Pattern Recognition 124 (2022) 108454

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Sketches by MoSSaRT: Representative selection from manifolds with gross sparse corruptions

Mahlagha Sedghi [a], Michael Georgiopoulos [a], George K. Atia [a,b,*]

[a] Department of Electrical and Computer Engineering, University of Central Florida
[b] Department of Computer Science, University of Central Florida

ARTICLE INFO

Article history:
Received 27 August 2020
Revised 29 May 2021
Accepted 25 November 2021
Available online 27 November 2021

Keywords:
Representative selection
Gross sparse corruption
Manifold learning
Reproducing kernel Hilbert spaces

ABSTRACT

Conventional sampling techniques fall short of selecting representatives that encode the underlying conformation of non-linear manifolds. The problem is exacerbated if the data is contaminated with gross sparse corruptions. In this paper, we present a data selection approach, dubbed MoSSaRT, which draws robust and descriptive sketches of grossly corrupted manifold structures. Built upon an explicit randomized transformation, we obtain a judiciously designed representation of the data relations, which facilitates a versatile selection approach accounting for robustness to gross corruption, descriptiveness and novelty of the chosen representatives, simultaneously. Our model lends itself to a convex formulation with an efficient parallelizable algorithm, which coupled with our randomized matrix structures gives rise to a highly scalable implementation. Theoretical analysis guarantees probabilistic convergence of the approximate function to the desired objective function and reveals insightful geometrical characterization of the chosen representatives. Finally, MoSSaRT substantially outperforms the state-of-the-art algorithms as demonstrated by experiments conducted on both real and synthetic data.

1. Introduction

The ever-increasing availability of large-scale and high-dimensional data offers unprecedented opportunities for data-driven studies across widely differing domains ranging from marketing and web mining, to bioinformatics and space exploration. Yet, it also poses formidable challenges in face of storing, organizing and analyzing such data.

With regard to dimensionality, there has been enormous progress in devising solutions for the analysis and visualization of high-dimensional data through low-dimensional embedding methods, e.g., Principal Component Analysis, Isomap, feature selection and dictionary learning algorithms, and embedding techniques via random projections [1–6].

Another line of research focuses on extracting knowledge from a sheer volume of data by tapping into the sample space while keeping the dimension intact. This paper focuses on the problem of representative selection, which has elicited strong interest from the data sciences communities in recent years. Selecting representative samples aims at reducing the problem size by subsampling the data points independently of the dimension, while minimizing the information loss. A major distinction from other methods obtaining compact representations in the sample space such as dictionary learning approaches is that the chosen representative subsets consist of actual data points, thereby affording easy interpretations in various application domains. For instance, these subsets could consist of distinct images in a collection and specific words in a document, or particular sensors and bands in a system and hyperspectral imaging [7–9]. The advantages of representative selection are multifold. Notably, substantial savings in storage and computation can be derived from the development of inference algorithms around descriptive and concise data sketches in lieu of the full-scale data. This is particularly relevant with the emergence of edge machine learning paradigms in which complex algorithms are required to run locally on tiny and resource-constrained devices with minimal information centralization. For instance, advancements in virtual and augmented reality such as Oculus and HoloLens [10,11] and smart wearable devices necessitate the efficient integration of state-of-the-art models and algorithms into these portable computing platforms, such as deep learning models, whose computational/memory burden and power consumption substantially overtax the resources of smaller devices. For example, a widely used classification network known as AlexNet [12] performs 1.5 billion high precision operations through 61 M parameters and takes 249 MB of memory per image. The requirements are even more considerable for more

* Corresponding author.
E-mail addresses: mahlagha.s@knights.ucf.edu (M. Sedghi), michaelg@ucf.edu (M. Georgiopoulos), George.Atia@ucf.edu (G.K. Atia).

https://doi.org/10.1016/j.patcog.2021.108454
0031-3203/

complex networks such as VGG and GANs [13,14]. Another domain is in the automative industry where mobility platforms have to realize efficient data management solutions to address the complexity underlying Advanced Driver-Assistance Systems (ADAS) and autonomous driving given a sheer volume of sensory data from Radar, Lidar, Cameras, Sonar, and GPS, among others [15,16]. Other advantages facilitated through concise and informative representatives include insightful summaries of complex systems, deeper grasp of complex underlying interactions, simpler data annotation and cleansing processes, and even better generalization and enhanced phase transitions for supervised and unsupervised learning algorithms [17].

Despite notable progress in developing compelling approaches to representative selection, some important limitations of prior work motivate the work of this paper. First, the vast majority of existing approaches rest upon linearity assumptions about the data. One commonly made assumption is that it lies in a union of low-dimensional linear subspaces. In many real-world scenarios, however, the underlying data patterns can be modeled more accurately by non-linear manifold structures of lower intrinsic dimensionality, rather than linear subspaces. Second and most important, while there exist numerous methods that are robust to various data perturbations such as gross corruptions, outliers, and noise under linear data models, no principled approach is known to date to handle such perturbations in the presence of non-linear data structures. Sparse gross corruptions, a central focus of this work, can be caused by occlusions, measurement errors, and adversarial interference and can easily jeopardize the validity of the existing methods due to their arbitrary magnitude and unknown support [18,19]. Therefore, selecting descriptive and compact samples under these practical circumstances remains largely unexplored. Motivated by this, here we study the problem of representative selection from manifold structures with gross sparse corruptions.

1.1. Summary of contributions

This paper makes five main contributions. First, for the first time we formalize the problem of representative selection from non-linear manifolds in presence of gross sparse corruptions in a principled and mathematically rigorous framework. Based on a constrained optimization formulation in a transformed space, we obtain an encoding of the data relations, termed *reproduction profile*, which we leverage to draw a representative, diverse and concise sketch of the data.

Second, we leverage the rich representation power of Reproducing Kernel Hilbert Spaces (RKHSs) to capture the non-linearities in the data structure. Much of the existing work in kernel settings is based on merely replacing the original inner products with kernel evaluations. However, as our formulation relies on sparsity-inducing norms to adequately handle sparse corruptions, the use of the standard kernel trick is not feasible. To overcome this issue, we integrate an approximate feature mapping framework in our formulation to emulate a desired feature mapping associated with a RKHS. While any approximate feature can be potentially plugged into our method depending on the data specifics, we showcase the use of random Fourier features [20] due to the wide use of stationary kernels in machine learning applications. The utility of these features, which were introduced for accelerating kernel machines, rests upon a classic result in harmonic analysis. Here, we exploit similar features for the first time in the context of representative sampling to mirror the unknown mapping of the RKHS.

Third, we develop a highly scalable and parallelizable ADMM-based algorithm for representative sampling. Leveraging the special structures of the approximate feature maps, the algorithm exhibits nearly linear complexity in the data size.[1]

Our fourth contribution lies in establishing key theoretical results affording guarantees on the goodness of the approximation induced by random feature maps and a characterization of the sampled representative set. In particular, based on concentration of measure arguments, we show that the optimal value for the proxy objective function induced by the approximate features converges to the true optimal value exponentially fast, thereby establishing the effectiveness of our proxy formulation (Theorem 1). In addition, we present a characterization rooted in geometric functional analysis of the sampled subset, which provides the theoretical underpinning of an interpretable mechanism for sampling informative representatives. In particular, it turns out that the sampled subset of representatives consists of the vertices of the symmetrized convex hull of all samples in a transformed space (Theorem 2).

As our final contribution, we demonstrate the effectiveness of the proposed approach using both synthetic and real data in a broad range of supervised and unsupervised applications, including classification, clustering, and face pose generation using Generative Adversarial Networks (GANs).

Fig. 1 illustrates a conceptual diagram of the proposed framework, which will be explicated in further detail in Section 2.

1.2. Related work

Random selection approaches are ineffective in fully describing the entire set due to redundancy and corruptions in the data. On the other hand, optimal subset selection is generally NP-hard. Hence, various relaxations of the problem have been tackled by different approaches, which can be mainly categorized into three classes: linear, diversity-based and clustering-based methods.

Linear algebraic methods typically found their models on the low-rankness of the data collection. Rank-Revealing QR (RRQR) algorithms [21–23] aim to find a permutation matrix that, when multiplied by the data matrix, reveals the best conditioned submatrix as its first columns. Others have focused on choosing some columns that can best span the column space of the original dataset [24–27]. Missing entries and non-negative matrices are considered in Balzano et al. [28], Esser et al. [29] via a greedy algorithm and ℓ_1/ℓ_∞ optimization, respectively. Inspired by dictionary learning approaches, Elhamifar et al. [27] uses a linear model in which each point in the dataset is described as a linear combination of others and a sparsity constraint is enforced to get a few representatives. The authors in Wang et al. [30] diversify these chosen samples by employing multiple regularization terms. Intuitively, the approaches of the first class all seek to find a low-rank approximation of the data matrix to recover its column space. Hence, they are only suited for linear models and cannot capture the non-linearity properly.

Diversity-based approaches, on the other hand, focus mainly on information novelty. A-optimal and D-optimal approaches [8,31] build on convex relaxations of the original problem. A faster greedy optimization algorithm is suggested in Shamaiah et al. [32], however it yields a sub-optimal solution, since the actual cost function of the problem is not sub-modular [33]. In an effort to maximize the diversity, these methods are all negligent of the representation power of the chosen subset, and are highly prone to choosing irrelevant corrupted data points.

Alternatively, the clustering-based approaches typically use similarity relationship among data points, which makes them potentially more suitable for sampling from non-linear data. Centroids

[1] The code is available at https://github.com/Mahlagha/MoSSaRT.

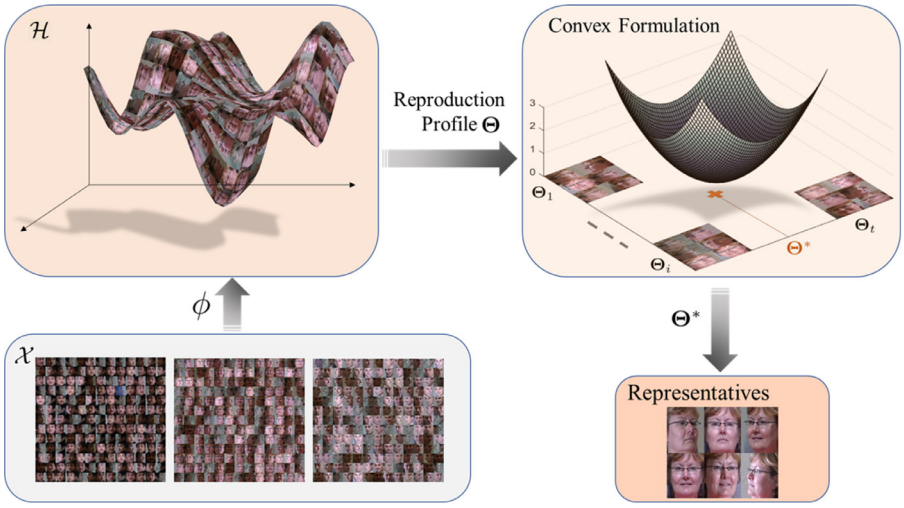

Fig. 1. Conceptual diagram of the proposed Representative Selection framework. First, the underlying patterns of the huge dataset are captured through non-linear manifold structures. The original collection is transformed into an explicit Hilbert space, emulating a desired implicit RKHS. Then, a reproduction profile is introduced for each sample, using which the combinatorial subset selection problem is formulated as a convex minimization. The optimization is corruption-aware, hence, the optimal reproduction profile indicates the best subset which negates the effects of gross corruptions in the data, while preserving the underlying structure of the whole collection.

of the clusters obtained by various clustering techniques are identified as representatives. In [34], exemplars are selected to minimize the total distance from all samples, and Charikar et al. [35] approximates the k-means algorithm. The efficacy of these algorithms is adversely affected by their high dependence on initialization. This issue was addressed in Frey and Dueck [36,37], where the cluster centroids were identified by a message passing procedure. Also, in Elhamifar et al. [17] a trace minimization program was suggested to find exemplars for a source and target set. These methods yield sub-optimal solutions and require restrictive conditions on the similarities to perform well.

