

Feature space approximation for kernel-based supervised learning

We propose a method for the approximation of high- or even infinite-dimensional feature vectors, which play an important role in supervised learning. The goal is to reduce the size of the training data, resulting in lower storage consumption and computational complexity. Furthermore, the method can be regarded as a regularization technique, which improves the generalizability of learned target functions. We demonstrate significant improvements in comparison to the computation of data-driven predictions involving the full training data set. The method is applied to classification and regression problems from different application areas such as image recognition, system identification, and oceanographic time series analysis.

Keywords: supervised learning, kernel-based methods, feature spaces, dimensionality reduction, system identification

MSC: 68Q27, 68Q32, 68T09

Patrick Gelß

Department of Mathematics
and Computer Science,
Freie Universität Berlin,
Berlin 14195, Germany

Stefan Klus

Department of Mathematics,
University of Surrey,
Guildford GU2 7XH, UK

Ingmar Schuster

Zalando Research,
Zalando SE Berlin,
Berlin 10243, Germany

Christof Schütte

Zuse Institute Berlin,
Berlin 14195, Germany /
Department of Mathematics
and Computer Science,
Freie Universität Berlin,
Berlin 14195, Germany

1 Introduction

Over the last years, supervised learning algorithms have become an indispensable tool in many different scientific fields. Supervised learning – a subbranch of machine learning – is the approach to make predictions and decisions based on given input-output pairs, so-called training data. These data sets can represent highly diverse types of information so that supervised learning techniques have been widely applied to real-world problems in, e.g., physics and chemistry [1, 2], medicine [3, 4], and finance [5]. Further popular examples are image classification [6, 7], spam filtering [8, 9], and the identification of governing equations of dynamical systems [10, 11]. Learning algorithms can be subdivided into regression and classification methods. The former are used to estimate the relationship between input and output vectors, while the latter are used to identify the category to which a given input vector belongs. There exist a lot of different supervised learning methods such as (non-)linear regression, support-vector machines, decision trees, and neural networks. For a detailed overview of different algorithms and applications, we refer to [12].

The field is continuously evolving and the demand for machine learning has grown significantly in the last decade. In this work, we will focus on one of the most powerful mathematical tools in this area, namely kernel-based regression and classification techniques. In this context, a kernel is a function that enables us to avoid explicit representations of high-dimensional feature maps which are used as a basis for learning nonlinear target functions. The main advantage of kernel methods is that inner products are not evaluated explicitly, but only implicitly through the kernel. Linear methods that can be entirely written in terms of kernel evaluations can be turned into nonlinear methods by replacing the standard inner product by a kernel function. This is sometimes referred to as *kernelization* and generally leads to increased performance and accuracy. Examples of such methods are kernel PCA [13], kernel TICA [14], kernel EDMD [15, 16], and kernel SINDy/MANdY [6].

If the training data set contains highly similar entries or if one feature vector can be written as a linear combination of other feature vectors, the resulting transformed data matrices and Gram matrices will be ill-conditioned or even singular, which can cause numerical instabilities. This is typically alleviated by regularization or other techniques like reduced set methods [17] and Nyström approximations [18]. In this paper, we will propose a set-reduction technique for kernel-based regression and classification methods that extracts important samples from the training data set – and therefore relevant information encoded in the feature vectors – in order to reduce the storage consumption as well as the computational costs. Additionally, the method can also be seen as a regularization technique.

The remainder of the paper is structured as follows: In Section 2, we give a brief overview of regression and classification problems as well as the use of feature maps and kernel functions for supervised

learning methods. In Section 3, the basic assumptions for the feature space approximation are outlined. The data reduction is then considered from an analytical as well as a heuristic point of view, where the latter results in a method called *kernel-based feature space approximation* (kFSA). Numerical results for benchmark problems from different applications areas are presented in Section 4. Section 5 concludes with a brief summary and a future outlook.

2 Preliminaries

We begin with a brief overview of regression and classification problems as well as their formulation in terms of feature spaces and kernels. The notation used throughout this paper is summarized in Table 1. At some points though, we will slightly abuse the notation as we do not distinguish between a data matrix X and the set of its columns, which is also simply denoted by X .

Table 1: Notation for special matrices, functions, and spaces.

Symbol	Description
\mathcal{S}	sample space (subset of \mathbb{R}^d)
\mathcal{H}	feature space (subset of \mathbb{R}^n , $n \gg d$, or ℓ^2)
$\Psi: \mathcal{S} \rightarrow \mathcal{H}$	feature map
$X = [x^{(1)}, \dots, x^{(m)}]$	input matrix in $\mathbb{R}^{d \times m}$
$Y = [y^{(1)}, \dots, y^{(m)}]$	output matrix in $\mathbb{R}^{d' \times m}$
Ψ_X	transformed data matrix in $\mathbb{R}^{n \times m}$
$k: \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$	kernel function
κ	kernel parameter in \mathbb{R}^+
$G_{X, X'}$	Gram matrix in $\mathbb{R}^{m \times m'}$
Θ	coefficient matrix
$f: \mathcal{S} \rightarrow \mathbb{R}^{d'}$	target (regression or decision) function
$E: \mathcal{S} \times \mathbb{R}^{d \times m'} \rightarrow \mathbb{R}_0^+$	approximation error
ε	threshold for feature space approximation in \mathbb{R}_0^+
γ	regularization parameter in \mathbb{R}^+

2.1 Regression and classification problems

Regression and classification problems are subsumed under *supervised learning*, i.e., the data-driven approach to construct a function that correctly maps input to output variables. Given a training set of samples in the form of vectors $x^{(j)} \in \mathbb{R}^d$ and corresponding output vectors $y^{(j)} \in \mathbb{R}^{d'}$, with $j = 1, \dots, m$, the aim of regression is to model the relationship between both sets by a function $f: \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ which enables us to predict the output y for any given test vector x . Classification problems are here seen as a special case of regression analysis where the output variable is categorical. That is, y determines the label of x , where the output vectors are typically given in the *one-hot encoding* format [19]:

$$y_i = \begin{cases} 1, & \text{if } x \text{ belongs to category } i, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where d' is the number of considered categories. The advantage of one-hot encoding is that the ‘distance’ between any two categories is always the same, thus spurious effects when the (non-ordinal) labels of two dissimilar samples are close to each other are avoided.

In this work, we want to construct regression and decision functions, respectively, as linear combinations of preselected basis functions ψ_i , i.e.,

$$f(x) = \sum_{i \in I} \lambda_i \psi_i(x), \quad (2)$$

where I is an appropriate index set. In many cases, however, the number of basis functions can be extremely large (or even countably infinite, i.e., $I \cong \mathbb{N}$) causing high numerical costs. Thus, we will exploit kernel functions in order to reduce the storage consumption and computational complexity. For more details on kernel functions and their application to supervised learning tasks, we refer to [17] and [20].

2.2 Feature maps and kernel functions

If the learning algorithm can be entirely written in terms of kernel evaluations, explicit basis function evaluations are not needed in practice. This is called the *kernel trick* and allows us to reduce the computational complexity significantly since we do not have to compute inner products in the potentially infinite-dimensional space. However, for the following derivations it is crucial to clarify the relationship between (implicitly) given basis functions and the corresponding kernel. Additionally, we will later consider examples where the explicit representation of the regression function in terms of basis functions is of interest. Consider a vector $x \in \mathbb{R}^d$ and a mapping $\Psi: \mathbb{R}^d \rightarrow \mathbb{R}^n$, typically with $n \gg d$, given by

$$\Psi(x) := [\psi_1(x), \dots, \psi_n(x)]^\top \in \mathbb{R}^n,$$

where the real-valued functionals ψ_1, \dots, ψ_n define the preselected basis functions.

Remark 1: *In what follows, we will focus on finite-dimensional feature maps to elucidate our approach. However, the concepts of kernels, Gram matrices, and pseudoinverses described below can be adapted to infinite-dimensional mappings $\Psi: \mathbb{R}^d \rightarrow \mathcal{H}$, e.g., $\mathcal{H} \subseteq \ell^2$, if Ψ is a bounded operator with closed range, see [21].*

Given a data set $X = [x^{(1)}, \dots, x^{(m)}] \in \mathbb{R}^{d \times m}$, we define the *transformed data matrix*

$$\Psi_X := \left[\Psi \left(x^{(1)} \right), \dots, \Psi \left(x^{(m)} \right) \right] \in \mathbb{R}^{n \times m}.$$

The transformation Ψ can be interpreted as a *feature map* and, thus, defines a so-called *kernel*.

Definition 1: *A function $k: \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ on a non-empty set \mathcal{S} is called a kernel if there exist a real Hilbert space \mathcal{H} and a feature map $\Psi: \mathcal{S} \rightarrow \mathcal{H}$ such that*

$$k(x, x') = \langle \Psi(x), \Psi(x') \rangle_{\mathcal{H}} \quad (3)$$

for all $x, x' \in \mathcal{S}$.

By definition, a kernel is a symmetric and positive semidefinite function. The space \mathcal{H} is called the *feature space* of k . In what follows, we will generally consider \mathcal{S} as the space of training and test samples, i.e., $\mathcal{S} \subseteq \mathbb{R}^d$. In this regard, we will slightly abuse the notation at some points by simply denoting the column set of a given data matrix X by the same symbol. Furthermore, we consider \mathcal{H} as a subspace of \mathbb{R}^n , see Remark 1, and use the Euclidean inner product. The *Gram matrix* corresponding to the kernel k and a given data matrix X is defined as

$$G_{X,X} = \Psi_X^\top \Psi_X = \begin{bmatrix} k(x^{(1)}, x^{(1)}) & \dots & k(x^{(1)}, x^{(m)}) \\ \vdots & \ddots & \vdots \\ k(x^{(m)}, x^{(1)}) & \dots & k(x^{(m)}, x^{(m)}) \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

In the following sections, we will also consider Gram matrices corresponding to two different data sets X, X' . These are then denoted by $G_{X,X'} = \Psi_X^\top \Psi_{X'}$.

