executed on these 6 components of the original full data $\mathcal{X}_{\text{full}}$. We transform the subsamples obtained by URS and DDS and conduct IBOSS, KH and SP based on the transformed full data sets $\mathcal{X}_{\text{full}}$ with above 10 components in models to obtain the corresponding subdata sets with 10 components.

Fig. 4(c) and (d) show the values of MSPE for the fitted regression model based on the transformed full data and subdata sets with different subsample size obtained by the five subsampling methods. It is obviously that the prediction performance of the regression model upon the transformed data set achieves the significant improvement compared with that in Fig. 4(a). DDS performs better than URS, IBOSS, KH for each subsample size $n$. The MSPE values of DDS is the lowest while the subsample size $n$ is relatively big. SP makes more accurate prediction than DDS while $n$ is relatively small. However, the MSPE value of SP arises as the subsampling size arises and it becomes lower than that of URS when $n \geq 150$, which shows SP failing to capture more information from the transformed data while the subsampling size arises. In contrast, the performance of DDS is consistently well.

From the performances of the DDS, URS, IBOSS, KH and SP under the linear regression and the generalized linear regression model, it is known that the model-based IBOSS performs worse than the three model-free subsampling methods URS, DDS and SP when the model does not fit the data very well. Therefore, model-free subsampling methods make a good presentation of the full data, which derives the benefit for the modeling procedure, especially for the cases that the true model is unknown. DDS is the most robust model-free subsampling method which performs well for both two models and different subsample sizes $n$.

**B. Real Case Study**

In this subsection, we use the DDS for a real case study of the physicochemical properties of protein tertiary structure data set from the UCI machine learning repository (Dheeru and Taniskidou [5]). It contains 45,730 samples with 9 continuous attributes and 1 response variable.

First, we get rid of some outliers which results 45,253 samples and then rescale the data as the original data to use. Different from the simulation example of the borehole experiment, there is no guidance on the model for this real data. We consider the two models, linear regression model and Gaussian process regression model for the subdata by DDS, URS, IBOSS, KH and SP. The MSPE values of the 5-fold cross-validation are calculated to compare the performance of different subsampling strategies. In the 5-fold cross-validation, there are 5 training data sets $\mathcal{X}_{\text{Train},1}, \ldots, \mathcal{X}_{\text{Train},5}$ and 5 test data sets $\mathcal{X}_{\text{Test},1}, \ldots, \mathcal{X}_{\text{Test},5}$. For each $\mathcal{X}_{\text{Train},t}$, in the process of DDS, we use the first two dominated components which may explain more than 85% of total variation in $\mathcal{X}_{\text{Train},t}$ by the principal component analysis. So the dimension of the rotated space is reduced to 2. Here we select some numbers from the leave-one-out Fibonacci sequence as the subsample size $n$ for the good uniformity of the corresponding 2-dimensional designs on $C^2$. The detail of the Fibonacci sequence can be seen in Fang et al. [6]. For an illustration, we take sample size $n = 986$ and perform DDS to one of the training data sets. The rotated data and rotated subsamples are shown in

| $n$ | DDS | IBOSS | KH | SP |
|-----|-----|-------|----|----|
| 33  | 0.0658 | 0.0191 | 10.5168 | 6.1491 |
| 54  | 0.0648 | 0.0197 | 10.5989 | 8.3069 |
| 88  | 0.0648 | 0.0192 | 10.7472 | 11.0574 |
| 143 | 0.0648 | 0.0195 | 10.7897 | 16.8222 |
| 232 | 0.0648 | 0.0193 | 10.9741 | 26.4306 |
| 376 | 0.0658 | 0.0197 | 11.1478 | 41.5587 |

Fig. 5. The subsamples selected by DDS in the rotated space of the UCI protein structure dataset.
ZHANG et al.: MODEL-FREE SUBSAMPLING METHOD BASED ON UNIFORM DESIGNS

Fig. 6. The MSPE values of the fitted linear regression model and Gaussian process regression model established on different subsample strategies for the real case dataset.

