EURASIAN MATHEMATICAL JOURNAL
volume 17 , number 1, 2026

EURASIAN NATIONAL UNIVERSITY

ISSN (Print): 2077-9879
ISSN (Online): 2617-2658

EURASIAN MATHEMATICAL JOURNAL

CONTENTS

ISSN 2077-9879

ISSN (Print): 2077-9879
ISSN (Online): 2617-2658

Eurasian Mathematical Journal

2026, Volume 17, Number 1

Founded in 2010 by
the L.N. Gumilyov Eurasian National University
in cooperation with
the M.V. Lomonosov Moscow State University
the Peoples' Friendship University of Russia (RUDN University)
the University of Padua

Starting with 2018 co-funded
by the L.N. Gumilyov Eurasian National University
and
the Peoples' Friendship University of Russia (RUDN University)

Supported by the ISAAC
(International Society for Analysis, its Applications and Computation)
and
by the Kazakhstan Mathematical Society

Published by

the L.N. Gumilyov Eurasian National University
Astana, Kazakhstan

EURASIAN MATHEMATICAL JOURNAL

Aims and Scope

The Eurasian Mathematical Journal (EMJ) publishes carefully selected original research papers in all areas of mathematics written by mathematicians, principally from Europe and Asia. However papers by mathematicians from other continents are also welcome.

From time to time the EMJ publishes survey papers.

The EMJ publishes 4 issues in a year.

The language of the paper must be English only.

The contents of the EMJ are indexed in Scopus, Web of Science (ESCI), Mathematical Reviews, MathSciNet, Zentralblatt Math (ZMATH), Referativnyi Zhurnal – Matematika, Math-Net.Ru.

The EMJ is included in the list of journals recommended by the Committee for Control of Education and Science (Ministry of Education and Science of the Republic of Kazakhstan) and in the list of journals recommended by the Higher Attestation Commission (Ministry of Education and Science of the Russian Federation).

Information for the Authors

Submission. Manuscripts should be written in LaTeX and should be submitted electronically in DVI, PostScript or PDF format to the EMJ Editorial Office through the provided web interface (www.enu.kz).

When the paper is accepted, the authors will be asked to send the tex-file of the paper to the Editorial Office.

The author who submitted an article for publication will be considered as a corresponding author. Authors may nominate a member of the Editorial Board whom they consider appropriate for the article. However, assignment to that particular editor is not guaranteed.

Copyright. When the paper is accepted, the copyright is automatically transferred to the EMJ. Manuscripts are accepted for review on the understanding that the same work has not been already published (except in the form of an abstract), that it is not under consideration for publication elsewhere, and that it has been approved by all authors.

Title page. The title page should start with the title of the paper and authors' names (no degrees). It should contain the Keywords (no more than 10), the Subject Classification (AMS Mathematics Subject Classification (2010) with primary (and secondary) subject classification codes), and the Abstract (no more than 150 words with minimal use of mathematical symbols).

Figures. Figures should be prepared in a digital form which is suitable for direct reproduction.

References. Bibliographical references should be listed alphabetically at the end of the article. The authors should consult the Mathematical Reviews for the standard abbreviations of journals' names.

Authors' data. The authors' affiliations, addresses and e-mail addresses should be placed after the References.

Proofs. The authors will receive proofs only once. The late return of proofs may result in the paper being published in a later issue.

Offprints. The authors will receive offprints in electronic form.

Publication Ethics and Publication Malpractice

For information on Ethics in publishing and Ethical guidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the EMJ implies that the work described has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The EMJ follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/NewCode.pdf). To verify originality, your article may be checked by the originality detection service CrossCheck http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the EMJ.

The Editorial Board of the EMJ will monitor and safeguard publishing ethics.