Remark 2: *The feature space representation (3) is in general not unique. That is, inner products of different feature maps can result in the same kernel. On the other hand, Mercer's theorem [22] states that we can find a (potentially infinite-dimensional) feature space representation for any given symmetric positive semidefinite kernel.*

2.3 Kernel-based regression and classification

It was shown in [6] that SINDy – originally developed for learning the governing equations of dynamical systems [10] – can be applied to classification problems using the transformed data matrices (or, more generally, tensors) and resulting kernels introduced above. This holds for supervised learning problems in general as we will illustrate in this section. Given a training set $X \in \mathbb{R}^{d \times m}$ and the corresponding output matrix $Y \in \mathbb{R}^{d' \times m}$, we consider the minimization problem

$$\min_{\Theta \in \mathbb{R}^{d' \times n}} \|Y - \Theta \Psi_X\|_F^2, \quad (4)$$

where $\Psi_X \in \mathbb{R}^{n \times m}$ is the transformed data matrix given by the feature map Ψ , see Section 2.2. A solution Θ of (4) then represents coefficients for expressing the output variables in terms of the preselected basis functions, i.e.,

$$y_i^{(j)} \approx \sum_{\mu=1}^n \Theta_{i,\mu} \psi_\mu(x^{(j)}).$$

If the number of basis functions is significantly larger than the number of training data points, typically equality holds. This might result in overfitting, i.e., the learned functions provide exact results for the training data but might not generalize well to new data. The regression/decision function $f: \mathcal{S} \rightarrow \mathbb{R}^{d'}$ is defined as

$$f(x) = \Theta \Psi(x), \quad (5)$$

where $x \in \mathcal{S} \subseteq \mathbb{R}^d$ is a sample from a given test set. For classification problems, the i th entry of the vector $f(x)$ is interpreted as the likelihood of x belonging to category i . The above approach in combination with hard thresholding is known as *sparse identification of nonlinear dynamical systems* (SINDy) in the context of system identification, i.e., when the data matrix X contains snapshots of a dynamical system and Y the corresponding derivatives.

Adding a regularization term of the form $\gamma \|\Theta\|_F^2$ to (4) yields the so-called ridge regression problem [23]. The solution would then be given by $\Theta = Y(\Psi_X^\top \Psi_X + \gamma \text{Id})^{-1} \Psi_X^\top$. As γ goes to zero, this converges to the solution of (4) with minimal Frobenius norm, i.e.,

$$\Theta = Y \Psi_X^+, \quad (6)$$

where Ψ_X^+ denotes the *Moore–Penrose inverse* (or simply called *pseudoinverse*) of Ψ_X , see [24]. The direct computation of Ψ_X^+ may be computationally expensive due to a large number of basis functions. Therefore, we use the kernel trick, i.e., the pseudoinverse is written as

$$\Psi_X^+ = (\Psi_X^\top \Psi_X)^+ \Psi_X^\top = G_{X,X}^+ \Psi_X^\top.$$

Now, the solution of (4) can be expressed as

$$\Theta = Y G_{X,X}^+ \Psi_X^\top. \quad (7)$$

The coefficient matrix $\Theta' := Y G_{X,X}^+$ is then the minimum-norm solution of

$$\min_{\Theta \in \mathbb{R}^{d' \times m}} \|Y - \Theta G_{X,X}\|_F^2. \quad (8)$$

This tells us that, equivalently to (4), we can consider the minimization problem (8) in order to construct the target function

$$f(x) = \Theta \Psi(x) = \Theta' \Psi_X^\top \Psi(x) = \Theta' G_{X,x},$$

where $G_{X,x}$ denotes the Gram matrix corresponding to the sets X and $\{x\}$.

In comparison to the coefficient matrix Θ , the matrix Θ' may have much smaller dimensions (assuming that $n \gg m$). Also, we only consider Gram matrices and kernel evaluations, there is no need to construct a transformed data matrix and we therefore may be able to mitigate storage problems. The above approach was introduced as kernel-based SINDy/MANDy in [6] and even works if the transformation Ψ is not given explicitly, but implicitly defined in terms of a kernel function.

3 Data reduction

As described in the previous section, the kernel-based construction of the regression/decision function allows us to omit explicit computations in the feature space. Under this premise, we are therefore not interested in methods for finding approximate kernel expressions or basis-function reductions as done for example in [25, 26, 27, 28]. The crucial aspect, in our case, is the dimension of the involved Gram matrix, whose size is given by the number of considered samples. In this section, we introduce an approach to filter unnecessary samples in order to speed up the computations, in particular in the testing phase.

3.1 Feature space approximation

We assume that the transformed data matrix is (nearly) rank-deficient or, in other words, there exists a $n \times \tilde{m}$ -submatrix of Ψ_X such that the column spaces of both matrices are (numerically) equal. Suppose there exist a subset $\tilde{X} \subset X$ and a matrix $C \in \mathbb{R}^{\tilde{m} \times m}$ with $\tilde{m} := |\tilde{X}| < m$ such that

$$\Psi_{\tilde{X}} C = \Psi_X, \quad (9)$$

where the columns of $\Psi_{\tilde{X}} \in \mathbb{R}^{n \times \tilde{m}}$ are a linearly independent subset of the columns of Ψ_X . In practice, (almost) linear dependencies among the columns of the transformed data matrix Ψ_X lead to ill-conditioned or singular Gram matrices. Thus, the proposed approach can also be seen as a regularization technique.

Lemma 1: *Given matrices $A \in \mathbb{R}^{k \times l}$ and $B \in \mathbb{R}^{l \times m}$ where A has full column rank and B full row rank, it holds that $(AB)^+ = B^+ A^+$.*

For a proof of the above lemma, we refer to [29]. By assuming that C has linearly independent rows, the pseudoinverse of Ψ_X is then given by

$$\Psi_X^+ = C^+ \Psi_{\tilde{X}}^+.$$

Using the kernel trick, the solution (6) can now be written as

$$\Theta = Y C^+ \Psi_{\tilde{X}}^+ = Y C^+ G_{\tilde{X}, \tilde{X}}^+ \Psi_{\tilde{X}}^\top.$$

It holds that $\Theta = \tilde{\Theta} \Psi_{\tilde{X}}^\top$, where the matrix $\tilde{\Theta} := Y C^+ G_{\tilde{X}, \tilde{X}}^+$ is the minimum-norm solution of the least-squares problem

$$\min_{\Theta \in \mathbb{R}^{d' \times \tilde{m}}} \|Y - \Theta G_{\tilde{X}, \tilde{X}}^+ C\|_F = \min_{\Theta \in \mathbb{R}^{d' \times \tilde{m}}} \|Y - \Theta G_{\tilde{X}, X}\|_F \quad (10)$$

because $G_{\tilde{X}, \tilde{X}}$ has full rank (equal to the column rank of $\Psi_{\tilde{X}}$) and therefore it holds that

$$(G_{\tilde{X}, \tilde{X}}^+ C)^+ = C^+ G_{\tilde{X}, \tilde{X}}^+.$$

Since $\tilde{m} < m \ll n$, this means we can – under the above assumptions – solve a minimization problem with potentially much lower dimensions than (7) and especially (6), but still can fully reconstruct the target function

$$f(x) = \tilde{\Theta} G_{\tilde{X}, x} = \Theta \Psi(x) \quad (11)$$

as defined in (5) (if the feature map is explicitly given). Note that even though the coefficient matrix $\tilde{\Theta}$ lives in a lower-dimensional space than Θ , the minimization problem (10) still includes all training data for supervised learning. Especially in the test phase this can be advantageous. If \tilde{m} is significantly smaller than m and n , the evaluation of the regression/classification function is much cheaper since it only requires the multiplication of a $d' \times \tilde{m}$ by a vector of \tilde{m} kernel evaluations.

In general, of course, there does not have to exist such an optimal set \tilde{X} satisfying (9). However, for practical applications we may find subsets $\tilde{X} \subset X$ which satisfy (9) up to a certain extent, i.e., any feature vector $\Psi(x)$, $x \in X$, can be closely approximated as a linear combination of the columns of $\Psi_{\tilde{X}}$.

3.2 Analytical formulation & existing approaches

As described above, we seek a subset \tilde{X} of the columns of the data matrix X such that (9) is (approximately) satisfied. That is, for each vector in $x \in X \setminus \tilde{X}$ there should exist a vector $c \in \mathbb{R}^{\tilde{m}}$ such that $\Psi_{\tilde{X}}c$ is equal to (or very close to) $\Psi(x)$. In other words, we try to find a smallest possible $\tilde{X} \subset X$ such that all column vectors of Ψ_X lie in the span of the column vectors of $\Psi_{\tilde{X}}$. For a single vector $\Psi(x)$ with $x \in X$, we define the approximation error by

$$E(\tilde{X}, x) := \min_{c \in \mathbb{R}^{\tilde{m}}} \|\Psi(x) - \Psi_{\tilde{X}}c\|_2^2, \quad (12)$$

where the minimum-norm solution can be written as $c = \Psi_{\tilde{X}}^+ \Psi(x)$, i.e., $E(\tilde{X}, x) = \Psi(x) - \Psi_{\tilde{X}} \Psi_{\tilde{X}}^+ \Psi(x)$. If $\Psi(x) \in \text{span } \Psi_{\tilde{X}}$, especially if $x \in \tilde{X}$, the approximation error (12) is equal to zero. Combining the error minimization with the constraint that $\tilde{m} = |\tilde{X}|$ should be small in comparison to $m = |X|$ (assuming that the number of given observations is large enough), we have to solve the following optimization problem:

$$\begin{aligned} \min_{\tilde{X} \subset X} \left(\sum_{x \in X} E(\tilde{X}, x) + \gamma \tilde{m} \right) &= \min_{\tilde{X} \subset X} \left(\sum_{x \in X} \min_{c \in \mathbb{R}^{\tilde{m}}} \|\Psi(x) - \Psi_{\tilde{X}}c\|_2^2 + \gamma \tilde{m} \right) \\ &= \min_{\tilde{X} \subset X} \left(\min_{C \in \mathbb{R}^{\tilde{m} \times m}} \|\Psi_X - \Psi_{\tilde{X}}C\|_F^2 + \gamma \tilde{m} \right), \end{aligned} \quad (13)$$

where γ is a regularization parameter such that high values for γ will enforce selecting fewer points.