To illustrate the property of the subsamples of DDS for the original data, we compare the models fitted by the subsamples obtained by different methods. For each fitted model and subsample size \( n \), the subsampling of URS, DDS and SP are repeated 100 times but no repetition for deterministic subsampling method IBOSS and KH. Table III shows the median computing time for each subsample size \( n \) and each subsampling method except URS. It can be seen that DDS is orders of magnitude faster than the other two model-free subsampling methods, Kernel herding and support points. Because of the simplicity of linear regression model, fitting such the model based on the training data is also considered. Fig. 6(a) and (b) present the MSPE values of the fitted linear regression models for the subdata obtained by different subsampling methods. To present the results more explicitly, the vertical axises in Fig. 6 are logarithmically transformed. The MSPE values of DDS are lower than that of URS from the versions of both mean and median. What’s more, the performance of DDS is better than IBOSS and KH. By Fig. 8(a) provided in Appendix A, available online, the upper bounds of the MSPE values of DDS are lower than that of IBOSS and KH for relatively large subsample size \( n \). SP leads to lower MSPE values than DDS for \( n \geq 232 \), however the performance of SP is the worst while \( n \) is relatively small, which is not robust for the subsample size.

For the Gaussian process regression model, training a model based on the training data in each fold is omitted because of its high complexity. For the subsampling methods, we set the same basis function, kernel function and hyper-parameter optimization method for each subdata obtained by each subsampling method. The corresponding MSPE values are shown in Fig. 6(c) and (d). Compared to the result of the simple linear regression model, the performance of IBOSS in this high complexity model even becomes worse. URS, DDS and SP have similar performance when the subsample size \( n \) is relatively large. However, DDS performs slightly better than URS when \( n \) is relatively small while other methods are significantly worse than URS.

These results of the real case study confirm that, if there is no a priori knowledge about the true model, model-free subsampling methods are more appropriate. Moreover, DDS performs more robustly than other model-free subsampling methods for different model specifications and subsample sizes.

VI. CONCLUSION

This article provides a novel model-free subsampling method based on the use of uniform designs. The goodness of subdata with respect to the original data is measured by the generalized empirical \( F \)-discrepancy (GEFD), which is asymptotically equivalent to the classical generalized \( \ell_2 \)-discrepancy in the theory of uniform designs. We run numerical experiments to show the efficiency, effectiveness and robustness of the proposed DDS method.

For future works there are several ways to extend and enhance the proposed DDS method. First, it may extend to the cases of decentralized data storage, via distributed sampling and parallel computation based on the use of sliced uniform designs. Second, the DDS method in the present form is developed for the Big Data that can be transformed to be approximately independent coordinates upon dimension reduction. When it is difficult to convert the underlying data to be independent in coordinates
(e.g., arbitrary data or manifolds), we may consider the segmented approach, e.g., by Rosenblatt transformation or recursive partitioning, then perform the DDS method in each local region. The overall goodness of subsampling can be either measured by an aggregated version of the GEFD criteria, or some central composite variant upon appropriate space segmentation.

ACKNOWLEDGMENTS

The authors would like to thank Hongquan Xu for his valuable comments.