Among all the developed techniques, only a few have specifically attempted to tackle the problem with a Manifold Learning (ML) approach. These methods mostly adopt graph-based distances as approximate measures of geodesic distances, or resemble manifolds by processing local neighborhood sets in a linear fashion. In [38], a geodesic measure minimization is included in the formulation of a RRQR-based factorization assuming a priori knowledge of the structure of the manifold. In [39], sampling of manifolds is tackled through an iterative scheme, where the spectrum of the Laplace–Beltrami operator on manifolds is approximated. In [40], a similarity-based quadratic criterion is optimized for high representability while rejecting column-wise outliers. A graph-based variant of the k-means algorithm is proposed in Tu et al. [41], where Euclidean distances are replaced by geodesic distances to account for the intrinsic characteristics of the manifold. These methods, either inherit the deficiency of the original methods such as dependency on initialization and complex iterations, or incorporate local information, which diminishes their ability to capture a global view of the collection.

Notation. Let $\mathbb{N}_k \triangleq \{1, \ldots, k\}$ for $k \in \mathbb{N}$. Column vectors and matrices are denoted in boldface lower-case and upper-case letters, respectively. Let $\mathbf{1}$ and $\mathbf{I}_n$ denote the all-ones vector of proper length, and the identity matrix of size n, respectively. For a scalar a, $|a|$ denotes its absolute value, while for a set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality. For a vector $\mathbf{a}$, $\|\mathbf{a}\|_p$ stands for its ℓ_p-norm, and $\mathbf{a}(i)$ its ith element. This notation is used for both finite-dimensional vectors and infinite sequences. When necessary, the distinction will be made explicit to avoid confusion. Accordingly, ℓ^p denotes the space of all sequences whose ℓ_p-norms are bounded. For a matrix $\mathbf{A}$, $\mathbf{a}_i, a_{ij}$ denote its ith column and (i, j)th element, respectively, $\|\mathbf{A}\|_F = \sum_i \|\mathbf{a}_i\|_2$ its Frobenius norm, and $\|\mathbf{A}\|_{1,p} = \sum_i \|\mathbf{a}_i\|_p$ its group Lasso norm. Similar to vectors, the notation is shared between matrices whose columns are finite-dimensional vectors or infinite sequences. Matrix $\mathbf{A_3} = [\mathbf{A_1}\ \mathbf{A_2}]$ denotes the concatenation of two matrices $\mathbf{A_1}, \mathbf{A_2}$, with equal number of rows. The hinge function denoted $[\cdot]_+$ is defined as $\max\{\cdot, 0\}$. For a random variable (RV) x, $M_x(\gamma)$ denotes its Moment Generating Function (MGF) with parameter γ. Also, the probability of realization of a random event $\mathcal{A}$, is denoted by $\mathbb{P}\{\mathcal{A}\}$.

2. Proposed method

In this section, we present the **M**anif**o**ld **S**ampling through **Sp**a**r**se **R**eproduction Profile of Randomized **T**ransformations (MoSSaRT) method, a powerful sampling approach for high-dimensional data governed by low-dimensional manifold structures. A key aspect is that the data is contaminated with gross sparse corruptions. Inspired by many real-world scenarios, the proposed method applies to both linear and non-linear models by choosing suitable settings. Formally, our adopted data model is as follows.

Data Model 1. The columns of matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ consist of corrupted observations from the set $\mathcal{X} \triangleq \{\mathbf{x}_i\}_{i \in \mathbb{N}_n}$. We assume that the clean data lies on a low-dimensional manifold $\mathcal{M}$, and each coordinate of the data points may be contaminated with gross corruption with a small probability p, resulting in a sparse corruption matrix $\mathbf{S}$, whose elements follow a Bernoulli distribution with probability $1 - p$. This gives rise to a natural decomposition of the data matrix as $\mathbf{X} = \mathbf{M} + \mathbf{S}$, where $\mathbf{M}$ refers to the collection of points drawn from the manifold $\mathcal{M}$.

Remark 1. Note that our adopted data model does not restrict the low-dimensional structures to linear settings; this element can come from a low-rank linear subspace, or a low-dimensional non-linear Riemannian manifold. The focus of this paper will be on the more challenging scenario for the non-linear settings, but as will be shown in the sequel, the linear case is a special case of our formulation.

In principle, there is a trade-off between the number of chosen representatives and the amount of information retrieved. We approach the problem noting these two confronting criteria. Our desirable exemplars are rich in representation power to maximize

the information content, but also not too similar to minimize redundancy. One natural choice would be a minimax type of formulation between these two criteria. However, this may yield an unduly aggressive strategy, given that in most realistic scenarios many data points in the collection could be redundant and individual data points may not be too informative. In addition, while such a formulation could implicitly reduce the amount of data, it may not meet explicit budget constraints for representative selection. In order to dictate these constraints more forcefully, we deliberately develop a two-stage strategy, wherein the first stage acts as the main building block, where we obtain thorough structural information of the underlying manifold, and the second step leverages the obtained encoding to impose any existing budget constraints explicitly.

Additionally, note that fulfilling the first criterion intrinsically affords robustness to the disturbances introduced by the gross corruptions, since otherwise, the chosen samples would not be able to represent the whole dataset with enough fidelity. Hence, the main goal is to choose few samples which can be descriptive representatives of the true data, in spite of observing the contaminated data.

For the sake of identifiability, we assume that the clean data is unlikely to be sparse. Otherwise, the decomposition problem becomes ill-posed since there is no unique solution to the problem (e.g., see Candès et al. [18]). We hypothesize that if there is a "good" representative subset for the data with few elements, then there should exist a low-cost projection of the data onto the span of that subset. The patterns underlying manifold data make it challenging to desirably model this behavior in the original domain, but one may achieve such a representation through a suitable transformation. To effectively capture the non-linear behavior of the data, we relax this criterion to any separable Hilbert space $\mathcal{H}$ up to a continuous transformation. In other words, we consider a possibly highly non-linear mapping function $\phi : \mathbb{R}^m \to \mathcal{H}$, where in the transformed domain the data points can be better represented by a small subset of the collection. It is worth noting that, in general, the elements of the Hilbert space are abstract vectors (such as functions), but since every separable Hilbert space has an orthonormal basis [42], any element can be uniquely specified by its coordinates w.r.t. that basis. In what follows, $\phi(\mathbf{x})$ denotes either the vector or the infinite sequence of its coordinates.

To elucidate our approach, knowing the data decomposition in hindsight, one can then formulate the oracle in (1) aiming at sparsifying the residual errors corresponding to the corruption matrix, while satisfying the reconstruction of the clean data.

$$\begin{aligned} &\min_{\Omega\subset\mathcal{X}} \|\phi(\mathbf{S}) - \pi_\Omega(\phi(\mathbf{S}))\|_0 \\ &\text{s.t.} \quad \phi(\mathbf{M}) = \pi_\Omega(\phi(\mathbf{M})), \quad |\Omega| = \varkappa \end{aligned} \tag{1}$$

where, for a matrix $\mathbf{A}$, $\phi(\mathbf{A})$ is defined as the matrix of element-wise evaluation of the function ϕ at the columns of $\mathbf{A}$, i.e. $\phi(\mathbf{A}) \triangleq [\phi(\mathbf{a}_1)\ \phi(\mathbf{a}_2)\ \dots\ \phi(\mathbf{a}_n)]$, and $\pi_S(\mathbf{A})$ stands for the projection of $\mathbf{A}$ onto the span of its selected columns indexed by $\mathcal{S}$.

To re-formulate this combinatorial optimization as a convex problem, we translate it into finding a real-valued matrix $\mathbf{\Theta} \in \mathbb{R}^{n\times n}$, which we call the *reproduction profile* of the dataset. The appellation is associated with the encoded information in this matrix which is delineated in Remark 2. The reproduction profile $\mathbf{\Theta}$ aims to emulate the projection operator when multiplied by the data matrix. To this end, $\mathbf{\Theta}$ is enforced to have sparse rows, such that the data is projected into the subspace spanned by the samples corresponding to the non-zero rows. Hence, it can be re-expressed as

$$\begin{aligned} &\min_{\mathbf{\Theta}} \|\phi(\mathbf{S}) - \phi(\mathbf{S})\mathbf{\Theta}\|_0 \\ &\text{s.t.} \quad \phi(\mathbf{M}) = \phi(\mathbf{M})\mathbf{\Theta}, \quad \|\mathbf{\Theta}^T\|_{0,l} = \varkappa, \quad \mathbf{\Theta} \neq \mathbf{I}_n \end{aligned} \tag{2}$$

Bearing in mind the successful employment of kernel methods in identifying the non-linear patterns hidden in the data, we would like our mapping function ϕ to resemble a feature mapping $\varphi_{\mathcal{H}} : \mathbb{R}^m \to \mathcal{H}$ associated with a RKHS $\mathcal{H}$. In this case, for a RKHS with the reproducing kernel k, a mapped feature $\varphi_{\mathcal{H}}(\mathbf{x})$ is itself a function from the input space to $\mathbb{R}$, such that $\varphi_{\mathcal{H}}(\mathbf{x}_i)(\mathbf{x}_t) = k(\mathbf{x}_i, \mathbf{x}_t),\ \forall \mathbf{x}_i, \mathbf{x}_t \in \mathcal{X}$. Then one can choose an orthonormal basis for this function space, collect the resulting coordinates for all elements in a matrix, and attempt to minimize its ℓ_0-norm as in (2). Note that, our formulation does not involve explicit inner products of the data points given our use of the ℓ_0-norm in order to capture the sparse structure of the corruption. Therefore, the common practice of substituting the inner products in the original space by those in the RKHS – a technique referred to as the *kernel trick* – is not feasible in our setting. Moreover, since the explicit feature mappings are not known in general, we obtain an approximate feature mapping function, such that it emulates that of the desired RKHS.