The procedure of reviewing a manuscript, established by the Editorial Board of the Eurasian Mathematical Journal

1. **Reviewing procedure**

1.1. All research papers received by the Eurasian Mathematical Journal (EMJ) are subject to mandatory reviewing.

1.2. The Managing Editor of the journal determines whether a paper fits to the scope of the EMJ and satisfies the rules of writing papers for the EMJ, and directs it for a preliminary review to one of the Editors-in-chief who checks the scientific content of the manuscript and assigns a specialist for reviewing the manuscript.

1.3. Reviewers of manuscripts are selected from highly qualified scientists and specialists of the L.N. Gumilyov Eurasian National University (doctors of sciences, professors), other universities of the Republic of Kazakhstan and foreign countries. An author of a paper cannot be its reviewer.

1.4. Duration of reviewing in each case is determined by the Managing Editor aiming at creating conditions for the most rapid publication of the paper.

1.5. Reviewing is confidential. Information about a reviewer is anonymous to the authors and is available only for the Editorial Board and the Control Committee in the Field of Education and Science of the Ministry of Education and Science of the Republic of Kazakhstan (CCFES). The author has the right to read the text of the review.

1.6. If required, the review is sent to the author by e-mail.

1.7. A positive review is not a sufficient basis for publication of the paper.

1.8. If a reviewer overall approves the paper, but has observations, the review is confidentially sent to the author. A revised version of the paper in which the comments of the reviewer are taken into account is sent to the same reviewer for additional reviewing.

1.9. In the case of a negative review the text of the review is confidentially sent to the author.

1.10. If the author sends a well reasoned response to the comments of the reviewer, the paper should be considered by a commission, consisting of three members of the Editorial Board.

1.11. The final decision on publication of the paper is made by the Editorial Board and is recorded in the minutes of the meeting of the Editorial Board.

1.12. After the paper is accepted for publication by the Editorial Board the Managing Editor informs the author about this and about the date of publication.

1.13. Originals reviews are stored in the Editorial Office for three years from the date of publication and are provided on request of the CCFES.

1.14. No fee for reviewing papers will be charged.

2. **Requirements for the content of a review**

2.1. In the title of a review there should be indicated the author(s) and the title of a paper.

2.2. A review should include a qualified analysis of the material of a paper, objective assessment and reasoned recommendations.

2.3. A review should cover the following topics:

- compliance of the paper with the scope of the EMJ;
- compliance of the title of the paper to its content;
- compliance of the paper to the rules of writing papers for the EMJ (abstract, key words and phrases, bibliography etc.);
- a general description and assessment of the content of the paper (subject, focus, actuality of the topic, importance and actuality of the obtained results, possible applications);
- content of the paper (the originality of the material, survey of previously published studies on the topic of the paper, erroneous statements (if any), controversial issues (if any), and so on);
- exposition of the paper (clarity, conciseness, completeness of proofs, completeness of bibliographic references, typographical quality of the text);

- possibility of reducing the volume of the paper, without harming the content and understanding of the presented scientific results;

- description of positive aspects of the paper, as well as of drawbacks, recommendations for corrections and complements to the text.

2.4. The final part of the review should contain an overall opinion of a reviewer on the paper and a clear recommendation on whether the paper can be published in the Eurasian Mathematical Journal, should be sent back to the author for revision or cannot be published.

Web-page

The web-page of the EMJ is www.emj.enu.kz. One can enter the web-page by typing Eurasian Mathematical Journal in any search engine (Google, Yandex, etc.). The archive of the web-page contains all papers published in the EMJ (free access).

Subscription

Subscription index of the EMJ 76090 via KAZPOST.

E-mail

eurasianmj@yandex.kz

The Eurasian Mathematical Journal (EMJ)
The Astana Editorial Office
The L.N. Gumilyov Eurasian National University
Building no. 3
Room 306a
Tel.: +7-7172-709500 extension 33312
13 Kazhymukan St
010008 Astana, Republic of Kazakhstan

The Moscow Editorial Office
The Patrice Lumumba Peoples' Friendship University of Russia
(RUDN University)
Room 473
3 Ordzonikidze St
117198 Moscow, Russian Federation

BOKAYEV NURZHAN ADILKHANOVICH

(to the 70th birthday)

January 5, 2026, marks the 70th birthday of Nurzhan Adilkhanovich Bokayev, Doctor of Physical and Mathematical Sciences (1996), Professor (2002), member of the Editorial Board of the Eurasian Mathematical Journal (2010).