Finding an optimal solution of (13) is a complex combinatorial optimization problem and there exist different iterative and greedy approaches. A description of *reduced set methods*, where samples are removed from the set X using, e.g., kernel PCA or ℓ_1 penalization, can be found in [17]. However, these approaches are based on a stepwise reduction of the whole set of feature vectors. That is, in each iteration step the sample $x \in \tilde{X}$ with minimum error $E(\tilde{X} \setminus x, x)$ is removed from \tilde{X} (starting with $\tilde{X} = X$). In particular, if the set of feature vectors is too large to handle at once, this method cannot be applied. Depending on the number of samples and the dimension of the feature space, reduced set methods can be computationally expensive in general. Another efficient way to find important samples in the given data set is the *Nyström method* [30]. It can be generally applied to symmetric positive semidefinite matrices A in order to compute a decomposition of the form $A = BC^+B^\top$, where B is a matrix formed by a subset of the columns of A and C is an intersection matrix. In particular, the Nyström method can be used to approximate Gram matrices [18, 31]. That is, for a fixed number of samples \tilde{m} , the Nyström algorithm seeks to find a subset $\tilde{X} \subset X$ with $|\tilde{X}| = \tilde{m}$ such that the Gram matrix $G_{X,X} = \Psi_X^\top \Psi_X$ can be approximately written as

$$G_{X,X} \approx G_{X,\tilde{X}}C, \quad (14)$$

with $C = G_{\tilde{X},\tilde{X}}^+ G_{\tilde{X},X}$. On the one hand, however, we here have a restriction on C so that there generally may exist better choices for \tilde{X} and C to minimize the error $\|G_{X,X} - G_{X,\tilde{X}}C\|_F$. On the other hand, the Nyström approximation does not provide a feature space approximation as described in Section 3.1 since (14) does not imply $\Psi_X \approx \Psi_{\tilde{X}}C$. Note that only the other direction is true.

In what follows, we will therefore propose a heuristic approach to iteratively construct a subset \tilde{X} such that any feature vector corresponding to a sample $x \in X$ can be approximated by a linear combination of the set $\{\Psi(x) : x \in \tilde{X}\}$ up to an arbitrarily small error ε . In Section 4.3, results obtained by applying the (recursive) Nyström method as introduced in [32] and our approach will be compared.

Remark 3: *The Nyström method as well as, e.g., the iterative spectral method (ISM) [33] can also be used for the reduction of the feature space dimension. However, our aim is to reduce the number of samples/feature vectors m , not the dimension n .*

3.3 Heuristic approach

In order to find a subset as described in Section 3.2, we introduce a heuristic approach based on a *bottom-up* principle – in contrast to the reduced set methods described in [17], which are based on

a *top-down* selection of the samples. Suppose we are given a subset \tilde{X} of the columns of X and the corresponding transformed data matrix $\Psi_{\tilde{X}}$. For the following approach – which we call *kernel-based feature space approximation* or in short *kFSA* – we expand the subset \tilde{X} step by step by adding the vector x that has the largest approximation error (12). The idea is to increase the dimension of the subspace spanned by the columns of $\Psi_{\tilde{X}}$ iteratively without including linearly dependent feature vectors (or even duplicates).

As the first data point, we choose the sample whose feature vector approximates all other columns of Ψ_X in an optimal way, i.e., we consider the extreme case of the minimization problem (13) where the regularization parameter γ is sufficiently large so that only a single sample is selected. We then obtain

$$x_0 = \arg \min_{x \in X} \min_{c \in \mathbb{R}^{1 \times m}} \|\Psi_X - \Psi(x)c\|_F^2.$$

It holds that

$$c = \Psi(x)^+ \Psi_X = \frac{1}{\|\Psi(x)\|_2^2} \Psi(x)^\top \Psi_X = \frac{1}{k(x, x)} G_{x, X}$$

and, therefore, x_0 is given by

$$\begin{aligned} x_0 &= \arg \min_{x \in X} \|\Psi_X - \frac{1}{k(x, x)} \Psi(x) G_{x, X}\|_F^2 \\ &= \arg \min_{x \in X} \text{tr} \left(\left(\Psi_X - \frac{1}{k(x, x)} \Psi(x) G_{x, X} \right)^\top \left(\Psi_X - \frac{1}{k(x, x)} \Psi(x) G_{x, X} \right) \right) \\ &= \arg \min_{x \in X} \text{tr} \left(G_{X, X} - \frac{2}{k(x, x)} G_{X, x} G_{x, X} + \frac{1}{k(x, x)^2} G_{X, x} k(x, x) G_{x, X} \right) \\ &= \arg \min_{x \in X} \text{tr} \left(G_{X, X} - \frac{1}{k(x, x)} G_{X, x} G_{x, X} \right) \\ &= \arg \min_{x \in X} \sum_{x' \in X} k(x', x') - \frac{k(x, x')^2}{k(x, x)} \\ &= \arg \max_{x \in X} \sum_{x' \in X} \frac{k(x, x')^2}{k(x, x)}. \end{aligned} \tag{15}$$

As long as $\Psi_{\tilde{X}} \in \mathbb{R}^{n \times \tilde{m}}$, $n \geq \tilde{m}$, has full column rank, it holds that $G_{\tilde{X}, \tilde{X}}^+ = G_{\tilde{X}, \tilde{X}}^{-1}$ since $G_{\tilde{X}, \tilde{X}}$ is then a non-singular matrix. Thus, we can rewrite the approximation error (12) in terms of kernel functions and corresponding Gram matrices.

Lemma 2: *Using the fact that $\Psi_{\tilde{X}}^+ = G_{\tilde{X}, \tilde{X}}^{-1} \Psi_{\tilde{X}}^\top$, the approximation error (12) for any $x \in X$ can be written as*

$$E(\tilde{X}, x) = k(x, x) - G_{x, \tilde{X}} G_{\tilde{X}, \tilde{X}}^{-1} G_{\tilde{X}, x}. \tag{16}$$

Proof: *Using $c = \Psi_{\tilde{X}}^+ \Psi(x)$, we obtain*

$$\begin{aligned} E(\tilde{X}, x) &= \|\Psi(x) - \Psi_{\tilde{X}} \Psi_{\tilde{X}}^+ \Psi(x)\|_2^2 \\ &= \|\Psi(x) - \Psi_{\tilde{X}} G_{\tilde{X}, \tilde{X}}^{-1} \Psi_{\tilde{X}}^\top \Psi(x)\|_2^2 \\ &= \left(\Psi(x) - \Psi_{\tilde{X}} G_{\tilde{X}, \tilde{X}}^{-1} \Psi_{\tilde{X}}^\top \Psi(x) \right)^\top \left(\Psi(x) - \Psi_{\tilde{X}} G_{\tilde{X}, \tilde{X}}^{-1} \Psi_{\tilde{X}}^\top \Psi(x) \right) \\ &= \Psi(x)^\top \Psi(x) - 2 \Psi(x)^\top \Psi_{\tilde{X}} G_{\tilde{X}, \tilde{X}}^{-1} \Psi_{\tilde{X}}^\top \Psi(x) + \Psi(x)^\top \Psi_{\tilde{X}} G_{\tilde{X}, \tilde{X}}^{-1} \underbrace{\Psi_{\tilde{X}}^\top \Psi_{\tilde{X}}}_{=G_{\tilde{X}, \tilde{X}}} G_{\tilde{X}, \tilde{X}}^{-1} \Psi_{\tilde{X}}^\top \Psi(x) \\ &= k(x, x) - G_{x, \tilde{X}} G_{\tilde{X}, \tilde{X}}^{-1} G_{\tilde{X}, x}. \end{aligned} \quad \square$$

Let us denote the vector containing all approximation errors for the samples in $X \setminus \tilde{X}$ by Δ . Using the Hadamard product, this vector can be expressed in a compact way as the following lemma shows.

Lemma 3: The vector $\Delta \in \mathbb{R}^{m-\tilde{m}}$ can be written as

$$\Delta = \text{diag} \left(G_{X \setminus \tilde{X}, X \setminus \tilde{X}} \right)^\top - \mathbf{1}^\top \cdot \left(G_{\tilde{X}, X \setminus \tilde{X}} \odot Z \right), \quad (17)$$

where $\mathbf{1} \in \mathbb{R}^{\tilde{m}}$ is a vector of ones and Z is the unique solution of

$$G_{\tilde{X}, \tilde{X}} Z = G_{\tilde{X}, X \setminus \tilde{X}}. \quad (18)$$

Proof: For any $x \in X$, the approximation error (16) can be expressed as

$$\begin{aligned} E(\tilde{X}, x) &= k(x, x) - \sum_{i=1}^{\tilde{m}} \left(G_{\tilde{X}, x} \right)_i \left(G_{\tilde{X}, \tilde{X}}^{-1} G_{\tilde{X}, x} \right)_i \\ &= k(x, x) - \sum_{i=1}^{\tilde{m}} \left(G_{\tilde{X}, x} \odot \left(G_{\tilde{X}, \tilde{X}}^{-1} G_{\tilde{X}, x} \right) \right)_i \\ &= k(x, x) - \mathbf{1}^\top \cdot \left(G_{\tilde{X}, x} \odot \left(G_{\tilde{X}, \tilde{X}}^{-1} G_{\tilde{X}, x} \right) \right). \end{aligned}$$

For the error vector Δ , it follows that

$$\Delta = \text{diag} \left(G_{X \setminus \tilde{X}, X \setminus \tilde{X}} \right)^\top - \mathbf{1}^\top \cdot \left(G_{\tilde{X}, X \setminus \tilde{X}} \odot \underbrace{G_{\tilde{X}, \tilde{X}}^{-1} G_{\tilde{X}, X \setminus \tilde{X}}}_{=: Z} \right). \quad \square$$

Note that we have to construct the complete Gram matrix $G_{X, X}$ corresponding to all samples only once. For the computation of the approximation errors in each step, we could then simply extract the necessary submatrices. However, using Lemma 3 for the error computation would result in a high complexity since the dominant computational costs arise from solving the systems of linear equations given in (18). That is, supposing we have already found \tilde{m} samples, we then have to solve a system with an $\tilde{m} \times \tilde{m}$ -matrix and $m - \tilde{m}$ right-hand sides. Thus, the overall complexity can be estimated by $O \left(\sum_{\tilde{m}=1}^M \tilde{m}^3 (m - \tilde{m}) \right)$ if we repeat the procedure M times until the desired accuracy is reached.

In order to reduce the computational effort, we can use an alternative way of calculating the approximation errors. Instead of computing the matrix Z that is needed for the construction of Δ in (17) as the solution of the system of linear equations given in (18), it is possible to iteratively calculate $G_{\tilde{X}, \tilde{X}}^{-1}$ in order to compute Z directly by matrix multiplication.