Existing feature approximation methods are primarily developed to accelerate the classical kernel methods. By contrast, here we exploit such approximations to overcome the foregoing issues, namely, the lack of explicit inner products in our formulation and the unknown feature mapping of the RKHS. Various approximations have been developed to provide an explicit feature mapping associated with different types of kernels, such as random Fourier features [20], fast random binning features [43], additive kernel approximates [44], locality sensitive binary codes [45], and compact random features [46] (e.g., see Liu et al. [47] for a recent survey of these methods). While any approximate feature map can be plugged in our proposed approach, we focus on the class of stationary positive-definite (pd) kernels (for which the random Fourier features were proposed) due to their wide use in machine learning applications. For a stationary pd kernel k, a result from harmonic analysis by Bochner [48] is applicable, asserting the existence of a probability measure $\mu(\boldsymbol{\zeta})$, with k as its Fourier transform. Accordingly, to approximate the RKHS features, we use the vector-valued function $\phi(\cdot; \boldsymbol{\zeta}, \beta) : \mathbb{R}^m \to \mathbb{R}^r$, where each element is calculated as $\sqrt{2/r}\sin(\boldsymbol{\zeta}_i{}^T\mathbf{x} + \beta_i)$, and $\{\boldsymbol{\zeta}_i, \beta_i\}_{i\in\mathbb{N}_r}$ are i.i.d. realizations from the independent distributions $\mu(\boldsymbol{\zeta})$, the inverse Fourier transform of the kernel function, and $U[0, \pi]$, respectively.

Now, recall that the formulation in (2) involves hindsight, as the data decomposition of Data Model 1 is not available explicitly, and this in fact, poses a core subtlety to our problem. Henceforth, inspired by (2), we propose the alternative formulation (3) expressed in terms of the observed contaminated data, where the problem has been also convexified by replacing the ℓ_0-norms by their tight ℓ_1 surrogates.

$$\min_{\mathbf{\Theta}} \underbrace{\sum_{t=1}^{n} \|\sin(\boldsymbol{\zeta}^T\mathbf{x}_t + \beta) - \sin(\boldsymbol{\zeta}^T\mathbf{x}_t + \beta)\boldsymbol{\theta}_t\|_1 + \lambda\|\mathbf{\Theta}^T\|_{1,l}}_{\triangleq f(\phi,\mathbf{\Theta})} \tag{3}$$

The first term amounts to a representation constraint, and the employed regularization automatically avoids the trivial solution of identity, hence eliminating the need for the constraint $\mathbf{\Theta} \neq \mathbf{I}_n$. Inspired by the oracle non-convex and constrained optimization in (2), our formulation in (3) yields excelling performance as shown in Section 4.

Remark 2. The optimal reproduction profile $\mathbf{\Theta}^*$ contains structural information about the collective behavior of the data points, which enables us to not only draw representative sketches, but also to ensure novelty. More specifically, each row of this matrix encodes how a given sample participates to reproduce the whole collection under the presumed constraints of adhering to manifold structures, while negating the impact of the gross corruptions. Therefore, sam-

ples are associated with an elaborate profile describing their reproducing ability. Hence, the representative points can be identified by the non-zero row-norms of the optimal profile, while hard constraints can be satisfied by choosing the most distinctive ones among the identified samples.

A secondary step ensures the selected set is compact so that each element contains novel information, otherwise gets eliminated. Beside offering variability, this step allows us to impose the budget constraint explicitly at no extra cost. We consider the samples to be analogous if they are close in the transformed space as

$$\delta(\mathbf{x}_i, \mathbf{x}_t) = \sqrt{\hat{k}(\mathbf{x}_i, \mathbf{x}_i) - 2\hat{k}(\mathbf{x}_i, \mathbf{x}_t) + \hat{k}(\mathbf{x}_t, \mathbf{x}_t)}\,, \tag{4}$$

where $\hat{k}(\mathbf{x}_i, \mathbf{x}_t) \triangleq \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_t) \rangle$ acts as the inner product in the transformed space, as a proxy to k, the actual inner product in the RKHS. Our measure to choose between two similar points is then their level of representation power. Scrutinizing the obtained encoding reveals that samples corresponding to higher row-norms of the encoding matrix contribute more to the reconstruction of the whole dataset, and hence, can be regarded as more influential representatives of the dataset. Exploiting this information, the procedure avoids the effort of fine-tuning hyper-parameters, and ensures maximal novelty in the chosen subset without sacrificing its representativeness.

2.1. Algorithm, complexity and scalablity

Generic solvers for convex problems such as CVX [49,50] have cubic or higher complexities, thus do not scale well with the problem size. To alleviate this problem, we develop an Alternating Direction Method of Multipliers (ADMM)-based algorithm [51], which reduces the computational costs and also enables parallel implementation of this program. As will be shown later in the section, employing the involved matrix structures as well as the ADMM approach yields a near-linear computational complexity of $\mathcal{O}(r^2 n^{1.373})$, where n is the number of samples and $r \ll n$ is the model parameter for the dimension of the proposed feature. Algorithm 1 illustrates the big picture of the sampling process, where the superscript inside the parenthesis is the iteration indicator, the convergence conditions for primal and dual feasibility are derived according to Boyd et al. [51], and the proximal operators of different norms are derived as follows: $\mathscr{T}_{\epsilon}^{(1)}(\mathbf{X}) \triangleq \text{sgn}(x)[|x| - \epsilon]_+$, and $\mathscr{T}_{\epsilon}(\mathbf{X})^{(1,1)} \triangleq [x-\epsilon]_+ - [-x-\epsilon]_+$ apply to the elements of a matrix. $\mathscr{T}_{\epsilon}(\mathbf{X})^{(1,2)}$ applies to the rows of a matrix and is taken to be $\mathbf{x} - \epsilon\mathbf{x}/\|\mathbf{x}\|_2$, if $\|\mathbf{x}\|_2 > \epsilon$, and the zero vector otherwise. Also, the matrix $\phi(\mathbf{X})$ is denoted by $\mathbf{\Phi}$ for simplicity, and $\hat{\mathbf{K}}$ stands for the pairwise inner product of the transformed features.

It is not hard to show that the overall computational complexity of our algorithm is dominated by the matrix inversion and multiplication in step 6, since the remaining steps consist of matrix summations or scalar multiplications. In general, efficient multiplication of matrices of the size $p \times q$ and $q \times t$ is shown to be $\mathcal{O}(pq^{0.373}t)$ [52]. Accordingly, the complexity of the inversion of a $n \times n$ matrix follows from the same algorithm and is $\mathcal{O}(n^{2.373})$ [52,53]. Since our algorithm is parallelizable, using P parallel processors leads to $\mathcal{O}(n^{1.373}\lceil n/P \rceil)$ computational complexity [54]. However, we illustrate in the following that via the use of specific structures of the matrices involved in our computations, we are able to reduce the complexity of our algorithm even further. Observe that the matrix whose inverse is calculated is of the form $\mathbf{I}_n + \hat{\mathbf{K}}$. Suppose we have the Singular Value Decomposition of the low-rank feature matrix as $\mathbf{\Phi} = \mathbf{SVD}^T$. Then, utilizing some matrix manipulations, one can show that

$$[\mathbf{I}_n + \hat{\mathbf{K}}]^{-1} = \mathbf{D}_1(\mathbf{I}_r + \mathbf{\Sigma})^{-1}\mathbf{D}_1{}^T + \mathbf{I}_n$$

Algorithm 1 Proposed sampling scheme using ADMM optimization.

Require: Data Matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, kernel k, desired number of samples $\varkappa$,$l \in \{1, 2\}$, Optimization parameters λ, ρ, feasibility tolerances $\epsilon_{abs}, \epsilon_{rel}$ (both set to 10^{-5} in our experiments).

Ensure: Collection of the representative samples $\mathcal{R}$

1: Construct the feature matrix $\mathbf{\Phi} \in \mathbb{R}^{r \times n}$ as explained in Section 2, using $\mathbf{X}, k$.
2: **Initialization:**Set $\mathbf{\Theta}^{(0)}$ and $\mathbf{U}_1^{(0)}, \mathbf{U}_2^{(0)} \in \mathbb{R}^{r \times n}$ to zero matrices, $converged = False$
3: **while** not $converged$ **do**
{Update Primal variables}
4: $\mathbf{Q}^{(i+1)} = \mathscr{T}_{1/\rho}^{(1)}(\mathbf{\Phi} - \mathbf{\Phi\Theta}^{(i)} - \mathbf{U_2}^{(i)})$
5: $\mathbf{B}^{(i+1)} = \mathscr{T}_{\lambda/\rho}^{(1,p)}(\mathbf{\Theta}^{(i)} - \mathbf{U_1}^{(i)})$
6: $\mathbf{\Theta}^{(i+1)} = (\mathbf{I}_n + \hat{\mathbf{K}})^{-1}(\mathbf{B}^{(i+1)} + \mathbf{\Phi}^T(\mathbf{\Phi} - \mathbf{Q}^{(i+1)} - \mathbf{U}_2^{(i)} + \mathbf{U}_1^{(i)}))$
{Update Dual variables}
7: $\mathbf{U}_1^{(i+1)} = \mathbf{U}_1^{(i)} + \mathbf{B}^{(i+1)} - \mathbf{\Theta}^{(i+1)}$
8: $\mathbf{U}_2^{(i+1)} = \mathbf{U}_2^{(i)} + \mathbf{Q}^{(i+1)} - \mathbf{\Phi} + \mathbf{\Phi\Theta}^{(i+1)}$
{Calculate Primal and Dual residuals}
9: $\mathbf{PR_1}^{(i)} = \mathbf{Q}^{(i)} - \mathbf{\Phi} + \mathbf{\Phi\Theta}^{(i)}$, $\mathbf{PR_2}^{(i)} = \mathbf{B}^{(i)} - \mathbf{\Theta}^{(i)}$
10: $\mathbf{DR_1}^{(i)} = \rho(\mathbf{\Phi}(\mathbf{\Theta}^{(i)} - \mathbf{\Theta}^{(i-1)}))$, $\mathbf{DR_2}^{(i)} = \rho(\mathbf{\Theta}^{(i)} - \mathbf{\Theta}^{(i-1)})$
{Calculate Primal and Dual feasibility tolerances}
11: $\epsilon_{p1} = n\epsilon_{abs} + \epsilon_{rel}\max(\|\mathbf{Q}^i\|_F, \|\mathbf{\Phi} - \mathbf{\Phi\Theta}^{(i)}\|_F)$, $\epsilon_{d1} = n\epsilon_{abs} + \rho\epsilon_{rel}\|\mathbf{U}_1^{(i)}\|_F$
12: $\epsilon_{p2} = n\epsilon_{abs} + \epsilon_{rel}\max(\|\mathbf{Q}^i\|_F, \|\mathbf{\Theta}^{(i)}\|_F)$, $\epsilon_{d2} = n\epsilon_{abs} + \rho\epsilon_{rel}\|\mathbf{U}_2^{(i)}\|_F$
{Check Convergence}
13: **If** $\|\mathbf{PR_1}^{(i)}\|_F \le \epsilon_{p1} \wedge \|\mathbf{PR_2}^{(i)}\|_F \le \epsilon_{p1} \wedge \|\mathbf{DR_1}^{(i)}\|_F \le \epsilon_{d1} \wedge \|\mathbf{DR_2}^{(i)}\|_F \le \epsilon_{d2}$
14: $converged = True$
15: **end while**
16: $\mathbf{\Theta}^* = \mathbf{\Theta}^{(i)}$
17: $\mathcal{R} = \left\{\mathbf{x}_j \in \mathbf{X} \;\middle|\; \|\theta^{*j}\|_2 \neq 0\right\}$
18: Calculate the pairwise distances acc. to Equation (4)
19: **while** $|\mathcal{R}| > \varkappa$ **do**
20: Remove less active samples similar to the highly contributing ones from $\mathcal{R}$
21: **end while**
22: **Return** $\mathcal{R}$

where $\mathbf{D}_1$ is the first r columns of the matrix $\mathbf{D}$, and $\mathbf{\Sigma}$ is the non-zero $r \times r$ sub-matrix of $\mathbf{V}^T\mathbf{V}$, with elements σ_i on the diagonals. Now we need to take the inverse of a diagonal matrix, which can be easily represented by a diagonal matrix with entries $\frac{1}{1+\sigma_i}$, which we denote by $\text{diag}\frac{1}{1+\sigma_i}$. The developed scheme effectively reduces the computational complexity of the original inversion down to $\mathcal{O}(r^2 n)$, exhibiting a linear complexity in n. Then, having the resultant matrix of the form $\mathbf{D}_1(\text{diag}\frac{1}{1+\sigma_i})\mathbf{D}_1{}^T$ multiplied by an $n \times n$ matrix yields the near-linear complexity of $\mathcal{O}(r^2 n^{1.373})$. The analysis is validated in our experiments reported in Section 4.7. Again, utilizing the P multi-processing cores reduces the complexity to $\mathcal{O}(r^2 n^{0.373}\lceil n/P \rceil)$.