REFERENCES

[1] L. Bottou, F. E. Curtis, and J. Nocedal, “Optimization methods for large-scale machine learning,” SIAM Rev., vol. 60, no. 2, pp. 223–311, 2018.
[2] M. Blum, R. W. Floyd, V. R. Pratt, R. L. Rivest, and R. E. Tarjan, “Time bounds for selection,” J. Comput. Syst. Sci., vol. 7, no. 4, pp. 448–461, 1973.
[3] H. Chen, H. Z. Huang, D. K. J. Lin, and M. Q. Liu, “Uniform sliced Latin hypercube designs,” Appl. Stochastic Models Bus. Ind., vol. 32, no. 5, pp. 574–584, 2016.
[4] Y. Chen, M. Welling, and A. Smola, “Super-samples from kernel herding,” in Proc. 26th Conf. Uncertainty Artif. Intell., 2010, pp. 109–116.
[5] D. Dheeru and E. Karra Taniskidou, “UCI machine learning repository,” 2017. [Online]. Available: http://archive.ics.uci.edu/ml
[6] K. T. Fang, M. Q. Liu, H. Qin, and Y. D. Zhou, Theory and Application of Uniform Experimental Designs. Singapore & Beijing, China: Springer & Science Press, 2018.
[7] K. T. Fang, R. Li, and A. Sudjianto, Design and Modeling for Computer Experiments. London, U.K.: Chapman and Hall/CRC, 2006.
[8] K. T. Fang and Y. Wang, Number-Theoretic Methods in Statistics. London, U.K.: Chapman and Hall, 1994.
[9] F. J. Hickernell, “A generalized discrepancy and quadrature error bound,” Math. Comput., vol. 67, no. 221, pp. 299–322, 1998.
[10] F. J. Hickernell, “Lattice rules: How well do they measure up?,” in Random and Quasi-Random Point Sets, vol. 138, P. Hellekalek and G. Larcher, Eds., New York, NY, USA: Springer, 1998, pp. 109–166.
[11] W. M. Ho and Z. Q. Xu, “Applications of uniform design to computer experiments,” J. Chin. Statist. Assoc., vol. 38, pp. 395–410, 2000.
[12] C. L. Liu and P. Y. Hsieh, “Model-based synthetic sampling for imbalanced data,” IEEE Trans. Knowl. Data Eng., vol. 32, no. 8, pp. 1543–1556, Aug. 2020.
[13] P. Ma, M. Mahoney, X. Xing, and B. Yu, “A statistical perspective on algorithmic leveraging,” J. Mach. Learn. Res., vol. 16, pp. 861–911, 2015.
[14] P. Ma, X. L. Zhang, X. J. Xing, J. X. Ma, and M. Mahoney, “Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms,” in Proc. Int. Conf. Artif. Intell. Statist., 2020, pp. 1026–1035.
[15] M. W. Mahoney, “Randomized algorithms for matrices and data,” Found. Trends Mach. Learn., vol. 3, no. 2, pp. 123–224, 2011.
[16] S. Mak and V. R. Joseph, “Support points,” Ann. Statist., vol. 46, no. 6A, pp. 2562–2592, 2018.
[17] C. Meng, R. Xie, A. Mandal, K. L. Zhang, W. X. Zhong, and P. Ma, “LowCon: A design-based subsampling approach in a misspecified linear model,” J. Comput. Graph. Statist., vol. 30, no. 3, pp. 694–708, 2021.
[18] M. D. Morris, T. J. Mitchell, and D. Ylivieska, “Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction,” Technometrics, vol. 35, no. 3, pp. 243–255, 1993.
[19] H. Niederreiter, Random Number Generation and Quasi-Monte Carlo Methods. Philadelphia, PA, USA: SIAM, 1992.
[20] H. Weyl, “On the equal distribution of the number mod. one in (Germany),” Math. Ann., vol. 77, no. 3, pp. 313–352, 1916.
[21] H. Y. Wang, R. Zhu, and P. Ma, “Optimal subsampling for large sample logistic regression,” J. Amer. Statist. Assoc., vol. 113, no. 522, pp. 829–844, 2018.
[22] H. Y. Wang, M. Yang, and J. Stufken, “Information-based optimal subdata selection for Big Data linear regression,” J. Amer. Statist. Assoc., vol. 114, no. 525, pp. 393–405, 2019.
[23] B. A. Worley, Deterministic uncertainty analysis, Oak Ridge National Lab., TN, USA, Tech. Rep. ORNL-6428, 1987.
[24] C. A. R. Hoare, “Quicksort,” Comput. J., vol. 5, no. 1, pp. 10–16, 1962.