Lemma 4: Given matrices A, B, C, D of appropriate size such that A and $D - CA^{-1}B$ are non-singular square matrices, the blockwise inversion formula (also called Banachiewicz inversion formula) states

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BS^{-1}CA^{-1} & -A^{-1}BS^{-1} \\ -S^{-1}CA^{-1} & S^{-1} \end{bmatrix},$$

where the matrix S – called the Schur complement – is given by $S = D - CA^{-1}B$.

For a proof of the above statement, we refer to [34]. Suppose we want to expand the (linearly independent) set \tilde{X} by a sample x_{new} and the inverse of $G_{\tilde{X}, \tilde{X}}$ is known – which is of course the case for the initial sample. The Gram matrix corresponding to the set $\tilde{X} \cup \{x_{\text{new}}\}$ is then given by

$$G_{\tilde{X} \cup \{x_{\text{new}}\}, \tilde{X} \cup \{x_{\text{new}}\}} = \begin{bmatrix} \Psi_{\tilde{X}}^\top \\ \Psi(x_{\text{new}})^\top \end{bmatrix} \begin{bmatrix} \Psi_{\tilde{X}} & \Psi(x_{\text{new}}) \end{bmatrix} = \begin{bmatrix} G_{\tilde{X}, \tilde{X}} & G_{\tilde{X}, x_{\text{new}}} \\ G_{x_{\text{new}}, \tilde{X}} & G_{x_{\text{new}}, x_{\text{new}}} \end{bmatrix}, \quad (19)$$

where $G_{x_{\text{new}}, x_{\text{new}}}$ is simply the kernel evaluation at x_{new} , i.e., $G_{x_{\text{new}}, x_{\text{new}}} = k(x_{\text{new}}, x_{\text{new}})$. If the feature vector $\Psi(x_{\text{new}})$ is not an element of the span of the columns of $\Psi_{\tilde{X}}$, then the matrix (19) is non-singular and, furthermore, symmetric positive definite. Applying Lemma 4, we can write

$$G_{\tilde{X} \cup \{x_{\text{new}}\}, \tilde{X} \cup \{x_{\text{new}}\}}^{-1} = \begin{bmatrix} G_{\tilde{X}, \tilde{X}}^{-1} + \frac{TT^\top}{S} & -\frac{T}{S} \\ -\frac{T^\top}{S} & \frac{1}{S} \end{bmatrix} \quad (20)$$

with $T = G_{\tilde{X}, \tilde{X}}^{-1} G_{\tilde{X}, x_{\text{new}}}$ and $S = k(x_{\text{new}}, x_{\text{new}}) - G_{x_{\text{new}}, \tilde{X}} G_{\tilde{X}, \tilde{X}}^{-1} G_{\tilde{X}, x_{\text{new}}} = E(\tilde{X}, x_{\text{new}}) > 0$. Using (20) would already avoid having to solve a system of linear equations. In this case, the dominant computational effort, namely the multiplication of $G_{\tilde{X}, \tilde{X}}^{-1}$ and $G_{\tilde{X}, X \setminus \tilde{X}}$ in every iteration step, can be estimated by $O\left(\sum_{\tilde{m}=1}^M \tilde{m}^2(m - \tilde{m})\right)$. However, we can exploit the block inverse (20) to directly construct an iterative scheme for Z and therefore reduce the complexity of our method even further. Each column of the matrix $Z = G_{\tilde{X}, \tilde{X}}^{-1} G_{\tilde{X}, X \setminus \tilde{X}} \in \mathbb{R}^{\tilde{m} \times (m - \tilde{m})}$ is from now on considered as the evaluation of a function H of a set \tilde{X} and a sample $x \in X \setminus \tilde{X}$, i.e., $H(\tilde{X}, x) = G_{\tilde{X}, \tilde{X}}^{-1} G_{\tilde{X}, x}$.

Theorem 1: For any x_{new} with $\Psi(x_{\text{new}}) \notin \text{span } \Psi_{\tilde{X}}$ and $x \in X$, it holds that

$$H(\tilde{X} \cup \{x_{\text{new}}\}, x) = \begin{bmatrix} H(\tilde{X}, x) - H(\tilde{X}, x_{\text{new}})\lambda \\ \lambda \end{bmatrix}$$

with $\lambda = \frac{k(x_{\text{new}}, x) - G_{x_{\text{new}}, \tilde{X}} H(\tilde{X}, x)}{E(\tilde{X}, x_{\text{new}})}$.

Proof: By definition, we have

$$H(\tilde{X} \cup \{x_{\text{new}}\}, x) = G_{\tilde{X} \cup \{x_{\text{new}}\}, \tilde{X} \cup \{x_{\text{new}}\}}^{-1} G_{\tilde{X} \cup \{x_{\text{new}}\}, x}$$

Using (20), it follows that

$$H(\tilde{X} \cup \{x_{\text{new}}\}, x) = \begin{bmatrix} G_{\tilde{X}, \tilde{X}}^{-1} + \frac{TT^\top}{S} & -\frac{T}{S} \\ -\frac{T^\top}{S} & \frac{1}{S} \end{bmatrix} \begin{bmatrix} G_{\tilde{X}, x} \\ k(x_{\text{new}}, x) \end{bmatrix},$$

with $T = G_{\tilde{X}, \tilde{X}}^{-1} G_{\tilde{X}, x_{\text{new}}} = H(\tilde{X}, x_{\text{new}})$ and $S = E(\tilde{X}, x_{\text{new}})$, see Lemma 2. Therefore, we obtain

$$\begin{aligned} H(\tilde{X} \cup \{x_{\text{new}}\}, x) &= \begin{bmatrix} G_{\tilde{X}, \tilde{X}}^{-1} G_{\tilde{X}, x} + H(\tilde{X}, x_{\text{new}}) \left(\frac{G_{x_{\text{new}}, \tilde{X}} G_{\tilde{X}, \tilde{X}}^{-1} G_{\tilde{X}, x} - k(x_{\text{new}}, x)}{E(\tilde{X}, x_{\text{new}})} \right) \\ \frac{k(x_{\text{new}}, x) - G_{x_{\text{new}}, \tilde{X}} G_{\tilde{X}, \tilde{X}}^{-1} G_{\tilde{X}, x}}{E(\tilde{X}, x_{\text{new}})} \end{bmatrix} \\ &= \begin{bmatrix} H(\tilde{X}, x) - H(\tilde{X}, x_{\text{new}}) \left(\frac{k(x_{\text{new}}, x) - G_{x_{\text{new}}, \tilde{X}} H(\tilde{X}, x)}{E(\tilde{X}, x_{\text{new}})} \right) \\ \frac{k(x_{\text{new}}, x) - G_{x_{\text{new}}, \tilde{X}} H(\tilde{X}, x)}{E(\tilde{X}, x_{\text{new}})} \end{bmatrix}, \end{aligned}$$

which yields the assertion. Note that $H(\tilde{X}, x) \in \mathbb{R}^{\tilde{m}}$ is not a subvector of $H(\tilde{X} \cup x_{\text{new}}, x) \in \mathbb{R}^{\tilde{m}+1}$. \square

Considering $Z = [H(\tilde{X}, x)]_{x \in X \setminus \tilde{X}} \in \mathbb{R}^{\tilde{m} \times (m - \tilde{m})}$ as given in (18), we define

$$\Lambda = [\lambda]_{x \in X \setminus \tilde{X}} = \frac{1}{E(\tilde{X}, x_{\text{new}})} \left(G_{x_{\text{new}}, X \setminus \tilde{X}} - G_{x_{\text{new}}, \tilde{X}} Z \right) \in \mathbb{R}^{m - \tilde{m}}. \quad (21)$$

Then, the matrix

$$Z_{\text{new}} = G_{\tilde{X} \cup \{x_{\text{new}}\}, \tilde{X} \cup \{x_{\text{new}}\}}^{-1} G_{\tilde{X} \cup \{x_{\text{new}}\}, X \setminus (\tilde{X} \cup \{x_{\text{new}}\})} \in \mathbb{R}^{\tilde{m}+1 \times (m - \tilde{m} - 1)}$$

required for the error computation (17) when considering the updated set $\tilde{X} \cup \{x_{\text{new}}\}$ instead of \tilde{X} can be written as

$$Z_{\text{new}} = \begin{bmatrix} Z - Z_{|x_{\text{new}}} \Lambda^\top \\ \Lambda^\top \end{bmatrix}_{|X \setminus (\tilde{X} \cup \{x_{\text{new}}\})}, \quad (22)$$

where $Z_{|x_{\text{new}}} = H(\tilde{X}, x_{\text{new}})$ denotes the column of Z corresponding to the sample $x_{\text{new}} \in X \setminus \tilde{X}$. Note that the matrix computed in (22) is also restricted to the columns corresponding to the set $X \setminus (\tilde{X} \cup \{x_{\text{new}}\})$. We combine the iterative construction of the matrix Z with the error-vector computation given in (17). The above derivations are summarized in Algorithm 1.

Lemma 5: Neglecting the computational costs for the construction of $G_{X,X}$, the complexity of Algorithm 1 can be estimated by $O(m^2 + mM^2)$, where m is the number of given training samples and M the number of extracted samples.

Proof: Determining the initial sample in line 1 needs $O(m^2)$ operations. For every execution of line 4, the dominant computational costs arise from calculating the Hadamard product in (17), i.e., $O(\tilde{m}(m - \tilde{m}))$. Since the error $E(\tilde{X}, x_{\text{new}})$ is already known in form of δ_{new} , the complexity of line 9 is divided into the computations of Δ and Z_{new} as given in (21) and (22), respectively. Both can be estimated by $O(\tilde{m}(m - \tilde{m}))$ as well. Repeating lines 4 to 9 of the above algorithm M times, i.e., until the desired accuracy is reached, and adding all computational costs yields

$$\begin{aligned} O\left(m^2 + \sum_{\tilde{m}=1}^M \tilde{m}(m - \tilde{m})\right) &= O\left(m^2 + m \sum_{\tilde{m}=1}^M \tilde{m} - \sum_{\tilde{m}=1}^M \tilde{m}^2\right) \\ &= O\left(m^2 + \frac{mM(M+1)}{2} - \frac{M(M+1)(2M+1)}{6}\right) \\ &= O\left(m^2 + M(M+1)\left(\frac{m}{2} - \frac{2M+1}{6}\right)\right), \end{aligned}$$

which can be roughly summarized as $O(m^2 + mM^2)$. Note that this estimation does not include a potential speed up due to removing samples for which the approximation error is already low enough, see line 6. \square

Algorithm 1: Kernel-based feature space approximation (kFSA).