3. Theoretical analysis

In this section, we present probabilistic and geometric analysis of the proposed method and the employed building blocks. As our main result, we first evaluate how well our proxy for the optimal encoding matrix minimizes the true cost in which the actual feature mapping $\varphi_{\mathcal{H}}$ is considered (Theorem 1). Then, we present a characterization of the obtained representatives based on geometric functional analysis (Theorem 2). The approximation of RKHS

features via random Fourier features was shown to uniformly converge to that of a given shift-invariant kernel originally in Rahimi and Recht [20], and later improved in Sutherland and Schneider [55]. The relatively tight and uniform bounds obtained in both approaches hinge upon some assumptions, including the compactness of the input space and the existence of the first two moments of $\mu(\boldsymbol{\zeta})$. Here, we dispense with these assumptions and obtain a concentration result that suffices to prove the result of Theorem 1. Relaxing the compactness requirement of the input domain is beneficial for convergence analysis in various optimization contexts where such restrictions do not hold. Also, the particular features utilized here are slight variants of the original random Fourier features. Thus, for coherency, the convergence behavior of the features utilized here is provided in Lemmas 1 and 2. For clarity, we denote the approximate inner product by $\hat{k}_{\boldsymbol{\zeta},\beta}(\mathbf{x}_i,\mathbf{x}_t)$, defined as $\phi(\mathbf{x}_i;\boldsymbol{\zeta},\beta)^T\phi(\mathbf{x}_t;\boldsymbol{\zeta},\beta)$. Also, we often use the short-hand notation $\phi(\mathbf{x})$ for the proposed features, where the two random variables $\boldsymbol{\zeta},\beta$ are omitted, when no confusion arises.

Lemma 1. *For a given stationary real-valued pd kernel k, the inner product of the associated proposed features $\phi(\mathbf{x}_i;\boldsymbol{\zeta},\beta)$ and $\phi(\mathbf{x}_t;\boldsymbol{\zeta},\beta)$ approximate the evaluation of the kernel k, i.e., $\mathbb{E}_{\boldsymbol{\zeta},\beta}[\hat{k}_{\boldsymbol{\zeta},\beta}(\mathbf{x}_i,\mathbf{x}_t)]=k(\mathbf{x}_i,\mathbf{x}_t)$.*

Proof. First, recall that the MGF of a uniform random variable β_l on $[a,b]$ is known to be $M_{\beta_l}(\gamma)=\frac{\exp(\gamma b)-\exp(\gamma b)}{\gamma(b-a)}$ for non-zero gamma values. Hence, one can show that for a fixed $\boldsymbol{\zeta}_l$,

$$\mathbb{E}_{\beta_l}[\cos(\boldsymbol{\zeta}_l{}^T(\mathbf{x}_i+\mathbf{x}_t)+2\beta_l)]=\Re\{\exp(j\boldsymbol{\zeta}_l{}^T(\mathbf{x}_i+\mathbf{x}_t))\overbrace{\mathbb{E}_{\beta_l}[\exp(j2\beta_l)]}^{M_{\beta_l}(2j)}\}$$
$$=\frac{\sin(\boldsymbol{\zeta}_l{}^T(\mathbf{x}_i+\mathbf{x}_t)+2\pi)-\sin(\boldsymbol{\zeta}_l{}^T(\mathbf{x}_i+\mathbf{x}_t))}{2\pi}=0\,. \tag{5}$$

Then, with $\hat{k}_{\boldsymbol{\zeta},\beta}$ defined as the inner product of the two feature maps one can write

$$\mathbb{E}_{\boldsymbol{\zeta},\beta}[\hat{k}_{\boldsymbol{\zeta},\beta}(\mathbf{x}_i,\mathbf{x}_t)]=\mathbb{E}_{\boldsymbol{\zeta}_l,\beta_l}[1/r\sum_{l=1}^{r}\cos(\boldsymbol{\zeta}_l{}^T(\mathbf{x}_i-\mathbf{x}_t))$$
$$-\cos(\boldsymbol{\zeta}_l{}^T(\mathbf{x}_i+\mathbf{x}_t)+2\beta_l)]$$
$$=\mathbb{E}_{\boldsymbol{\zeta}}[\cos(\langle\boldsymbol{\zeta},(\mathbf{x}_i-\mathbf{x}_t)\rangle)]$$
$$+1/r\sum_{l=1}^{r}\mathbb{E}_{\boldsymbol{\zeta}_l,\beta_l}[\cos(\boldsymbol{\zeta}_l{}^T(\mathbf{x}_i+\mathbf{x}_t)+2\beta_l)] \tag{6}$$

where the last equality follows from the identical distribution of the variables $\boldsymbol{\zeta}_i$'s, and independence of $\{\boldsymbol{\zeta}_i\}_{i\in\mathbb{N}_r}$ and $\{\beta_i\}_{i\in\mathbb{N}_r}$. Now, substituting (5) into (6) implies that

$$\mathbb{E}_{\boldsymbol{\zeta},\beta}[\hat{k}_{\boldsymbol{\zeta},\beta}(\mathbf{x}_i,\mathbf{x}_t)]=\mathbb{E}_{\boldsymbol{\zeta},\beta}[\cos(\langle\boldsymbol{\zeta},(\mathbf{x}_i-\mathbf{x}_t)\rangle)]=k(\mathbf{x}_i,\mathbf{x}_t)$$

where the last equality follows by Bochner's theorem for real-valued functions [48]. □

The above analysis guarantees convergence of the features in expectation. Next, we give a stronger result establishing exponentially fast concentration around the mean.

Lemma 2. *Let $\delta>0$ be a confidence level, then for any given points $\mathbf{x}_i,\mathbf{x}_t,\{i,t\}\in\mathbb{N}_n$, with probability at least $1-\delta$*

$$\hat{k}\in\left[k-4\sqrt{2/r\log(2/\delta)},k+4\sqrt{2/r\log(2/\delta)}\right] \tag{7}$$

where $\hat{k},k$ are short-hand notations for $\hat{k}_{\boldsymbol{\zeta},\beta}(\mathbf{x}_i,\mathbf{x}_t),k(\mathbf{x}_i,\mathbf{x}_t)$, respectively.

The theorem shows that the approximate inner products obtained from the inner product of any two proposed feature vectors concentrate around their expected value, i.e., the true inner product, with high probability. To prove this result, we will first characterize the tail behavior of the involved RVs in the following lemma, and then establish the result of the theorem. The reader is referred to Vershynin [56] for the common definitions regarding sub-Gaussian property, and related theorems such as Hoeffding, and Chernoff bounds. Let $\hat{k}_l(\mathbf{x}_i,\mathbf{x}_t)\triangleq\sin(\boldsymbol{\zeta}_l{}^T\mathbf{x}_i+\beta_l)\sin(\boldsymbol{\zeta}_l{}^T\mathbf{x}_t+\beta_l)$.

Lemma 3. *$\hat{k}_l(\mathbf{x}_i,\mathbf{x}_t)$ is a sub-Gaussian RV with parameter $\alpha=2$.*

Proof. Let $\hat{\kappa}_l$ be independent of $\hat{k}_l$ and with the same distribution, where we have omitted the two arguments for simplicity. Observe that $\hat{k}_l-\hat{\kappa}_l$ matches $\epsilon(\hat{k}_l-\hat{\kappa}_l)$ in distribution, where ϵ is the Rademacher RV. Hence,

$$\mathbb{E}[\exp(\gamma(\hat{k}_l-\hat{\kappa}_l))]=\mathbb{E}_{\hat{k}_l,\hat{\kappa}_l}[\underbrace{\mathbb{E}_\epsilon[\exp(\epsilon\gamma(\hat{k}_l-\hat{\kappa}_l))]}_{M_\epsilon(\gamma(\hat{k}_l-\hat{\kappa}_l))}]\,.$$

It is not hard to show that the Rademacher RV is itself sub-Gaussian with parameter 1, and hence, together with the boundedness of $\hat{k}_l$ and $\hat{\kappa}_l$ we conclude that

$$\mathbb{E}_{\hat{k}_l,\hat{\kappa}_l}[\exp(\gamma(\hat{k}_l-\hat{\kappa}_l))]\le\exp(\gamma^2 2^2/2)$$

Finally, Jensen's inequality implies

$$\mathbb{E}_{\hat{k}_l}[\exp(\gamma(\hat{k}_l-\mathbb{E}_{\hat{\kappa}_l}\hat{\kappa}_l))]<\mathbb{E}_{\hat{k}_l,\hat{\kappa}_l}[\exp(\gamma(\hat{k}_l-\hat{\kappa}_l))]$$

which concludes the proof. □

Proof of Lemma 1. $\hat{k}_{\boldsymbol{\zeta},\beta}(\mathbf{x}_i,\mathbf{x}_t)$ is proportional to the sum of r independent RVs as

$$\hat{k}_{\boldsymbol{\zeta},\beta}(\mathbf{x}_i,\mathbf{x}_t)=1/r\sum_{l=1}^{r}\hat{k}_l(\mathbf{x}_i,\mathbf{x}_t)\,. \tag{8}$$

The Sum Rule, together with the result of Lemma 1, imply that $\hat{k}_{\boldsymbol{\zeta},\beta}$ is sub-Gaussian with parameter $\alpha=4/\sqrt{r}$. Recalling that $\mathbb{E}_{\boldsymbol{\zeta},\beta}[\hat{k}_{\boldsymbol{\zeta},\beta}(\mathbf{x}_i,\mathbf{x}_t)]=k(\mathbf{x}_i,\mathbf{x}_t)$ by Lemma 1, the concentration of the RV of interest around its mean is inferred from Chernoff bound,

$$\mathbb{P}\{|\hat{k}-k|\ge\tau\}\le 2\exp(-r\tau^2/32)\,.$$

Finally, setting the obtained bound to a desired confidence level δ completes the proof. □

Theorem 1. *The proxy optimal function of the problem (3) concentrates around the true optimal value with exponentially high probability, i.e.,*

$$\mathbb{P}\{|f(\phi,\boldsymbol{\Theta}^*)-f(\varphi_{\mathcal{H}},\boldsymbol{\Theta}^*_{\mathcal{H}})|<c_1(n\tau+n^2\tau^2c_2)\}$$
$$\ge 1-2\exp(-r\tau^2/32) \tag{9}$$

where $\boldsymbol{\Theta}_{\mathcal{H}}{}^*$ denotes the optimizer of $f(\varphi_{\mathcal{H}},\boldsymbol{\Theta})$, and c_1,τ,c_2 are positive constants.