Nurzhan Adilkhanovich Bokayev was born on 5 January, 1956 in the village of Urnek, Karabalyk District, Kostanay Region. He graduated in 1972, with a gold medal from the Burlin Secondary School in the district. That same year, he entered the Mathematics Department of Karaganda State University and graduated with honors in 1977. From 1978 to 1979, he served in the Soviet Army. In 1980, he completed an internship, and from 1981 to 1984, he studied in the graduate program at Lomonosov Moscow State University in the Department of Function Theory and Functional Analysis. In 1985, he defended his candidate's dissertation there under the supervision of Corresponding Member of the Academy of Sciences of the USSR D.E. Menshov and Professor V.A. Skvortsov. In 1996, he defended his doctoral dissertation, "Fourier Coefficients and Uniqueness Theorems for Series in Generalized Walsh and Haar Systems", at the Institute of Mathematics of the Ministry of Education and Science of the Republic of Kazakhstan, speciality Mathematical Analysis (01.01.01).

After completing his postgraduate studies, he worked as a lecturer, senior lecturer, associate professor, and professor in the Department of Mathematical Analysis at E.A. Buketov Karaganda State University (1985-1999). He headed the Department of Mathematics and Mathematical Modeling (1996-1999), and was a dean of the Faculty of Mathematics at E.A. Buketov Karaganda State University (1999-2005). Since 2005, he has been a professor in the Faculty of Mechanics and Mathematics at the L.N. Gumilyov Eurasian National University. From 2009 to 2018, he was the Head of the Department of Higher Mathematics at the L.N. Gumilyov Eurasian National University, and from 2018 to the present, he has been a professor in the Department of Fundamental Mathematics.

Professor Bokaev's research focuses on problems in function theory and functional analysis, the theory of orthogonal series for generalized Walsh and Haar systems, and operator theory in various function spaces. He has proved renewal and uniqueness theorems for series with respect to periodic multiplicative systems and Haar-type systems, and constructed continual sets of uniqueness (U-sets) and sets of non-uniqueness (M-sets) for multiplicative systems. He obtained conditions for functions to belong to various functional classes in terms of the Fourier coefficients of generalized Haar and Walsh systems, and embedding criteria for Nikol'skii-Besov spaces constructed on the basis of multiplicative systems. He also obtained conditions for the boundedness and compactness of the commutator of the Riesz potential in general Morrey-type spaces, and conditions for boundedness of generalizedRiesz and Bessel potentials and generalized fractional-maximal operators in rearrangement-invariant spaces.

His co-authors include Professor V.A. Skvortsov (Moscow State University, Moscow), Professors V.I. Burenkov and M.L. Goldman (Peoples' Friendship University of Russia (RUDN University). Moscow), Dr. A. Gogatishvili (Institute of Mathematics of the Czech Academy of Sciences, Prague). His doctoral students' foreign advisors include Professors W. Sickel (Friedrich-Schiller-Universitity, Jena, Germany), Massimo Lanza de Cristoforis (University of Padova, Padova, Italy), V. Ruzhansky (Ghent University, Ghent, Belgium), U. Goginava (United Arab Emirates University, Al Ain, United Arab Emirates), and E. Panakhov (Institute of Applied Mathematics at Baku State University, Baku, Azerbaijan).

Under his supervision, 15 dissertations (4 candidate's and 11 PhD) were defended. He has published over 220 scientific papers, 2 monographs and 2 textbooks.

He is a three-time recipient of the state grant "Best University Teacher" of the Republic of Kazakhstan (2006, 2010, 2024) and served as Vice President of the Mathematical Society of Turkic-Speaking Countries (2014-2023). He was awarded the "For Contribution to the Development of Science" badge (2022).