Input: data matrix X , kernel function k , and threshold $\varepsilon > 0$.

Output: subset $\tilde{X} \subset X$ with $E(\tilde{X}, x) < \varepsilon$ for all $x \in X$.

- 1: Set $\tilde{X} := \{x_0\}$, see (15), and $X_{\text{left}} := X \setminus \{x_0\}$.
- 2: Define $Z := G_{\tilde{X}, \tilde{X}}^{-1} G_{\tilde{X}, X_{\text{left}}} = (1/k(x_0, x_0))G_{x_0, X_{\text{left}}}$.
- 3: **while** X_{left} is not empty **do**
- 4: Use Z to compute the vector $\Delta \in \mathbb{R}^{|X_{\text{left}}|}$ of approximation errors, see (17).
- 5: Define $x_{\text{new}} \in X$ to be the sample corresponding to the maximum error δ_{new} in Δ .
- 6: Remove all x from X_{left} with corresponding approximation errors smaller than ε .
- 7: **if** $\delta_{\text{new}} \geq \varepsilon$ **then**
- 8: Add sample vector x_{new} to \tilde{X} and remove it from X_{left} .
- 9: Update Z as described above in (22).

When Algorithm 1 terminates, all given samples $x \in X$ which are not an element of \tilde{X} can be approximated with an error $E(\tilde{X}, x)$ smaller than ε . In case the Gram matrix corresponding to the whole training set consumes too much memory, the algorithm may be applied iteratively to subsets of the training set in order to filter unnecessary samples. Moreover, suppose we have found a subset $\tilde{X} \subset X$ by applying Algorithm 1. If we subsequently want to extend the training data set, one of the main advantages of our method in comparison to, e.g., reduced set methods is that it is easy to decide whether or not a new sample is added to \tilde{X} . We do not have to apply the algorithm again to the extended set but can simply continue our bottom-up approach.

4 Results

In this section, we will present three examples for the application of data-reduced regression and classification, namely the MNIST data set, the Fermi–Pasta–Ulam–Tsingou problem, and hydrographic data from the California Cooperative Oceanic Fisheries Investigations. The numerical experiments

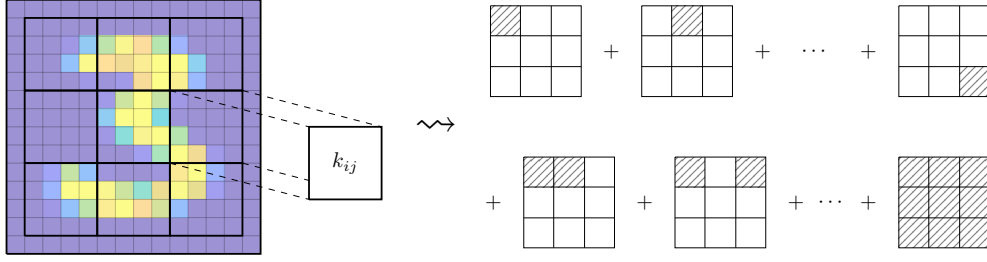


Figure 1: Block decomposition of the MNIST data set: Each image is divided into 9 disjoint 4×4 blocks where the pixels at the margin are neglected since their corresponding values are zero for almost all images in the data set. Then, a kernel function is defined for each subgroup of the pixels. The kernel combinations encoded in the Gram matrix G (24) are depicted on the right-hand side.

have been performed on a Linux machine with 128 GB RAM and a 3 GHz Intel Xeon processor with 8 cores. The algorithms have been implemented in Matlab R2018b and are available on GitHub: <https://github.com/PGelss/kFSA>.

4.1 MNIST data set

Let us consider the classification of the MNIST data set [35] as a first example. The set contains grayscale images of handwritten digits and their corresponding labels from 0 to 9 (represented using one-hot encoding). Originally, the images have a resolution of 28×28 pixels and are divided into 60000 training and 10000 test samples.¹ We downsample the images to 14×14 pixels by averaging groups of four pixels, cf. [36]. Additionally, we normalize the samples such that the largest pixel value in each image is equal to 1. This has shown to yield better classification rates in our experiments. In [6], we already used the kernel-based methods described in Section 2.3 to classify the downsampled MNIST images. Due to the large amount of training data, this example set is an ideal candidate for applying our proposed set-reduction method.

Given a subset of mutually different pixels $i_1, \dots, i_r \in \{1, \dots, 14^2\}$, we again use feature maps $\Psi: \mathbb{R}^r \rightarrow \mathbb{R}^{2 \times 2 \times \dots \times 2}$ of the form

$$\Psi(x) = \Psi(x_{i_1}) \otimes \Psi(x_{i_2}) \otimes \dots \otimes \Psi(x_{i_r}) = \begin{bmatrix} \cos(\kappa x_{i_1}) \\ \sin(\kappa x_{i_1}) \end{bmatrix} \otimes \begin{bmatrix} \cos(\kappa x_{i_2}) \\ \sin(\kappa x_{i_2}) \end{bmatrix} \otimes \dots \otimes \begin{bmatrix} \cos(\kappa x_{i_r}) \\ \sin(\kappa x_{i_r}) \end{bmatrix}, \quad (23)$$

where x_{i_1}, \dots, x_{i_r} are the corresponding pixel values and $\kappa > 0$ is the kernel parameter. In contrast to, e.g., [6, 36], we do not consider all pixels of a given image at once. Instead, we partition the images into 9 blocks, see Figure 1, and then apply feature maps as given in (23) separately. The feature map (23) is defined in terms of tensor products. That is, the entries of $\Psi(x)$ are given by products of r cosine or sine functions evaluated at the given pixels, i.e.,

$$\Psi(x)_{b_1, \dots, b_r} = (\Psi(x_{i_1}))_{b_1} \cdot (\Psi(x_{i_2}))_{b_2} \cdot \dots \cdot (\Psi(x_{i_r}))_{b_r},$$

where

$$(\Psi(x_{i_j}))_{b_j} = \begin{cases} \cos(\kappa x_{i_j}), & \text{if } b_j = 1, \\ \sin(\kappa x_{i_j}), & \text{if } b_j = 2. \end{cases}$$

The corresponding kernel is then given by

$$\Psi(x)^\top \Psi(x') = \bigotimes_{j=1}^r \left(\begin{bmatrix} \cos(\kappa x_{i_j}) & \sin(\kappa x_{i_j}) \end{bmatrix} \cdot \begin{bmatrix} \cos(\kappa x'_{i_j}) \\ \sin(\kappa x'_{i_j}) \end{bmatrix} \right) = \prod_{j=1}^r \cos(\kappa(x_{i_j} - x'_{i_j})).$$

Thus, we consider 9 kernel functions k_1, \dots, k_9 corresponding to different subgroups of the pixels as shown in Figure 1, compute the associated Gram matrices $G_1, \dots, G_9 \in \mathbb{R}^{m \times m}$ and define

$$G = \frac{1}{511} ((G_1 + \mathbb{1}) \odot \dots \odot (G_9 + \mathbb{1}) - \mathbb{1}) = \frac{1}{511} \sum_{p \in I} G_1^{\odot p_1} \odot G_2^{\odot p_2} \odot \dots \odot G_9^{\odot p_9}, \quad (24)$$

¹<http://yann.lecun.com/exdb/mnist/>

where $\mathbf{1} \in \mathbb{R}^{m \times m}$ denotes the matrix of ones and

$$I = \{p = (p_1, \dots, p_9) \in [0, 1]^9 : \text{at least one } p_i \neq 0\}$$

is a set of binary multi-indices. By the notation $\odot p_i$ in (24) we mean the element-wise exponentiation of the matrices G_i , i.e., $G_i^{\odot 0} = \mathbf{1}$ and $G_i^{\odot 1} = G_i$. The matrix G is the sum over all possible q -combinations, $q \geq 1$, of Hadamard products of the matrices G_1, \dots, G_9 as depicted in the right part of Figure 1. Each combination represents the application of the kernel corresponding to the feature map (23) on a subgroup of the different blocks. That is, we consider

$$\sum_{q=1}^9 \binom{9}{q} = 2^9 - 1 = 511$$

different kernel functions since the product of kernels again forms a kernel, see [37]. The same holds for the sum of kernels. Let us denote the kernel function corresponding to the Gram matrix G in (24) by k .

In [36], the proposed value for the kernel parameter is $\kappa = \pi/2$. However, as discussed in [6], the optimal choice can vary. Thus, we consider different values for the kernel parameter κ as well as for the threshold ε in Algorithm 1. Since it is reasonable to assume that the feature vectors corresponding to different categories are linearly independent, we split the training set into the different classes and apply Algorithm 1 to each category separately. The number of samples extracted by our method is shown in Figure 2 (a).

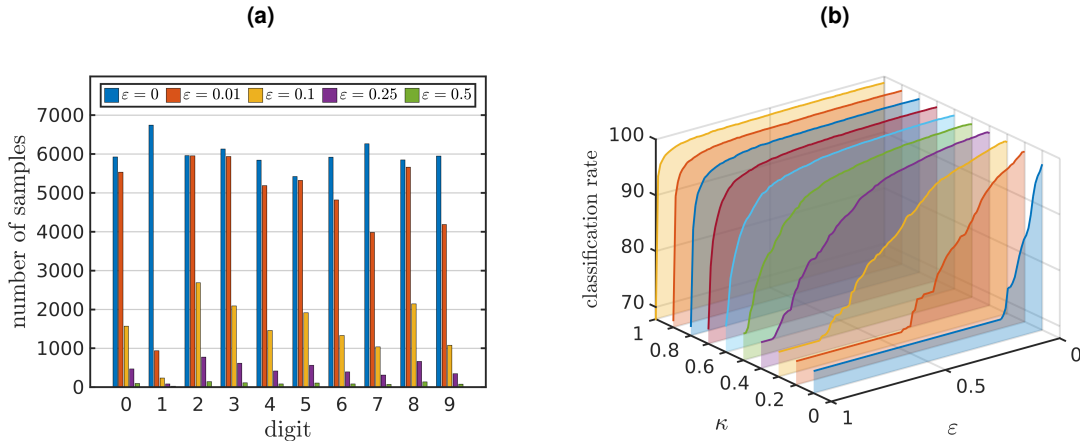


Figure 2: Application of $kFSA$ to the MNIST data set: (a) Number of extracted samples per digit for $\kappa = 0.5$ and different thresholds. (b) Classification rates obtained on the test set as functions of ε for different values of κ , both between 0 and 1. The intervals in which the classification rates are constant for small values of κ indicate that only one sample is extracted for each category.