Proof. Consider $f_F(\phi,\boldsymbol{\Theta})\triangleq\|\phi(\mathbf{X})-\phi(\mathbf{X})\boldsymbol{\Theta}\|_F+\lambda\|\boldsymbol{\Theta}^T\|_{1,l}$. The following holds uniformly over $\boldsymbol{\Theta}$.

$$|f_F(\phi,\boldsymbol{\Theta})-f_F(\varphi_{\mathcal{H}},\boldsymbol{\Theta})|=\left|\operatorname{tr}\left\{\hat{\mathbf{K}}-\mathbf{K}+\boldsymbol{\Theta}^T(\hat{\mathbf{K}}-\mathbf{K})\boldsymbol{\Theta}-2(\hat{\mathbf{K}}-\mathbf{K}\boldsymbol{\Theta})\right\}\right|$$
$$\le\left|\operatorname{tr}\left\{\boldsymbol{\Delta}\right\}\right|+\left|\operatorname{tr}\left\{\boldsymbol{\Delta}\right\}\boldsymbol{\Theta}\boldsymbol{\Theta}^T\right|+2\left|\operatorname{tr}\left\{\boldsymbol{\Delta}\right\}\boldsymbol{\Theta}\right|$$
$$\le\left|\operatorname{tr}\left\{\boldsymbol{\Delta}\right\}\right|+\sqrt{\operatorname{tr}\left\{\boldsymbol{\Delta}^T\boldsymbol{\Delta}\right\}}\left(\|(\boldsymbol{\Theta}^T\boldsymbol{\Theta})^2\|_F+2\|\boldsymbol{\Theta}^T\boldsymbol{\Theta}\|_F\right)$$
$$\le\left|\operatorname{tr}\left\{\boldsymbol{\Delta}\right\}\right|+\sqrt{\operatorname{tr}\left\{\boldsymbol{\Delta}^T\boldsymbol{\Delta}\right\}}\left(\|\boldsymbol{\Theta}\|_F^4+2\|\boldsymbol{\Theta}\|_F^2\right) \tag{10}$$

where $\mathbf{K}$ and $\boldsymbol{\Delta}$ denote the matrix of pairwise evaluations of the kernel function k, and the difference matrix $|\hat{\mathbf{K}}-\mathbf{K}|$, respectively. Since both objective functions are coercive and lower-bounded by zero, the global optimizer of $f(\phi,\cdot)$ and $f(\varphi_{\mathcal{H}},\cdot)$ are attained. Hence, both $\|\boldsymbol{\Theta}^*\|_F$ and $\|\boldsymbol{\Theta}^*_{\mathcal{H}}\|_F$ should be bounded

by a positive number $\sqrt{\nu}$. Then the bound in (10) can be written as

$$\left|\|\phi(\mathbf{X})-\phi(\mathbf{X})\mathbf{\Theta}^*\|_F-\|\varphi_{\mathcal{H}}(\mathbf{X})-\varphi_{\mathcal{H}}(\mathbf{X})\mathbf{\Theta}^*_{\mathcal{H}}\|_F\right|\leq\left|\operatorname{tr}\left\{\mathbf{\Delta}\right\}\right| +\sqrt{\operatorname{tr}\left\{\mathbf{\Delta}^T\mathbf{\Delta}\right\}}(\nu^2+\nu)\leq n\tau+n^2\tau^2c_2 \quad (11)$$

with probability at least $1-2\exp(-r\tau^2/32)$, where $c_2=\nu^2+\nu$ is a constant, and the last inequality is a result of a union bound on the concentration bound of Lemma 2.

Note that since any RKHS is separable, it is isometric to either $\mathbb{R}^m$ for a finite m, or the space of square summable sequences, i.e. ℓ^2 [42]. Herein, the arguments consider the case where the transformed space is of infinite dimensions, and as shown later, the finite-dimensional case follows with no extra effort.

We will work with the equivalent ℓ^2 space with the standard orthonormal basis $\{\mathbf{e}_i\}_{i=1}^{\infty}$. We have already shown that $\phi(\mathbf{x})-\phi(\mathbf{X})\theta\in\ell^2$, but we know that for the problem to be well-defined, it also needs to be in ℓ^1 space at optimality. We will denote the sequence at optimality by $\mathbf{h}^*\triangleq\sum_{t=1}^{n}\phi(\mathbf{x}_t)-\phi(\mathbf{X})\theta_t^*$. Note that $\|\cdot\|_2<\|\cdot\|_1$, hence, ℓ^1 is a subset of ℓ^2, but as it is dense in ℓ^2 [42], this does not impose a restrictive condition on the feasible space.

The ℓ_1-norm of the optimal sequence $\mathbf{h}^*$ introduces a convergent series, for which the convergence theorems guarantee the existence of an integer μ, such that for a given tolerance ϵ, $\sum_{i>\mu}|\mathbf{h}^*(i)|<\epsilon$ [57]. Then, $\|\mathbf{h}^*\|_1$ can be upper-bounded by

$$\|\mathbf{h}^*\|_1=\left\|\sum_{i=1}^{\infty}\mathbf{h}^*(i)\mathbf{e}_i\right\|_1=\left\|\sum_{i=1}^{\mu}\mathbf{h}^*(i)\mathbf{e}_i\right\|_1+\epsilon\leq\sum_{i=1}^{\mu}|\mathbf{h}^*(i)|\|\mathbf{e}_i\|_1+\epsilon \leq\sqrt{\sum_{i=1}^{\mu}|\mathbf{h}^*(i)|^2}\sqrt{\sum_{i=1}^{\mu}\|\mathbf{e}(i)\|_1^2}+\epsilon=\|\mathbf{h}^*\|_2\sqrt{\mu}+\epsilon\,. \quad (12)$$

Letting $c=\mu+\epsilon/\|\mathbf{h}^*\|_2^2$ yields the desired constant, hence, showing the boundedness of the ℓ_1-norm for the optimal point of the induced RKHS by a constant factor of its ℓ_2-norm, and the proof is complete by the equivalence of these two norms at optimality. □

Remark 3. The analysis in the proof of Theorem 1 is applicable to kernels whose induced RKHS are finite dimensional, at no extra effort. The arguments follow closely the proof above, where the equivalence of the Hilbert space with the finite dimensional Euclidean space is considered, and the ℓ_1-norm of the RKHS vector at optimality is simply a finite sum.

Remark 4. As explained in Section 2, the mapped features in the Hilbert space are functions, and the choice of the orthonormal basis maps them to the Euclidean space or the space of infinite square-summable sequences. For many cases though, specific choices of the basis connect to well-known fundamental concepts in signal processing fields. We illustrate such an example in the following. Consider a stationary kernel $(k(\mathbf{x}_i,\mathbf{x}_t)=\kappa(\boldsymbol{\delta})$, where $\boldsymbol{\delta}:=\mathbf{x}_i-\mathbf{x}_t)$ which is defined over the interval $[0,2\pi]^m$, such that it can be extended to a symmetric and periodic function on $\mathbb{R}^m$. Then, if we choose the Fourier basis for the induced function space, an element of the RKHS can be written by the Fourier series of this function as $\sum_{\mathbf{i}}\hat{\kappa}_{\mathbf{i}}\exp(j2\pi\mathbf{i}^T\boldsymbol{\delta})$, where the vector $\mathbf{i}=(\mathbf{i}(1),\mathbf{i}(2),\ldots,\mathbf{i}(m))$ is the integer lattice in $\mathbb{R}^m$, typically known as $\mathbb{Z}^m$, and the Fourier coefficients can be obtained by $\hat{\kappa}_{\mathbf{i}}=\int_{[0,2\pi]^m}\exp(-j2\pi\mathbf{i}^T\boldsymbol{\delta})\kappa(\boldsymbol{\delta})d\boldsymbol{\delta}$. A one-dimensional example boils down to the familiar Fourier basis as the orthogonal functions of $\{1,\sin(x),\cos(x),\sin(2x),\cos(2x),\ldots\}$. Then, the mapped ℓ^2 sequence is nothing but the Fourier coefficients, and the condition on the optimality can be translated to the Truncated Fourier series of the original function for infinite dimensional spaces. With Fourier coefficients going to zero at a faster speed, μ is smaller and the ℓ_1 norm is closer to the ℓ_2 norm. In the finite-dimensional case, only the elements of the space corresponding to the finite non-zero coefficients are present in constructing the transformed feature, which defines the geometry of the associated RKHS.

Next, to characterize the representative subset we select, consider the following. We denote the symmetrized convex hull of all transformed samples by $\mathcal{P}(\mathbf{\Phi})$. Also, let the positive ray of a vector $\mathbf{v}$ be $\bar{\mathbf{v}}=\{t\mathbf{v}:t>0\}$.

Theorem 2. *When the representation constraint in (3) is fully enforced, our method samples the vertices of $\mathcal{P}(\mathbf{\Phi})$ as representatives.*

Proof. Consider the following definitions. Given the convex hull $\mathcal{P}(\mathbf{\Phi})$, its polar set is defined as

$$\mathcal{P}^o=\{\mathbf{h}\in\mathcal{H}:\langle\mathbf{y},\mathbf{h}\rangle\leq 1,\ \forall\mathbf{y}\in\mathcal{P}(\mathbf{\Phi})\}\,. \quad (13)$$

Also, we call the face of the convex hull passing through the positive ray of $\mathbf{y}$, the *closest face* to $\mathbf{y}$. Now, note that when regularizing by the ℓ_1-norms of the encoding matrix, our objective function is decomposable to columns of the optimization variable, and hence, the minimization can be done in a separate fashion.

$$\min_{\theta_t}\|\theta_t\|_1$$
$$\text{s.t. }\sin(\boldsymbol{\zeta}^T\mathbf{x}_t+\beta)=\sin(\boldsymbol{\zeta}^T\mathbf{x}_t+\beta)\boldsymbol{\theta}_t\ \forall t\in\mathbb{N}_n \quad (14)$$

When the reconstruction constraint is fully enforced, the dual can be obtained as the following linear program

$$\max_{\mathbf{h}\in\mathcal{H}}\langle\sin(\boldsymbol{\zeta}^T\mathbf{x}_t+\beta),\mathbf{h}\rangle$$
$$\text{s.t. }\|\mathbf{\Phi}^T\mathbf{h}\|_\infty\leq 1\,. \quad (15)$$

The constraint in (15) can be equivalently expressed as $\mathbf{h}$ belonging to the polar set of the convex hull.