Over the last ten years, he has been and continues to be a head of more than 5 national and international funded projects.

The Editorial Board of the Eurasian Mathematical Journal, his friends and colleagues cordially congratulate Nurzhan Adilkhanovich on the occasion of his 70th birthday and wish him good health, happiness and new achievements in mathematics and mathematical education.

EURASIAN MATHEMATICAL JOURNAL
ISSN 2077-9879
Volume 17, Number 1 (2026), 58 – 76

COMBINING UNSUPERVISED DIMENSION REDUCTION WITH SUFFICIENT DIMENSION REDUCTION

Zh. Mukanov, A. Sharafudinov, R. Takhanov, A. Bekembayev

Communicated by K.T. Mynbayev

Key words: unsupervised dimension reduction, sufficient dimension reduction, complex measures, hybrid setting.

AMS Mathematics Subject Classification: 62-07, 62H25, 68T10.

Abstract. We present a new method for dimension reduction that combines unsupervised dimension reduction (UDR) with sufficient dimension reduction (SDR). In unsupervised dimension reduction the goal is to find a low-dimensional linear subspace that approximates the support of a data distribution. If data is supervised, then in sufficient dimension reduction the goal is to find a low-dimensional linear subspace, called the effective subspace, such that the projection of an input vector onto that subspace maximally captures information on correlations between an input and an output.

The objective that we suggest to minimize consists of two parts. The first one is responsible for the UDR part, it forces a low-dimensional probabilistic measure μ to approximate a distribution over inputs. The second one is responsible for the SDR part, it forces a regression function f to be consistent with supervised data. Additionally, we require the support of μ and the effective subspace of f to be equal. In this hybrid setting we solve two problems, UDR and SDR, so that the UDR term serves as a regularizer of the SDR term.

We reformulate the problem as an optimization task of finding a k-dimensional linear subspace S and a pair of complex measures (μ, μ') supported in S. Instead of optimizing over complex measures, we suggest minimizing over ordinary functions (g_1, g_2) but with an additional term R that penalizes a distortion of the common support of g_1, g_2 from a k-dimensional linear subspace. The algorithm that we develop can be formulated for functions (g_1, g_2) as well as for their inverse Fourier transforms. Eventually, we report results of numerical experiments on well-known datasets.

DOI: https://doi.org/10.32523/2077-9879-2026-17-1-58-76

1 Introduction

Unsupervised dimension reduction (UDR) is a classical problem in data science that has many non-equivalent formulations coming from different contexts, such as principal component analysis [13], factor analysis, linear multidimensional scaling [10], Fisher's linear discriminant analysis [12], canonical correlations analysis [18], sufficient dimension reduction (SDR) [14], maximum autocorrelation factors [29], slow feature analysis [43], kernel methods [27, 14, 34], methods based on autoencoders [38, 3] and more.

In UDR we are given a finite number of points in $\mathbb{R}^n$ (sampled according to some unknown distribution) and our goal is to find a "low-dimensional" affine (or linear) subspace that approximates "the support" of the distribution. As was pointed out in [35], the study field currently achieved a saturation level at which developing a unifying framework to the problem becomes highly demanding.

In SDR (sometimes called supervised dimension reduction), we are given a finite number of pairs $(\mathbf{x}_i, y_i), \mathbf{x}_i \in \mathbb{R}^n, y_i \in \mathbb{R}$, also generated according to some unknown joint distribution $p(\mathbf{x}, y)$, and our goal is to find k vectors (where $k \ll n$) $\mathbf{w}_1, \cdots, \mathbf{w}_k \in \mathbb{R}^n$ such that symbolically:

$$y \perp\!\!\!\perp \mathbf{x} | \mathbf{w}_1^T\mathbf{x}, \cdots, \mathbf{w}_k^T\mathbf{x}.$$

This means that an output y is conditionally independent of $\mathbf{x}$, given $\mathbf{w}_1^T\mathbf{x}, \cdots, \mathbf{w}_k^T\mathbf{x}$ or, the conditional distribution $p(y|\mathbf{x})$ is the same as $p(y|\mathbf{w}_1^T\mathbf{x}, \cdots, \mathbf{w}_k^T\mathbf{x})$.