The numbers of extracted samples for different thresholds show that the digit 1 is subject to the least variation since even for thresholds close to zero only a relatively small number of samples remains. For instance, if we consider $\kappa = 1$ and set $\varepsilon = 0.05$, nearly all samples from the training data set are extracted by Algorithm 1 except for the samples corresponding to digit 1 where only 1820 out of 6742 images are included in the set \tilde{X} . In contrast to that, the feature vectors of the samples corresponding to the digit 2 show the highest degree of linear independence, i.e., a comparatively large number of samples is needed in order to reach the desired approximation errors (12) among all respective feature vectors. Since the considered kernel is normalized, i.e., $k(x, x) = 1$ for all $x \in X$, the approximation errors are smaller than 1 for any initial sample x_0 (15) in a given category. Thus, only one sample is extracted per category if we choose a threshold of $\varepsilon = 1$, whereas all given samples in the training set are taken into account if we set $\varepsilon = 0$. In the latter case, the minimization problems (8) are ill-conditioned and additional regularization is necessary. Therefore, we apply *kernel ridge regression* [23], i.e., we solve the minimization problem

$$\min_{\Theta \in \mathbb{R}^{d_{\text{prime}} \times m}} \|Y - \Theta(G_{X,X} + \gamma \text{Id})\|_F^2$$

with $\gamma = 10^{-10}$.

For the test phase, we solve the minimization problem (10) and construct the corresponding target function (11). Given a test vector x , the index of the largest entry of $f(x)$ (interpreted as a vector of likelihoods) then defines the predicted category. Figure 2 (b) shows the amount of correctly identified images in the test set for different kernel parameters and thresholds. As the kernel parameter approaches 1, the classification rate tends to increase even for larger thresholds. On the one hand, we observe the effect that for a fixed threshold more and more samples are extracted with increasing kernel parameter. On the other hand, however, note that we achieve high classification rates using only a fraction of the training set instead the whole set. In fact, Table 2 shows that in most cases the generalizability increases for some thresholds $\varepsilon > 0$.

Table 2: Application of kFSA to the MNIST data set: For each choice of κ , we show the threshold $\varepsilon > 0$ and the numbers of samples per digit corresponding to the highest classification rates (CR) obtained by applying Algorithm 1. Except for the digit 1 – where we get a classification rate of 97.53% when using all 60000 samples – the results are better than the respective rates for $\varepsilon = 0$.

κ	ε	number of samples per digit										CR
		0	1	2	3	4	5	6	7	8	9	
0.1	0.01	74	41	91	81	76	82	74	69	88	73	96.36%
0.2	0.01	510	142	751	623	473	592	443	375	669	401	98.39%
0.3	0.01	1892	340	3027	2426	1773	2237	1608	1317	2430	1319	98.82%
0.4	0.02	2763	418	4336	3477	2559	3210	2310	1814	3426	1837	98.95%
0.5	0.04	3301	449	4972	4062	3024	3760	2708	2117	3980	2178	98.99%
0.6	0.07	3747	469	5370	4515	3399	4179	3055	2377	4393	2468	99.01%
0.7	0.16	3137	368	4901	3974	2872	3650	2566	1964	3920	2038	98.98%
0.8	0.19	3923	441	5527	4742	3581	4349	3196	2464	4613	2576	98.99%
0.9	0.19	4951	582	5884	5614	4582	5058	4140	3265	5372	3470	98.98%
1.0	0.27	4927	554	5881	5604	4555	5038	4100	3222	5359	3436	98.98%

For example, the highest classification rate of 99.01% is obtained by only using approximately 57% of the given training samples. Moreover, only 2035 of the 60000 samples are needed to achieve a classification rate larger than 98% ($\kappa = 0.6$, $\varepsilon = 0.54$, CR= 98.03%). Thus, small numbers of samples, extracted by Algorithm 1, may be sufficient to obtain satisfying or even increased classification rates. The best result without applying kFSA, 98.93%, is obtained when setting $\kappa = 0.7$.

4.2 Fermi–Pasta–Ulam–Tsingou model

As a second example for the application of kFSA, we consider the Fermi–Pasta–Ulam–Tsingou model, see [38], which represents a vibrating string by a system of d coupled oscillators fixed at the end points as shown in Figure 3.

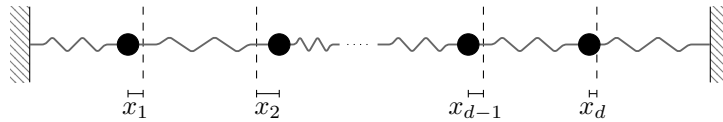


Figure 3: Fermi–Pasta–Ulam–Tsingou model: Representation of a vibrating string by a set of masses coupled by springs.

The governing equations of this dynamical system are given by second-order differential equations of the form

$$\ddot{x}_i = (x_{i+1} - 2x_i + x_{i-1}) + \beta ((x_{i+1} - x_i)^3 - (x_i - x_{i-1})^3), \quad (25)$$

for $i = 1, \dots, d$, where we assume cubic coupling terms between the variables x_i representing the displacements from the stationary positions of the oscillators, see Figure 3. We set $x_0 = x_{d+1} = 0$ since the end points of the chain are supposed to be fixed. The parameter $\beta \in \mathbb{R}$ defines the coupling strength. In [11], we applied MANDy – the tensor-based version of SINDy [10] – to randomly generated training data in order to recover the governing equations (25). Here, we want to consider the same problem, but apply kernel-based regression as described in Section 3.1 using a polynomial kernel. That is, we consider the kernel function

$$k(x, x') = (\kappa + x^\top x')^q, \quad (26)$$

with $\kappa = 1$ and $q = 3$. We construct the data matrices $X, Y \in \mathbb{R}^{d \times m}$ by varying the number of oscillators and generating random displacements in $[-0.1, 0.1]$ for each oscillator as well as the corresponding derivatives according to (25) with $\beta = 0.7$. Due to the random sampling, we expect the (implicitly given) feature space to be sampled sufficiently so that the feature vectors span the whole feature space. Thus, we try to use kFSA to determine the feature space dimension n . Figure 4 shows the number of indices that are extracted by Algorithm 1 when applied to randomly generated training data with different dimensions d and a threshold of 10^{-10} .

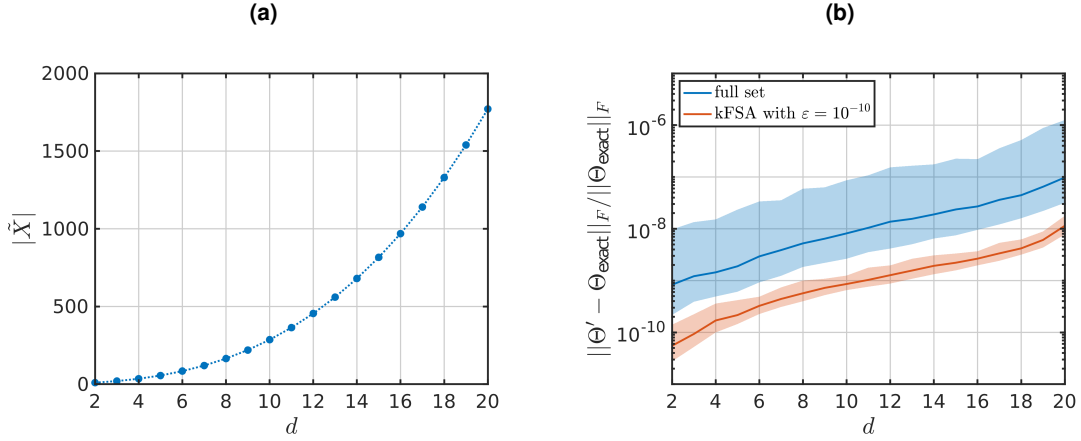


Figure 4: Application of kFSA to the Fermi–Pasta–Ulam–Tsingou problem: (a) Number of samples extracted by Algorithm 1 when applied to 2000 randomly generated d -dimensional displacement vectors. (b) Median of the relative errors of the approximate coefficient matrices over d for the full set as well as the reduced set with 1000 realizations for the former and 100 realizations for latter case. The semi-transparent areas comprise the 5th to the 95th percentile of the respective results.