Guaranteed by the strong duality, the optimal value of each original optimization problem in (14) is equal to that of (16).

$$\max_{\mathbf{h}\in\mathcal{P}^o}\langle\sin(\boldsymbol{\zeta}^T\mathbf{x}_t+\beta),\mathbf{h}\rangle\,. \quad (16)$$

This problem can be easily solved using linear programming techniques [58,59]. Using the aforementioned definitions, the problem is equivalent to finding the closest faces of the convex hull of transformed dataset to the given point $\sin(\boldsymbol{\zeta}^T\mathbf{x}_t+\beta)$. The extreme points of this face are indicated by the indices with non-zero entries. Finally, note that this holds for all columns of the encoding matrix as shown in (14), through which, the vertices of the convex hull are identified. □

Specifically, by solving the optimization problem with zero reconstruction error constraint, we indeed find those faces of the polytope $\mathcal{P}(\mathbf{\Phi})$ which intersect the positive rays of transformed data points, $\{t\phi(\mathbf{x}):t>0,\ \forall\mathbf{x}\in\mathcal{X}\}$. Also, the vertices of this polytope, being the extreme points of the aforementioned faces are indicated by the rows of the optimal encoding matrix $\mathbf{\Theta}^*$ with non-zero norms. Sampling the vertices of this convex hull meets our intuition in restoring the information of the dataset via few samples. Vertices are considered as the most critical points of a convex body; while they are not reconstructible by the others, any other point of the set can be represented by a convex combination of these points.

Finally, note that enforcing the reconstruction error to be zero does not inactivate the regularization, rather provides the most interesting case to analyze theoretically. Indeed, if the problem has a feasible representative subset, the transformed data matrix $\mathbf{\Phi}$ is implied to be inherently low-rank. Consequently, there exist multiple potential encoding matrices which result in precisely zero reconstruction error, and the choice of regularizer rules one of those candidates as the optimal solution of the program. As it will be

clear from the proof of the theorem, when $\lambda = \infty$, the regularizer actually plays as the main cost function of the problem, constrained by the precise reconstruction.

Corollary 1. *The optimal solution of our program in Theorem 2 is of the form*

$$\mathbf{\Theta}^* = \begin{bmatrix} \tilde{\mathbf{\Theta}} & \mathbf{I}_v \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{T} \tag{17}$$

where v is the number of vertices of the convex hull $\mathcal{P}(\mathbf{\Phi})$, and $\tilde{\mathbf{\Theta}}$ is a non-zero sub-matrix.

Proof. By definition, every point inside the convex hull can be reconstructed by a linear combination of the vertices, i.e. the chosen representatives. This reconstruction results in the encoding sub-matrix $\tilde{\mathbf{\Theta}}$. On the other hand, each vertex can only be written as its own coordinates by an encoding of 1, which gives rise to the $\mathbf{I}_v$ sub-matrix. □

4. Experiments

In this section, we conduct a set of experiments using both synthetic and real data for various applications of the proposed method, and also compare its performance to state-of-the-art sketching algorithms. In the first set of experiments, we study representative sampling from a database of face images (Section 4.3). We provide both a qualitative and quantitative analysis of the ability of different methods to select a balanced set of images, sample sufficiently from different face expressions and poses, and pick faces with minimal shading/occlusion effects. The second set of experiments in Section 4.4 focuses on fair sampling from group-forming and clustered data; we assess if our method samples sufficiently from each group/cluster in data with multiple overlapping manifold structures of various dimensions and sparse corruption, and yields a balanced sub-sample. In Section 4.5, we consider the downstream task of classification using representatives in various domains, including face identification and hand-written digit recognition. The classification model is trained on a small subset of the training data selected using different selection methods, then the performance is evaluated on the original test data. In the fourth set of experiments, we investigate the application of our model in the task of face pose generation, where the goal is to generate new poses/angles of the face images using a set of observed images from multiple views (See Section 4.6). The pipeline consists of choosing representatives of the face images of each subject using different selection methods, training a deep generative model using only said representatives, and evaluating the performance of each trained model on the test set. We make use of a two-path GAN architecture [60], and provide samples of the generated images and each method's average identity error. Lastly, we compare the running time of different algorithms for a different number of data points, which confirms the computational superiority of the proposed algorithm.

4.1. Data

We use both real and synthetic manifold datasets for the experiments. *5Spirals* refers to a 5-class 10-dimensional artificial dataset with 500 points sampled from surfaces of arithmetic 5 spirals, sampled from an involute of a circle with parametric equations $x_1 = \cos\phi + \phi\sin\phi$ and $x_2 = \sin\phi - \phi\cos\phi$, by a uniform distribution of the angle ϕ. The manifolds are then rotated, shifted, and embedded in the ambient space. Similarly, *3Spirals-2Knots* $\in \mathbb{R}^{10\times 500}$ is another 5-class synthetic data with 300 points from three spiral manifolds, and the other two from embedded Trefoil Knots, where 100 points from each knot are sampled uniformly over the parameter θ of the curve $[x_1, x_2, x_3] = [\sin\theta + 2\sin 2\theta, \cos\theta - 2\cos 2\theta, -\sin 3\theta]$. Lastly, *Sphere-SwissRoll* contains 30-dimensional points that lie on one of the two low-dimensional manifolds of Sphere (100 points uniformly sampled over θ and ϕ from the curve $[x_1, x_2, x_3] = [\cos\theta\sin\phi, \sin\theta\sin\phi, \cos\phi]$) and SwissRoll, constructed from 150 points on arithmetic spirals with three different heights. We randomly replace 5 percent of the data with random values in the data range, i.e., we add a sparse matrix $\mathbf{S}$ with $p = 0.05$ to the data points to affect gross corruption in the data. As for the real data, the *Extended Yale Database* [61] consists of face images of 38 human subjects, with 64 different illuminations per person. We have resized the images to 32×32. *Multiple Features* is another real dataset from the UCI repository [62], consisting of 1000 data points, each with 649 features of handwritten digits, covering their different characterizations. Similarly, the *MNIST* dataset [63] contains images of hand-written images. Lastly, *Multi-pie* Face Database [64] contains images of 250 persons under various variations of illumination (20 settings), pose (13 angles), and expression (4 sessions). In the experiments, the preprocessed data is used with 128×128 images of two expressions from the first session, under all lighting and pose variations.

4.2. Experimental setup

We compare against state-of-the-art methods that can handle non-linear data relations and are widely used in the related literature, including Spec [39], MKM [41], Kmed [34], SRS [27], and DS3 [17]. For these methods, the parameters were set as suggested by their authors, while for the MoSSaRT, we avoid the use of data-dependent feature specifications as well as hyper-parameter tuning, by only using the two generic kernel functions, namely the Gaussian kernel, $k(\mathbf{x}, \mathbf{x}') = \exp\frac{-\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\sigma^2}$, and the Laplacian kernel $k(\mathbf{x}, \mathbf{x}') = \exp\frac{-\|\mathbf{x}-\mathbf{x}'\|_1}{\sigma}$, with spread parameter σ. For σ, we adopt the standard procedure of choosing it from $\{0.1, 1, 10\}$ of average data variance. The datasets are randomly split to training, testing, and validation sets by $70-20-10$ ratios. We choose the kernel/parameter settings over their performance on the validation set, and the results are averaged over 50 runs.

4.3. Representatives of face poses

Here, we present our findings on experimental results of the *Multi-PIE* dataset as a case study for our methodology. In Fig. 2a selected images from 520 images of a subject based on different selection methods are displayed in different rows. Some interesting observations can be made from this visualization. First, other methods tend to sample dominantly from one expression (smile) versus the other (neutral). Other methods fall short of preserving the representation space of the expressions as evidenced by the unbalanced selection from the different expressions. In sharp contrast, MoSSaRT selects sufficient samples from both gestures, covering the whole space of different expressions. Second, as our model tackles corruption, it avoids the misinformation caused by the lighting conditions, by selecting the illumination with minimal shading/occlusion effects. Third, except our method, others choose samples that focus on specific angles, hence, deteriorating their representation power. MoSSaRT, however, selects samples that span the angle space of the poses in a balanced way. To evaluate this feature quantitatively, we allow the methods to select 13 samples from each subject for 50 randomly chosen subjects, and report the average value for selected angle intervals by each method in Fig. 2b. The even distribution of our results as opposed to the others' validates our visual observations.

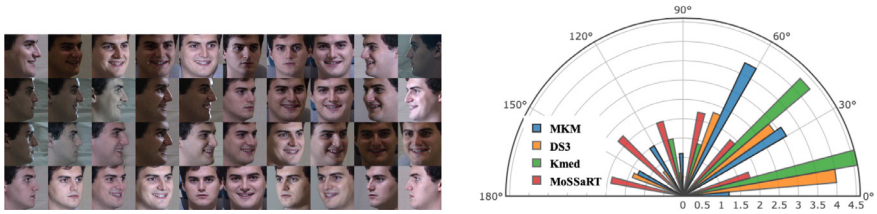

Fig. 2. Left: Visualization of representatives of a subject in *Multi-pie* dataset selected by different algorithms (MKM, DS3, KMed, and MoSSaRT, from top to bottom row). Right: Average number of selected images based on their view angle. Only our method selects samples from diverse angles evenly.

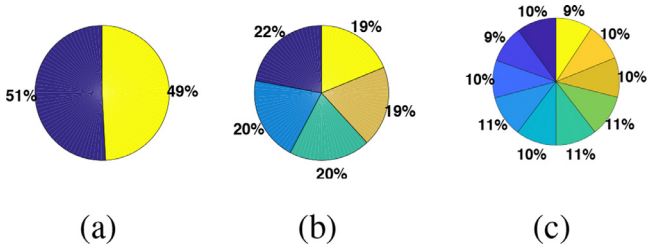

Fig. 3. Percentage of sampled representatives from different clusters. Our method fairly samples from different clusters for (a) *Sphere-SwissRoll*, (b) *3Spirals-2Knots*, and (c) for *Multiple Features* dataset.

4.4. Fair sampling of clustered data

One expects a representative subset to capture key characteristics of a collection. Group-forming collections are some of the widely encountered data types that arise in clustering problems. In this setting, the ability to contain sufficient information forming the clusters becomes pivotal, which can be significantly distorted by gross corruptions. Many clustering algorithms exhibit considerably better performance when the input data are clustered in balanced sizes [65]. Here, we experiment how our sampling scheme can fairly sample from multiple grossly corrupted clusters. For a more challenging scenario, the sizes of the clusters are chosen to be different. More specifically, *3Spirals-2Knots* contains 5 equally-sized clusters, and the number of points in the two clusters of *Sphere-SwissRoll* are 100–150. For the *Multiple Features* dataset, we randomly choose three groups of digits of the size $3-3-4$, and set their corresponding cluster sizes to $500-750-1000$ data points. Then, our representative selection algorithm is applied to the obtained datasets, and the percentage of chosen samples from each cluster is illustrated in Fig. 3. The results show that our method is capable of fairly sampling from both balanced and unbalanced clusters for different synthetic and real datasets.

4.5. Training classifiers on reduced sketches

Having in mind the burden of classification tasks for large datasets, we consider the problems of face identification, handwritten digit recognition, as well as classification of synthetic datasets by training the classifier only using the chosen representatives. To this end, 10% of the training set is first selected by different selection methods. Using these reduced sketches, we then train a SVM model [66] on those subsets and evaluate their performance on the original test sets for multiple real and synthetic datasets. We also include the results for training the classifier with complete training sets as guidelines, denoted by "No selection". The comparison of classification accuracy trained on samples obtained from our method vs. other sampling methods is shown in Table 1. Confirmed by its lowest classification errors, MoSSaRT achieves the best performance uniformly over all datasets. While offering significant savings in the computational/storage requirements, one can infer that the sampled chosen by our proposed method are indeed good representatives of the whole collection, as our performance is close to (if not better than) training the model with the full training set. In fact, since MoSSaRT is designed to handle gross corruptions, it exhibits considerable improvements over the full training set in most cases.