Obviously, the last formulation is not precise if we do not make any assumptions about the joint distribution, or more specifically about the conditional distribution $p(y|\mathbf{x})$. Typically, it is assumed that

$$y = g(\mathbf{w}_1^T\mathbf{x}, \cdots, \mathbf{w}_k^T\mathbf{x}) + \varepsilon, \tag{1.1}$$

where ε is the Gaussian noise with $\mathbb{E}\varepsilon = 0$ and $\mathbb{E}\varepsilon^2 = \delta^2$. The function g is an unknown smooth function. Then, the function $f(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}] = g(\mathbf{w}_1^T\mathbf{x}, \cdots, \mathbf{w}_k^T\mathbf{x})$ is called the regression function.

Many methods have been proposed for estimation of the parameters of model (1.1) such as: sliced inverse regression [22],[9], methods based on an analysis of gradient and Hessian of the regression function [23, 44, 26], methods based on combining local classifiers [17, 28], kernel-based methods [14] and more. In such methods for the SDR problem as the Sliced Inverse Regression [22], the Principal Hessian Direction [23], the Sliced Average Variance Estimation [6], an effective subspace is recovered from the Singular Value Decomposition applied to a certain matrix that is constructed from a training set in a straightforward way. Other methods, such as the Principal Fitted Components [7], the Likelihood Acquired Direction [8], the Kernel Dimensionality Reduction [14], are based on analytic expressions measuring the affinity of a k-dimensional subspace to the effective subspace. In the second type of methods the SDR problem is reduced to an optimization problem over the Stiefel manifold, or the Grassmanian. For other methods we refer to a tutorial on SDR methods [15]. Again, an important aspect of all these methods is that, given a fixed effective subspace, the regression function that predicts an output variable has a relatively straightforward structure and is not optimized by any additional supervised learning procedure.

The SDR problem is tightly connected with the unsupervised dimension reduction problem. In [42] it was shown how a method originally developed for SDR can be turned into a UDR method, i.e. applied to unsupervised data by simply setting an output to be equal to an input. In [11], the SDR problem, together with UDR problems, is cast as an optimization problem over the Stiefel manifold. Taking into account deep connections between UDR and SDR problems, the current study's goal is to develop an approach to a hybrid setting, i.e. when we target to find a low-dimensional linear subspace that both approximates the support of data and is close to the effective subspace (that allows to predict an output). Note that one can solve these two problems independently and obtain two solutions — the span of their union maintains some of their desirable properties (at the expense of increasing the dimension to $2k$). The goal of the paper is to develop better approaches to the problem.

In the hybrid setting our goal is to approximate the empirical inputs distribution μ_{emp} by a distribution μ, supported in some k-dimensional subspace L, and find the regression function $f = g(\mathbf{w}_1^T\mathbf{x}, \cdots, \mathbf{w}_k^T\mathbf{x})$ such that the effective subspace $\text{span}\,(\mathbf{w}_1, \cdots, \mathbf{w}_k)$ is equal to L. This will guarantee that after projecting the input vector $\mathbf{x}$ onto L we are changing the geometry of the dataset only slightly and, additionally, do not lose the ability to predict the output y. The objective that is minimized in this setting consists of two parts, the first considering the dependence on μ and the second considering the dependence on f. Thus, for a target space L there is a trade-off between the requirement to support the whole data and the sufficiency to predict y. Note that if the main goal is SDR, then the first part of the objective can be interpreted as a regularization part that is dedicated to avoiding over-fitting. If alternatively, the main goal is UDR, then the second part can play its role in many interesting contexts. For example, let the output y be an indicator of an

outlier, i.e., $y = 1$ indicates that the input $\mathbf{x}$ is an outlier. Then, it is desirable that after projecting onto the low-dimensional subspace L we are still able to distinguish outliers from typical points.