As we can see in Figure 4 (a), our method extracts less than 2000 samples for each dimension d up to 20. Note that for different realizations of X , the set \tilde{X} varies but the number of samples $|\tilde{X}|$ remains unchanged. The cubic dependency of the number of important samples on d can be explained as follows: For the polynomial kernel, we can directly write down an explicit expression of a feature map. As shown in, e.g., [39], the kernel (26) can be expressed as

$$k(x, x') = \sum_{|p|=q} \left(\sqrt{a_p} \prod_{i=1}^d x_i^{p_i} \right) \left(\sqrt{a_p} \prod_{i=1}^d x_i'^{p_i} \right) = \Psi(x)^\top \Psi(x'), \quad (27)$$

where $p = (p_0, p_1, \dots, p_d) \in \mathbb{N}_0^{d+1}$ is a multi-index and $|p| = p_0 + \dots + p_d$. The prefactors a_p are given in terms of multinomial coefficients:

$$a_p = \binom{q}{p} \alpha^{p_0} = \frac{q!}{p_0! p_1! \dots p_d!} \alpha^{p_0}. \quad (28)$$

Thus, the feature vector $\Psi(x)$ in (27) contains all monomials of order up to and including three in the variables x_i , scaled by the square roots of the prefactors defined in (28). The dimension of the feature space is then given by the number of all possible vectors p with $|p| = q$, which is

$$n = \binom{d+q}{q}.$$

For $q = 3$, this reduces to

$$n = \frac{1}{6}(d+1)(d+2)(d+3),$$

which are the exact same numbers of samples extracted by our method for different dimensions d . That is, Algorithm 1 detects the feature space dimension by choosing the correct amount of feature vectors to form a basis, given in form of column subsets $\tilde{X} \subset X$. As described in Section 3.1, we can use the transformed data matrix $\Psi_{\tilde{X}}$ to reconstruct the coefficient matrix (6) representing the solution of (4), i.e., to recover the governing equations. For a given set $\tilde{X} \subseteq X$, the regression function is defined as

$$f(x) = \Theta \Psi_{\tilde{X}}^{\top} \Psi(x) = \underbrace{\Theta \Psi_{\tilde{X}}^{\top}}_{=\Theta'} \underbrace{D^{-1} \Psi(x)}_{=\Psi'(x)},$$

where Θ is the solution of (10) and D is a diagonal matrix containing the corresponding prefactors given in (28). $\Psi'(x)$ then only contains monomials with prefactor 1 and the exact coefficient matrix defined by (25) becomes

$$\Theta_{\text{exact}} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_2 & x_1 x_2 & x_1^2 x_2 & & x_{d-2} x_d^2 & x_{d-1} x_d^2 & x_d^3 \\ 0 & -2 & 0 & -1.4 & 1 & 0 & 2.1 & & 0 & 0 & 0 \\ 0 & 1 & 0 & 0.7 & -2 & 0 & -2.1 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & & 0 & 0 & 0 \\ & & & \vdots & & & & \ddots & & \vdots & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & -2.1 & 0.7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & & 0 & 2.1 & -1.4 \end{bmatrix}.$$

For the considered values of d and randomly generated training sets, Figure 4 (b) shows that the relative approximation error in the Frobenius norm between Θ' and the exact coefficient matrix is more stable and approximately one order of magnitude lower for the reduced set than for the full set. Particularly, we observe more extreme outliers in the results for $\varepsilon = 0$, where we do not apply any regularization technique, than for $\varepsilon = 10^{-10}$. This indicates that the set generated by kFSA can be used to improve system identification performance since, in addition to reduced numerical costs, we also benefit from regularization effects.

4.3 CalCOFI

The third example is an analysis of time series data taken from the *CalCOFI* (California Cooperative Oceanic Fisheries Investigations) data base [40]. CalCOFI is a partnership of several institutions founded in 1949 to investigate the sardine collapse off California and today provides one of the longest-running time series that exist. These data – containing different seawater measurements such as depth, pressure, and temperature – are the most reliable oceanographic time series available. For the demonstration of our method, we randomly extract a training and a test set from CalCOFI's final upcast CTD (Conductivity, Temperature, Depth) data.² The training set contains 25000 measurement vectors while the test set contains 5000 samples. Various relationships in the data have been examined using supervised learning techniques, see for instance [41, 42, 43]. In this work, the aim is to find a regression function that predicts the dissolved oxygen in the seawater as a function of the depth, pressure, temperature, and salinity.

Without any knowledge of the relationship between the input and output variables, we choose Gaussian kernels for the kernel-based regression, i.e.,

$$k(x, x') = \exp(-\kappa \|x - x'\|_2^2)$$

with $x, x' \in \mathcal{S} \subset \mathbb{R}^4$ and kernel parameter $\kappa > 0$. Since the input variables live on different scales, we normalize these measurements in order to avoid the effect of larger-scale variables dominating the others. That is, given a training data matrix $X = [x^{(1)}, \dots, x^{(m)}] \in \mathbb{R}^{4 \times m}$ we apply *min-max normalization* in the form of

$$\hat{X}_{i,j} = \hat{x}_i^{(j)} = \frac{x_i^{(j)} - l_i}{u_i - l_i} \in [0, 1]$$

²<https://calcofi.org/data/ctd/455-ctd-data-files.html>

for $i = 1, \dots, 4$ and $j = 1, \dots, m$, where

$$l_i = \min_{j \in \{1, \dots, m\}} x_i^{(j)} = \min X_{i,:} \quad \text{and} \quad u_i = \max_{j \in \{1, \dots, m\}} x_i^{(j)} = \max X_{i,:}.$$

In the test phase, we then normalize any given vector by using the same constants before applying the regression function $f: \mathbb{R}^4 \rightarrow \mathbb{R}$ to predict the oxygen concentration.

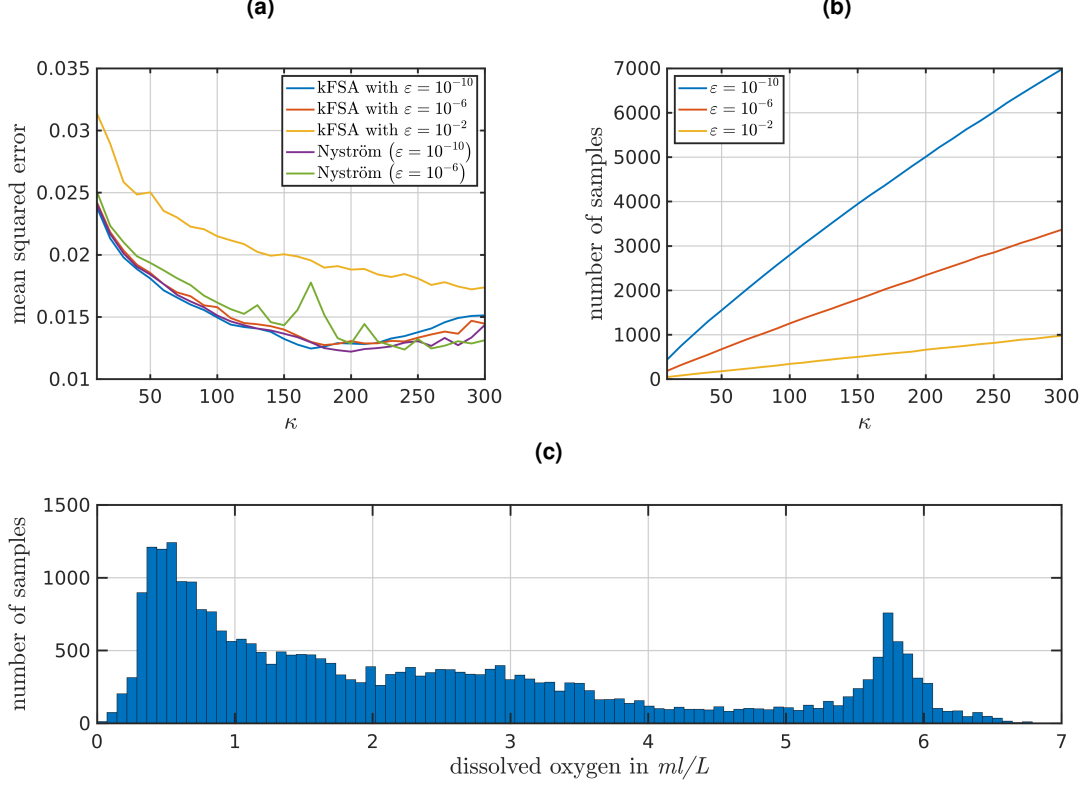


Figure 5: Application of kFSA method to the CalCOFI data set: (a) Mean squared errors in the predicted dissolved oxygen values for different thresholds and kernel parameters. (b) Number of samples extracted by Algorithm 1 for different thresholds and kernel parameters. (c) Histogram of the output data (including training and test set).

The feature space corresponding to the Gaussian kernel is infinite-dimensional and a Gram matrix is always non-singular for mutually different samples. That is, given $X = [x^{(1)}, \dots, x^{(m)}]$ with $x^{(i)} \neq x^{(j)}$ for $i \neq j$, the matrix $G_{X,X}$ has rank m . In particular, all submatrices $G_{\tilde{X},X}$ would therefore have rank $\tilde{m} = |\tilde{X}|$. However, our experiments show that the computed Gram matrices corresponding to subsets $\tilde{X} \subset X$ extracted by kFSA are still ill-conditioned. Thus, we use additional regularization in order to construct well-posed systems of linear equations. For the extraction of relevant samples, we will employ Algorithm 1 with different thresholds ϵ as well as, for comparison, the Nyström method described in Section 3.2. Suppose we have found a subset $\tilde{X} \subset X$, we then solve the minimization problem (10) by constructing a regularized system of linear equations

$$\tilde{\Theta} \left(G_{\tilde{X},X} G_{X,\tilde{X}} + \gamma \text{Id} \right) = Y G_{X,\tilde{X}},$$

where we used the normal equation of (10). Here, the (row) vector $Y \in \mathbb{R}^m$ contains the oxygen levels corresponding to the samples given in the matrix X . The regression function is then defined as

$$f(x) = \tilde{\Theta} \tilde{G}_{\tilde{X},\hat{x}},$$

see (11), where \hat{x} denotes the normalized version of x as described above. The results in form of the mean squared errors in the output variables for $\gamma = 10^{-10}$ and different values for κ as well as ϵ are shown in Figure 5 (a).

The mean squared errors lie between 0.0125 and 0.0314 for any choice of the parameters κ and ε . As we can see in Figure 5 (c), the calculated errors are negligibly small in comparison to the majority of the oxygen measurements in the CalCOFI data set. The results for $\varepsilon = 10^{-10}$ and $\varepsilon = 10^{-6}$ are quite similar, which indicates that even for smaller thresholds we will not get significantly lower errors. This assumption is also supported by the mean squared errors corresponding to the Nyström method proposed in [32], which is provably accurate for any kernel. In order to compare both methods, we used the Nyström method³ to extract subsets of the same size as obtained by applying kFSA with $\varepsilon = 10^{-10}$ and $\varepsilon = 10^{-6}$, respectively, see Figure 5 (b). If we allow the same amount of samples as determined by Algorithm 1 with $\varepsilon = 10^{-10}$, the results of the Nyström method are comparable to the results of Algorithm 1 for thresholds $\varepsilon = 10^{-10}$ and $\varepsilon = 10^{-6}$. If we only allow as many samples as determined by Algorithm 1 with $\varepsilon = 10^{-6}$, we see that the errors are slightly larger and erratic. For smaller numbers of samples, the results obtained by applying the Nyström method start to oscillate – even worse than they already do for $\varepsilon = 10^{-6}$. That is, for small thresholds Algorithm 1 extracts subsets \tilde{X} which are more suitable for a numerically stable regression than the subsets extracted by the Nyström method. As shown in Figure 5 (b), the number of extracted samples depends sublinearly on the kernel parameter κ . For the considered parameter combinations, that number is always smaller than 7000, which corresponds to less than 30% of the given training data.