4.6. Face pose generation

This experiment investigates the problem of generating face images from multiple views. Despite marked success in computer vision and robotics, face recognition in a pose-invariant manner remains a challenging problem, mainly due to the performance degradation caused by variability in pose, illumination, and noise. Here, to examine the effectiveness of the selected representatives in such a challenging setting, we train a deep generative model only on the sampled representatives, and evaluate its performance on the test set. Regular GANs are tuned to generate realistic instances but may learn incomplete representations due to their use of a single-pathway architecture, which consists of a generator (encoder E and decoder G) and a discriminator (D). As such, we make use of the two-path GAN architecture (CR-GAN) in Tian et al. [60], which introduces a second path in order to learn complete representations for multi-view generation. The idea is to use a second generation sideway to create view-specific images from randomly sampled embeddings (see Fig. 4 and its caption). By encoding the complement space of the first path, a complete representation space for the generator is obtained, so that more realistic outputs can be generated from single-view inputs.

In our experiment, we first use different sampling methods to draw a small representative subset of 13 images for each subject in the training data, each forming a different reduced training set. Then, a CR-GAN network is trained for 300 epochs on each of these reduced training sets, and the models are then evaluated using the same test examples. Fig. 4 illustrates a diagram of our pipeline in this experiment. For more details on the implementation and training of a CR-GAN we refer the reader to Tian et al. [60]. Fig. 5 shows a visual comparison of the images generated by the GAN models trained on different sets of representatives in different rows. The first row contains the output of the model trained by the samples chosen by our proposed method, and the next ones correspond to SRS, Kmed, DS3, Spec, and MKM, from top to bottom. Clearly, the results produced by the proposed method MoSSaRT are visually more appealing and realistic, testifying that our chosen samples are indeed better representatives of the whole training set. Others on the other hand, suffer from artifacts such as checkerboard, corrupt pixels, posterizing, blurring and ringing effects, which result in images that are visually less pleasant and perceptually less convincing for the human viewer. Among others, the two methods Spec and MKM generated better looking images, which may be caused by their manifold-specific approaches. Note that since all the training details including architecture, loss functions, and hyper-parameters are set the same for

Table 1
Comparison of classification accuracy for classifiers trained on reduced subsets obtained from different methods on various datasets."No selection" corresponds to the results of training the classifiers on the complete datasets. Our proposed method MoSSaRT outperforms all other methods.

Data \ Method	No selection	MoSSaRT	DS3	Kmed	MKM	SRS	Spec
Sphere-SwissRoll	0.725	0.800	0.725	0.700	0.800	0.775	0.625
5Spirals	0.450	0.447	0.370	0.430	0.430	0.300	0.340
Digits	0.523	0.610	0.575	0.355	0.523	0.418	0.548
MNIST	0.850	0.870	0.718	0.825	0.760	0.737	0.688
Yale	0.660	0.613	0.519	0.576	0.506	0.551	0.432

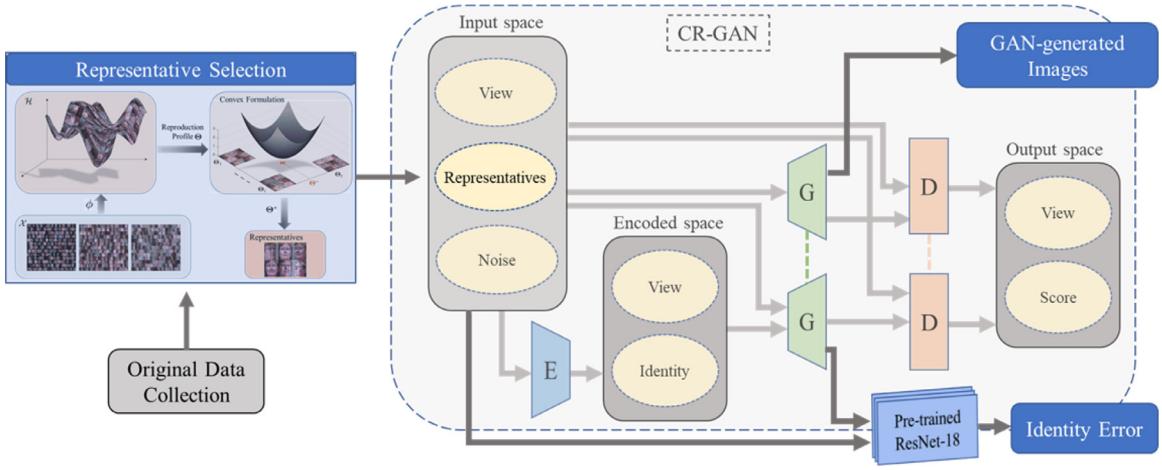

Fig. 4. Pipeline of the trained model for the experiment of Section 4.6, consisting of a two-way GAN architecture (CR-GAN) in the faded box, followed by the feature extraction by a pre-trained 18-layer ResNet for Identity Error calculation (reported in Table 2), and generation of synthetic face images in various angles (shown in Fig. 5), without the identity-preserving constraint. The upper generation path trains a generator G and discriminator D to produce realistic images. G generates images from random noise (without an encoder) to complete the latent space of the lower reconstruction path of a standard GAN, consisting of encoder E, decoder G, and discriminator D. The dashed-lines between the two generators and the two discriminators indicate weight-sharing.

Fig. 5. GAN generated images trained on samples obtained from different selection methods. From top to bottom row: MoSSaRT, SRS, Kmed, DS3, Spec, and MKM. As it chooses better representatives, MoSSaRT results in more photo-realistic outputs compared to the others.

all cases, these varied qualities can be solely traced back to the differences of the chosen samples by different methods. Moreover, we take a step further to avoid the subjective comparison of qualitative results, and monitor the identity error of a generated view for a given image. This error indicates the Euclidean distance between the features of real and generated images. For a given image, we extract a 256-dimensional feature vector from a 18-layer ResNet model [67] pre-trained on MS-Celeb-1M, a large-scale real world face dataset [68]. The reported results in the first row of Table 2 correspond to the average value and standard deviation of the normalized identity errors over the test set. As this error illustrates how close a generated image is to its real version, the lower value of the error with MoSSaRT indicates its better performance in generating more realistic images from a given pose. This in turn suggests the capability of our method in selecting more informative representatives, which give rise to a better trained GAN model.

4.7. Running time comparison

Lastly, we illustrate the efficiency of the developed algorithm on how scalable it is in the data size. Two subsets of the *Multi-PIE* dataset of size 1000 and 5000 are randomly selected, and multiple selection algorithms are run to select 13 samples from each subset. We report average running time of each algorithm over 50 runs in the last two rows of Table 2. For these experiments, a X64 machine with 2.4 GHz CPU and 32 GB RAM is used. While the ADMM algorithm of DS3 is faster than a general convex solver such as CVX, as can be seen from its run-time for 1000 samples, this algorithm is too computationally expensive (approximately $\mathcal{O}(n^3)$), hence the experiment with the larger subset of $n = 5000$ was intractable to run for this method. Among others, MoSSaRT demonstrates much faster running time (except Spec), illustrating our algorithm's higher efficiency. These results also validate our complexity analysis of near-linear computational complexity in terms of number of data points ($\mathcal{O}(n^{1.366})$). While Spec has sightly lower run-time that our algorithm, as shown in Tables 1 and 2, it considerably falls behind our method w.r.t. other performance measures.

Table 2
Performance analysis of various sampling methods on the *Multi-PIE* dataset. First row: Average (± standard deviation) normalized identity error on the test set for face pose generation. The GAN models are trained on reduced training sets (13 per subject) obtained by different sampling methods. Second row: Average runtime of sampling algorithms for two different numbers of data points ($n = 1000, 5000$).

Metric \ Method	MoSSaRT	SRS	Kmed	DS3	Spec	MKM
GAN Identity Error	0.537 ± 0.194	0.674 ± 0.209	0.632 ± 0.205	0.625 ± 0.186	0.618 ± 0.229	0.613 ± 0.231
Runtime ($n = 1000$) (s)	17.14	39.93	25.46	887.72	10.69	24.32
Runtime ($n = 5000$) (s)	157.08	1612.35	776.64	–	154.17	1685.72

5. Conclusion

Informative representatives allow for substantial computational and storage conservations. This paper tackles important limitations of existing methods under realistic and practical scenarios. More specifically, the proposed method is the first approach that offers the following advantages simultaneously: (i) ability to account for a versatile set of qualities in the chosen subset including representativeness, novelty, and conciseness, (ii) a global understanding of the prevailing non-linear manifold structures in high-dimensional data, (iii) robustness to gross sparse corruptions in non-linear settings, (iv) provable guarantees and interpretability, (v) computationally efficient and scalable implementation. We developed an approach tailored for non-linear manifold data without the use of local information or complex algebraic iterations. An approximate explicit transformation was built upon an implicit feature mapping of a desired RKHS. Based on the introduced reproduction profile, our formulation gave rise to a parallelizable convex minimization whose optimal solution provides a concise encoding of the data facilitating the realization of the aforementioned criteria. Finally, experiments on both synthetic and real datasets showed that our method improves upon the state-of-the-art on the problems of face identification, hand-written digit recognition, face pose generation using GANs, classification of artificial data, and running time analysis.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by NSF CAREER Award CCF-1552497, NSF Awards CCF-2106339, 1643835, 1161228, and 1356233, and the CSIT TEAm BOG grant.