The key observation of our analysis, stated in Theorem 3.1 of Section 3, is that a class of functions of the form $g(\mathbf{w}_1^T\mathbf{x}, \cdots, \mathbf{w}_k^T\mathbf{x})$ can be characterized as functions whose Fourier transforms are supported in a k-dimensional linear subspace. Thus, the main problem in the hybrid setting is to find a probabilistic measure μ that approximates the empirical measure μ_{emp} and the regression function f such that μ and $\mathcal{F}[f]$ are both supported in the same k-dimensional linear subspace. Instead of optimizing over generalized functions with a k-dimensional support (or, k-dimensional complex measures, in our terminology), we suggest minimizing over ordinary functions given as feed-forward neural networks but with an additional soft constraint. To force the function's support to be close to a k-dimensional subspace, in Section 4 we introduce a class of penalty functions R such that large values of R indicate a strong distortion of the support from any k-dimensional linear subspace. For a specific case of R, in Section 5 we develop an algorithm for our problem that can be formulated for functions given in the frequency coordinate form as well as in the initial coordinate form. The last section is dedicated to experiments.

2 Preliminaries

Throughout the paper we will use common terminology and notations from functional analysis. The Schwartz space of functions is denoted by $\mathcal{S}(\mathbb{R}^n)$ and the set of all tempered distributions is $\mathcal{S}'(\mathbb{R}^n)$, the dual space of $\mathcal{S}(\mathbb{R}^n)$. The Fourier and the inverse Fourier transforms are first defined by

$$\mathcal{F}[f](\boldsymbol{\xi}) = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} f(\mathbf{x}) e^{-i\boldsymbol{\xi}^T\mathbf{x}} d\mathbf{x},$$

$$\mathcal{F}^{-1}[f](\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} f(\boldsymbol{\xi}) e^{i\boldsymbol{\xi}^T\mathbf{x}} d\boldsymbol{\xi},$$

and then extended to continuous bijective linear operators $\mathcal{F}, \mathcal{F}^{-1} : \mathcal{S}'(\mathbb{R}^n) \to \mathcal{S}'(\mathbb{R}^n)$. A Borel complex measure is a mapping $\mu : \mathcal{E} \to \mathbb{C}$ where $\mathcal{E}$ is a sigma-algebra of all Borel sets on $\mathbb{R}^n$ and μ is sigma-additive. If μ is real-valued, then μ is called a finite Borel measure. A finite Borel measure with $\mu(\mathbb{R}^n) = 1$ is called the Borel probabilistic measure. The set of all Borel probabilistic measures on $\mathbb{R}^n$ is denoted by $\mathfrak{B}(\mathbb{R}^n)$. The symbol $X_1, \cdots, X_m \sim^{iid} \mu$ denotes the fact that random variables $X_1, \cdots, X_m$ are all independent and each has a distribution function μ.

If a function $f : \mathbb{R}^n \to \mathbb{C}$ is such that $\int_{\mathbb{R}^n} f(\mathbf{x})u(\mathbf{x})d\mathbf{x} < \infty$ for any $u \in \mathcal{S}(\mathbb{R}^n)$ then it induces an operator $T_f : \mathcal{S}(\mathbb{R}^n) \to \mathbb{C}$, where $T_f(u) = \int_{\mathbb{R}^n} f(\mathbf{x})u(\mathbf{x})d\mathbf{x}$. Analogously, a Borel complex measure μ on $\mathbb{R}^n$ defines a tempered distribution $T_\mu : \mathcal{S}(\mathbb{R}^n) \to \mathbb{C}$, where $T_\mu(u) = \int_{\mathbb{R}^n} u(\mathbf{x})d\mu$. For simplicity of our notation, we use f and T_f (μ and T_μ) interchangeably (from the context it will alway be clear what we mean). By $L_2(\mathbb{R}^n)$ we denote the Hilbert space of all square-integrable functions from $\mathbb{R}^n$ to $\mathbb{C}$, with the inner product $\langle u, v\rangle_{L_2(\mathbb{R}^n)} = \int u(\mathbf{x})^* v(\mathbf{x})d\mu$. The induced norm is then $\|u\|_{L_2(\mathbb{R}^n)} = \sqrt{\langle u, u\rangle}$.