Note that further improvements in terms of computational complexity of kFSA will be subject to future research. At this point, the runtime of kFSA is not competitive. For instance, the Nyström method we used for comparison requires only $O(mM^2)$ operations instead of $O(m^2 + mM^2)$, where M is the number of extracted samples, see Lemma 5.

5 Conclusion & outlook

We have proposed a data-reduction approach for supervised learning tasks. The resulting method – called *kernel-based feature space approximation* or, in short, kFSA – extracts relevant samples from a given data set whose corresponding feature vectors can be used as a (approximate) basis of the feature space spanned by the transformations of all samples up to an arbitrary accuracy. The aim is reduce the storage consumption and computational complexity as well as to achieve a regularization effect that improves the generalizability of the learned target functions. The introduced method is purely written in terms of kernel functions, which allows us to avoid explicit representations of high-dimensional feature maps and even to consider infinite-dimensional feature spaces.

The performance of kFSA was demonstrated using examples of classification and regression problems from different scientific areas such as image recognition, system identification, and time-series analysis. In our experiments, we observed that the computational effort may be reduced significantly while obtaining even better results than when considering all training data in the learning phase. Moreover, we showed that kFSA is able to detect the correct dimension of a feature space that is only implicitly given by a kernel. We also found cases where sample subsets extracted by kFSA can lead to numerically more stable results than subsets found by a state-of-the-art implementation of the Nyström method.

Future research will include the application of kFSA to a broader spectrum of problems in the field of data-driven analysis of dynamical systems, e.g., transfer operator approximation, model reduction, and system identification. Other open issues are algorithmic improvements for the bottom-up construction of relevant subsets, imposing additional or alternative constraints such as sparsity of the coefficient matrices, and the artificial generation of input-output pairs to further reduce the number of needed basis vectors in the feature space.

Acknowledgments

This research has been funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1114 “Scaling Cascades in Complex Systems” (project ID: 235221301, project B06).

³<https://github.com/cnmusco/recursive-nystrom>

References

- [1] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91:045002, 2019. doi:10.1103/RevModPhys.91.045002.
- [2] H. M. Cartwright. *Machine Learning in Chemistry: The Impact of Artificial Intelligence*. Theoretical and Computational Chemistry Series. The Royal Society of Chemistry, Cambridge, UK, 2020. doi:10.1039/9781839160233.
- [3] R. C. Deo. Machine learning in medicine. *Circulation*, 132(20):1920–1930, 2015. doi:10.1161/CIRCULATIONAHA.115.001593.
- [4] J. I. Glaser, A. S. Benjamin, R. Farhoodi, and K. P. Kording. The roles of supervised machine learning in systems neuroscience. *Progress in Neurobiology*, 175:126–137, 2019. doi:10.1016/j.pneurobio.2019.01.008.
- [5] A. M. Ozbayoglu, M. U. Gudelek, and O. B. Sezer. Deep learning for financial applications: A survey. *Applied Soft Computing*, 93:106384, 2020. doi:10.1016/j.asoc.2020.106384.
- [6] S. Klus and P. Gelß. Tensor-based algorithms for image classification. *Algorithms*, 12(11), 2019. doi:10.3390/a12110240.
- [7] S. H. S. Basha, S. R. Dubey, V. Pulabaigari, and S. Mukherjee. Impact of fully connected layers on performance of convolutional neural networks for image classification. *Neurocomputing*, 378:112–119, 2020. doi:10.1016/j.neucom.2019.10.008.
- [8] B. Alghamdi, Y. Xu, and J. Watson. A hybrid approach for detecting spammers in online social networks. In H. Hacid, W. Cellary, H. Wang, H.-Y. Paik, and R. Zhou, editors, *Web information systems engineering – WISE 2018*, pages 189–198, Basel, CH, 2018. Springer International Publishing. doi:10.1007/978-3-030-02922-7_13.
- [9] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa. Machine learning for email spam filtering: Review, approaches and open research problems. *Heliyon*, 5(6):e01802, 2019. doi:10.1016/j.heliyon.2019.e01802.
- [10] S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113:3932–3937, 2016. doi:10.1073/pnas.1517384113.
- [11] P. Gelß, S. Klus, J. Eisert, and C. Schütte. Multidimensional approximation of nonlinear dynamical systems. *Journal of Computational and Nonlinear Dynamics*, 14(6):061006, 2019. doi:10.1115/1.4043148.
- [12] S. Tufféry. *Data Mining and Statistics for Decision Making*. John Wiley & Sons, Ltd, Hoboken, NJ, USA, 2011. doi:10.1002/9780470979174.
- [13] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998. doi:10.1162/089976698300017467.
- [14] S. Harmeling, A. Ziehe, M. Kawanabe, and K.-R. Müller. Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5):1089–1124, 2003. doi:10.1162/089976603765202677.
- [15] M. O. Williams, C. W. Rowley, and I. G. Kevrekidis. A kernel-based method for data-driven Koopman spectral analysis. *Journal of Computational Dynamics*, 2(2):247–265, 2015. doi:10.3934/jcd.2015005.
- [16] S. Klus, I. Schuster, and K. Muandet. Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces. *Journal of Nonlinear Science*, 2019. doi:10.1007/s00332-019-09574-z.
- [17] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, 2002.
- [18] P. Drineas and M. W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.
- [19] K. Potdar, T. S. Pardawala, and C. D. Pai. A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, 175(4):7–9, 2017. doi:10.5120/ijca2017915495.
- [20] G. Bonaccorso. *Machine Learning Algorithms: A Reference Guide to Popular Algorithms for Data Science and Machine Learning*. Packt Publishing, Birminham, UK, 2017.
- [21] A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications*. CMS Books in Mathematics. Springer Science+Business Media, New York, NY, USA, 2nd edition, 2003. doi:10.1007/b97366.

- [22] J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London*, 209:415–446, 1909. doi:<https://doi.org/10.1098/rsta.1909.0016>.
- [23] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, MA, USA, 2012.
- [24] G. H. Golub and C. F. van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. The Johns Hopkins University Press, Baltimore, MD, USA, 4th edition, 2013.
- [25] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in neural information processing systems 20*, NIPS 2007, pages 1177–1184. Curran Associates, Inc., Red Hook, NY, USA, 2008.
- [26] Q. Le, T. Sarlos, and A. Smola. Fastfood – Computing Hilbert space expansions in loglinear time. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *ICML 2013*, pages 244–252, Atlanta, GA, USA, 2013.
- [27] F. Litzinger, L. Boninsegna, H. Wu, F. Nüske, R. Patel, R. Baraniuk, F. Noé, and C. Clementi. Rapid calculation of molecular kinetics using compressed sensing. *Journal of Chemical Theory and Computation*, 14(5):2771–2783, 2018. doi:[10.1021/acs.jctc.8b00089](https://doi.org/10.1021/acs.jctc.8b00089).
- [28] G. Santin and B. Haasdonk. Kernel methods for surrogate modeling. *ArXiv e-prints*, 2019. arXiv:abs/1907.10556.
- [29] S. L. Campbell and C. D. Meyer. *Generalized Inverses of Linear Transformations*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2009. doi:[10.1137/1.9780898719048](https://doi.org/10.1137/1.9780898719048).
- [30] S. Wang and Z. Zhang. Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling. *Journal of Machine Learning Research*, 14(47):2729–2769, 2013.
- [31] C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS 2000, pages 661–667, Cambridge, MA, USA, 2000. MIT Press.
- [32] C. Musco and C. Musco. Recursive sampling for the Nyström method. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS 2017, pages 3836–3848, Red Hook, NY, USA, 2017. Curran Associates, Inc.
- [33] Solving interpretable kernel dimension reduction. *Journal of Computational Dynamics*, 2(2):247–265, 2015. doi:[10.3934/jcd.2015005](https://doi.org/10.3934/jcd.2015005).
- [34] Y. Tian and Y. Takane. More on generalized inverses of partitioned matrices with Banachiewicz—Schur forms. *Linear Algebra and its Applications*, 430(5):1641–1655, 2009. doi:[10.1016/j.laa.2008.06.007](https://doi.org/10.1016/j.laa.2008.06.007).
- [35] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi:[10.1109/5.726791](https://doi.org/10.1109/5.726791).
- [36] E. M. Stoudenmire and D. J. Schwab. Supervised learning with tensor networks. *Neural Information Processing Systems 29*, pages 4799–4807, 2016.
- [37] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 1st edition, 2008.
- [38] E. Fermi, J. Pasta, and S. Ulam. Studies of nonlinear problems. Technical Report LA-1940, Los Alamos Scientific Lab, 1955.
- [39] M. J. Zaki and W. Meira, Jr. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, Cambridge, UK, 2014. doi:[10.1017/CB09780511810114](https://doi.org/10.1017/CB09780511810114).
- [40] California Cooperative Oceanic Fisheries Investigations. CalCOFI hydrographic database, 2020. URL: <https://calcofi.org/ccdata.html>.
- [41] S. R. Alin, R. A. Feely, A. G. Dickson, J. M. Hernández-Ayón, L. W. Juranek, M. D. Ohman, and R. Goericke. Robust empirical relationships for estimating the carbonate system in the southern California current system and application to CalCOFI hydrographic cruise data (2005–2011). *Journal of Geophysical Research: Oceans*, 117(C5), 2012. doi:[10.1029/2011JC007511](https://doi.org/10.1029/2011JC007511).
- [42] S. Y. Kim and B. D. Cornuelle. Coastal ocean climatology of temperature and salinity off the Southern California Bight: Seasonal variability, climate index correlation, and linear trend. *Progress in Oceanography*, 138:136–157, 2015. doi:[10.1016/j.pocean.2015.08.001](https://doi.org/10.1016/j.pocean.2015.08.001).
- [43] C. M. Sakamoto, K. S. Johnson, L. J. Coletti, T. L. Maurer, G. Massion, J. T. Pennington, J. N. Plant, H. W. Jannasch, and F. P. Chavez. Hourly in situ nitrate on a coastal mooring: A 15-year record and insights into new production. *Oceanography*, 30(4):114–127, 2017. doi:[10.5670/oceanog.2017.428](https://doi.org/10.5670/oceanog.2017.428).