References

[1] K. Pearson, LIII. On lines and planes of closest fit to systems of points in space, Lond., Edinb., Dublin Philos. Mag. J. Sci. 2 (11) (1901) 559–572.
[2] J.B. Tenenbaum, V. De Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[3] K. Zheng, X. Wang, Feature selection method with joint maximal information entropy between features and class, Pattern Recognit. 77 (2018) 20–29.
[4] P. Zhou, L. Du, X. Li, Y.-D. Shen, Y. Qian, Unsupervised feature selection with adaptive multiple graph learning, Pattern Recognit. (2020) 107375.
[5] X. Luo, Y. Xu, J. Yang, Multi-resolution dictionary learning for face recognition, Pattern Recognit. 93 (2019) 283–292.
[6] Z. Zheng, H. Sun, Jointly discriminative projection and dictionary learning for domain adaptive collaborative representation-based classification, Pattern Recognit. 90 (2019) 325–336. http://www.sciencedirect.com/science/article/pii/S003132031930010X
[7] A. Kulesza, B. Taskar, *k*-DPPs: fixed-size determinantal point processes, in: Proceedings of the 28th International Conference on Machine Learning, 2011, pp. 1193–1200.
[8] S. Joshi, S. Boyd, Sensor selection via convex optimization, IEEE Trans. Signal Process. 57 (2) (2013) 451–462.
[9] W. Chen, Z. Yang, J. Ren, J. Cao, N. Cai, H. Zhao, P. Yuen, MIMN-DPP: maximum-information and minimum-noise determinantal point processes for unsupervised hyperspectral band selection, Pattern Recognit. 102 (2020) 107213. http://www.sciencedirect.com/science/article/pii/S0031320320300194
[10] V. Oculus, Oculus rift-virtual reality headset for 3D gaming(2020). https://www.oculus.com/.
[11] M. Gottmer, Merging reality and virtuality with MS hololens(2020).
[12] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[13] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations (ICLR), 2015.
[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in Neural Information Processing Systems, 2014.
[15] S.R. Kundur, D. Raviv, Active vision-based control schemes for autonomous navigation tasks, Pattern Recognit. 33 (2) (2000) 295–308. http://www.sciencedirect.com/science/article/pii/S0031320399000588
[16] E. Widyaningrum, R.Y. Peters, R.C. Lindenbergh, Building outline extraction from ALS point clouds using medial axis transform descriptors, Pattern Recognit. 106 (2020) 107447. http://www.sciencedirect.com/science/article/pii/S0031320320302508
[17] E. Elhamifar, G. Sapiro, S.S. Sastry, Dissimilarity-based sparse subset selection, IEEE Trans. Pattern Anal. Mach. Intell. 38 (11) (2016) 2182–2197.
[18] E.J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis? J. ACM 58 (3) (2011) 11:1–11:37, doi:10.1145/1970392.1970395.
[19] Y. Chen, C. Caramanis, S. Mannor, Robust sparse regression under adversarial corruption, in: Int. Conference on Machine Learning, 2013, pp. 774–782.
[20] A. Rahimi, B. Recht, Random features for large-scale kernel machines, in: Advances in Neural Information Processing Systems, 2008, pp. 1177–1184.
[21] G. Golub, Numerical methods for solving linear least squares problems, Numer. Math. 7 (3) (1965) 206–216, doi:10.1007/BF01436075.
[22] T.F. Chan, Rank revealing QR factorizations, Linear Algebra Appl. 88 (1987) 67–82.
[23] M. Gu, S. Eisenstat, Efficient algorithms for computing a strong rank-revealing QR factorization, SIAM J. Sci. Comput. 17 (4) (1996) 848–869.
[24] M.W. Mahoney, Randomized algorithms for matrices and data, Found. Trends Mach. Learn. 3 (2) (2011) 123–224, doi:10.1561/2200000035.
[25] P. Drineas, M.W. Mahoney, S. Muthukrishnan, Relative-error CUR matrix decompositions, SIAM J. Matrix Anal. Appl. 30 (2) (2008) 844–881.
[26] J.A. Tropp, Column subset selection, matrix factorization, and eigenvalue optimization, in: Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms, 2009, pp. 978–986.
[27] E. Elhamifar, G. Sapiro, R. Vidal, See all by looking at a few: sparse modeling for finding representative objects, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1600–1607.
[28] L. Balzano, R. Nowak, W. Bajwa, Column subset selection with missing data, NIPS Workshop on Low-Rank Methods for Large-Scale Machine Learning, 2010.
[29] E. Esser, M. Moller, S. Osher, G. Sapiro, J. Xin, A convex model for nonnegative matrix factorization and dimensionality reduction on physical space, IEEE Trans. Image Process. 21 (7) (2012) 3239–3252.
[30] H. Wang, Y. Kawahara, C. Weng, J. Yuan, Representative selection with structured sparsity, Pattern Recognit. 63 (2017) 268–278.
[31] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge university press, 2004.
[32] M. Shamaiah, S. Banerjee, H. Vikalo, Greedy sensor selection: leveraging submodularity, in: 49th IEEE Conference on Decision and Control (CDC), 2010, pp. 2572–2577.
[33] G.L. Nemhauser, L.A. Wolsey, M.L. Fisher, An analysis of approximations for maximizing submodular set functions I, Math. Program. 14 (1) (1978) 265–294.
[34] L. Kaufman, P. Rousseeuw, Clustering by Means of Medoids, North-Holland, 1987.
[35] M. Charikar, S. Guha, . Tardos, D.B. Shmoys, A constant-factor approximation algorithm for the *k*-median problem, J. Comput. Syst. Sci. 65 (1) (2002) 129–149. http://www.sciencedirect.com/science/article/pii/S0022000002918829
[36] B.J. Frey, D. Dueck, Mixture modeling by affinity propagation, in: Advances in Neural Information Processing Systems, MIT Press, 2006, pp. 379–386.
[37] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (5814) (2007) 972–976.
[38] N. Shroff, P. Turaga, R. Chellappa, Manifold precis: an annealing technique for

diverse sampling of manifolds, in: Advances in Neural Information Processing Systems, 2011, pp. 154–162.
[39] A.C. Öztireli, M. Alexa, M. Gross, Spectral sampling of manifolds, ACM Transactions on Graphics (TOG), vol. 29, ACM, 2015.
[40] M. Sedghi, M. Georgiopoulos, G. Atia, A multi-criteria approach for fast and robust representative selection from manifolds, IEEE Trans. Knowl. Data Eng. (Early Access) (2020), doi:10.1109/TKDE.2020.3024099. 1–1.
[41] E. Tu, L. Cao, J. Yang, N. Kasabov, A novel graph-based *k*-means for nonlinear manifold clustering and representative selection, Neurocomputing 143 (2017) 109–122. http://www.sciencedirect.com/science/article/pii/S0925231214007565
[42] G.B. Folland, Real Analysis: Modern Techniques and Their Applications, Wiley, 2013. https://books.google.com/books?id=wI4fAwAAQBAJ
[43] L. Wu, I.E.H. Yen, J. Chen, R. Yan, Revisiting random binning features: fast convergence and strong parallelizability, in: 22nd ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining, New York, NY, 2016, pp. 1265–1274, doi:10.1145/2939672.2939794.
[44] A. Vedaldi, A. Zisserman, Efficient additive kernels via explicit feature maps, IEEE Trans. Pattern Anal. Mach. Intell. 34 (3) (2012) 480–492.
[45] M. Raginsky, S. Lazebnik, Locality-sensitive binary codes from shift-invariant kernels, Advances in Neural Information Processing Systems, 2009.
[46] R. Hamid, Y. Xiao, A. Gittens, D. Decoste, Compact random feature maps, in: Int. Conference on Machine Learning, Bejing, China, vol. 32, 2014, pp. 19–27. http://proceedings.mlr.press/v32/hamid14.html
[47] F. Liu, X. Huang, Y. Chen, J.A.K. Suykens, Random features for kernel approximation: a survey in algorithms, theory, and beyond, arXiv preprint arXiv:2004.11154
[48] W. Rudin, Fourier Analysis on Groups, vol. 121967, Wiley Online Library, 1962.
[49] M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming, version 2.1, 2014,
[50] M. Grant, S. Boyd, Graph implementations for nonsmooth convex programs, in: Recent Advances in Learning and Control, Lecture Notes in Control and Information Sciences, Springer-Verlag Limited, 2008, pp. 95–110.
[51] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al., Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends® Mach. Learn. 3 (1) (2011) 1–122.
[52] V.V. Williams, Multiplying matrices in $\mathcal{O}(n^{2.373})$ time, preprint (2014).
[53] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms, third ed., The MIT Press, 2009.
[54] A. Gupta, V. Kumar, Scalability of parallel algorithms for matrix multiplication, in: International Conference on Parallel Processing, vol. 3, 1993, pp. 115–123.
[55] D.J. Sutherland, J. Schneider, On the error of random fourier features, in: Uncertainty in Artificial Intelligence, Arlington, Virginia, 2015, p. 862871.
[56] R. Vershynin, Four lectures on probabilistic methods for data science, Math. Data (2018) 231–271, doi:10.1090/pcms/025/05.
[57] W. Rudin, Principles of Mathematical Analysis, International Series in Pure and Applied Mathematics, McGraw-Hill, 1976. https://books.google.com/books?id=kwqzPAAACAAJ
[58] P.E. Gill, W. Murray, M. Wright, Numerical Linear Algebra and Optimization, vol. 1, Addison-Wesley, 1991.
[59] D.G. Luenberger, Y. Ye, Linear and Nonlinear Programming, Springer, 2015.
[60] Y. Tian, X. Peng, L. Zhao, S. Zhang, D.N. Metaxas, CR-GAN: learning complete representations for multi-view generation, in: International Joint Conference on Artificial Intelligence (IJCAI), 2018.
[61] A.S. Georghiades, P.N. Belhumeur, D.J. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 643–660.
[62] M. Lichman, UCI machine learning repository, 2013,
[63] Y. LeCun, C. Cortes, MNIST handwritten digit database(2010). http://yann.lecun.com/exdb/mnist/.
[64] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-PIE, Image Vis. Comput. 28 (5) (2010) 807–813.
[65] Y. Chen, A. Jalali, S. Sanghavi, H. Xu, Clustering partially observed graphs via convex optimization, J. Mach. Learn. Res. 15 (1) (2014) 2213–2238.
[66] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Pattern Classification and Scene Analysis: Pattern Classification, Wiley, 2001. https://books.google.com/books?id=YoxQAAAAMAAJ
[67] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, (2015). CoRR arXiv:1512.03385
[68] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, MS-Celeb-1M: a dataset and benchmark for large-scale face recognition, (2016). CoRR arXiv:1607.08221

Mahlagha Sedghi received her B.Sc. degree in Electrical Engineering from Sharif University of Technology, Tehran, Iran, in 2014, M.Sc. degree in Electrical Engineering from University of Central Florida (UCF), Orlando, FL, in 2017. She is currently pursuing her doctoral studies at UCF, working jointly in Machine Learning and Signal Processing Labs, with research interests including Manifold Learning and Unsupervised Learning and Kernel Methods. She is a recipient of Graduate Dean's fellowship and Daniel Hammond Engineering scholarship at UCF (2014–2018).

Michael Georgiopoulos received his Diploma in EE (National Technical University of Athens; 1981), and MS, Ph.D. in EE (University of Connecticut; 1983, 1986). He has been at the University of Central Florida (UCF) since 1986 and is currently a Professor in ECE. He serves as the dean of the College of Engineering and Computer Science (2013-present). His research expertise lies in machine learning with special emphasis on neural network algorithms and related applications. He has been involved in more than 50 grants and contracts, some of them of multi-million-dollar value. He has published and presented more than 270 papers in journals and conferences. He has served as Associate Editor of the Neural Networks journal and the IEEE Transactions on Neural Networks. In 2010, he was named a UCF Pegasus Professor, the most prestigious award bestowed to senior UCF faculty members for extraordinary contributions in teaching, research, and service.

George K. Atia received the B.Sc. and M.Sc. degrees from Alexandria University, Egypt, in 2000 and 2003, respectively, and the Ph.D. degree from Boston University, MA, in 2009, all in Electrical and Computer Engineering. He joined the University of Central Florida in Fall 2012, where he is currently an Associate Professor in the Department of Electrical and Computer Engineering. He was a Visiting Faculty at the Air Force Research Laboratory (AFRL) in 2019–2020. From 2009 to 2012, he was a Postdoctoral Research Associate at the Coordinated Science Laboratory (CSL) at the University of Illinois at Urbana-Champaign (UIUC). His research interests include machine learning and data analytics, statistical signal processing, stochastic control, detection and estimation theory, and information theory. He is an Associate Editor for the IEEE Transactions on Signal Processing. Dr. Atia is the recipient of many awards, including the UCF Reach for the Stars Award and the CECS Research Excellence Award in 2018, the Dean's Advisory Board Fellowship and the Inaugural UCF Luminary Award in 2017, the NSF CAREER Award in 2016, and the Charles Millican Faculty Fellowship Award (2015–2017).