A positive-definite function is a complex-valued function $f : \mathbb{R}^n \mapsto \mathbb{C}$ such that for any real numbers $\mathbf{x}_1, \cdots, \mathbf{x}_s$ the matrix $A = (a_{i,j})_{i,j=1}^n$ where $a_{i,j} = f(\mathbf{x}_i - \mathbf{x}_j)$ is positive-semidefinite.

For a matrix $A = \left[a_{ij}\right]_{1 \le i,j \le n}$ the Frobenius norm is $\|A\|_F = \sqrt{\sum_{ij} |a_{ij}|^2}$.

3 Problem formulation

We formulate the hybrid dimension reduction problem as an optimization task:

$$\inf_{(\mu,f)\in\mathfrak{D}} (1-\rho)I(\mu, \mu_{\text{emp}}) + \rho J(f). \tag{3.1}$$

In the last expression $\mu \in \mathfrak{B}(\mathbb{R}^n)$ denotes a Borel probabilistic measure on $\mathbb{R}^n$ that approximates the empirical distribution over inputs, μ_{emp}, and I denotes the distance function between measures (e.g. the maximum mean discrepancy or the Wasserstein distance) and $\rho \in [0, 1]$.

The object $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth real-valued function that can be given in an arbitrary form, keeping in mind the case of f defined by a feed-forward neural network. We assume that f is a candidate for the regression function and $J(f)$ is a cost function that values how strongly f fits in this role. In practice for the regression case we use the following cost function:

$$J(f) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)} |y_i - f(\mathbf{x}_i + \boldsymbol{\epsilon})|^2,$$

where the target variable y is normalized, i.e. the estimator of outputs variance $\hat{\text{Var}}(y) = 1$. We add the last remark only to make $I(\mu, \mu_{\text{emp}})$ and $J(f)$ to be of the same scale. For the binary classification case with 0-1 outputs we use:

$$J(f) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)} H\left(y_i, \frac{e^{f(\mathbf{x}_i + \boldsymbol{\epsilon})}}{1 + e^{f(\mathbf{x}_i + \boldsymbol{\epsilon})}}\right),$$

where $H(y, p) = -y \log p - (1 - y) \log(1 - p)$. Thus, $\rho = 0$ corresponds to the pure unsupervised dimension reduction task and $\rho = 1$ corresponds to the pure sufficient dimension reduction task.

We assume that f satisfies (for k fixed in advance):

$$f(\mathbf{x}) = g(\mathbf{w}_1^T \mathbf{x}, \cdots, \mathbf{w}_k^T \mathbf{x}),$$

where g is an arbitrary function and $\mathbf{w}_1, \cdots, \mathbf{w}_k \in \mathbb{R}^n$. Thus, given an input $\mathbf{x}$, the corresponding output depends on the projection of $\mathbf{x}$ onto $\text{span}(\mathbf{w}_1, \cdots, \mathbf{w}_k)$. Thus, $\text{span}(\mathbf{w}_1, \cdots, \mathbf{w}_k)$ serves as the effective subspace. We assume that the measure μ is also supported in that subspace. Thus, we define:

$$\begin{aligned} \mathfrak{D} = \{(\mu, f) | \mu \in \mathfrak{B}(\mathbb{R}^n), \underset{\mathbb{R}^n}{\exists} \mathbf{w}_1, \cdots, \mathbf{w}_k \; \exists g : \mathbb{R}^k \to \mathbb{R} \\ \text{such that} \underset{\text{Borel}}{\forall} A \; \mu(A) = \mu(A \cap \text{span}(\mathbf{w}_1, \cdots, \mathbf{w}_k)) \\ \text{and} \underset{\mathbb{R}^n}{\forall} \mathbf{x} \; f(\mathbf{x}) = g(\mathbf{w}_1^T \mathbf{x}, \cdots, \mathbf{w}_k^T \mathbf{x})\}. \end{aligned} \tag{3.2}$$

The parameter $\rho \geq 0$ regulates how strongly we prefer the sufficiency term J over the distance till the empirical distribution.

The following theorem is the key observation behind our approach to the problem (3.1).

Theorem 3.1. *A function $k(\mathbf{x})$ can be represented as $k(\mathbf{x}) = g(\mathbf{w}_1^T \mathbf{x}, \cdots, \mathbf{w}_k^T \mathbf{x}), g \in \mathcal{S}(\mathbb{R}^k)$ if and only if there is an orthonormal basis $\{\mathbf{a}_1, \cdots, \mathbf{a}_n\} \subseteq \mathbb{R}^n$ such that:*

$$\mathcal{F}[T_l] = r(\mathbf{a}_1^T \mathbf{x}, \cdots, \mathbf{a}_{k'}^T \mathbf{x}) \prod_{i=k'+1}^{n} \delta(\mathbf{a}_i^T \mathbf{x}), r \in \mathcal{S}(\mathbb{R}^k), \tag{3.3}$$

where $\delta(\cdot)$ is the Dirac delta-function and $\text{span}(\mathbf{a}_1, \cdots, \mathbf{a}_{k'}) = \text{span}(\mathbf{w}_1, \cdots, \mathbf{w}_k)$.

Sketch of the proof. Without loss of generality, we can assume that $\mathbf{w}_1, \cdots, \mathbf{w}_k$ are linearly independent. A rigorous proof of the theorem would require a careful checking of certain integral identities. Instead, we will present a sketch of the proof at the abstraction level common to theoretical physics papers.

($\Rightarrow$) We also can assume that $\mathbf{w}_1, \cdots, \mathbf{w}_k$ are orthonormal. Indeed, after every redefinition of g given by the rule $g(s_1, \cdots, s_k) \leftarrow g(s_1, \cdots, s_i + \alpha s_j, \cdots, s_k)$ we get the same function l if we simultaneously transform $\mathbf{w}_i$ to $\mathbf{w}_i - \alpha \mathbf{w}_j$. By making such redefinitions, we can always orthogonalize $\mathbf{w}_1, \cdots, \mathbf{w}_k$ by the Gramm-Schmidt process with a subsequent scaling of g's arguments.

Let us complete $\mathbf{w}_1, \cdots, \mathbf{w}_k$ with $\mathbf{w}_{k+1}, \cdots, \mathbf{w}_n$ to form an orthonormal basis in $\mathbb{R}^n$ and set:

$$Q = \left[\mathbf{w}_1, \cdots, \mathbf{w}_n\right] = \left[Q_1, Q_2\right], Q_1 \in \mathbb{R}^{n \times k}, Q_2 \in \mathbb{R}^{n \times (n-k)}.$$

Then, in the Fourier transform formula we make the change of variables $\mathbf{x} = Q \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = Q_1 \mathbf{y}_1 + Q_2 \mathbf{y}_2$, $\mathbf{y}_1 \in \mathbb{R}^k$, $\mathbf{y}_2 \in \mathbb{R}^{n-k}$ and get:

$$\mathcal{F}[l](\boldsymbol{\xi}) = \frac{1}{\sqrt{2\pi}^n} \int_{\mathbb{R}^n} g(\mathbf{w}_1^T \mathbf{x}, \cdots, \mathbf{w}_k^T \mathbf{x}) e^{-i \boldsymbol{\xi}^T \mathbf{x}} d\mathbf{x}$$

$$= \frac{1}{\sqrt{2\pi}^n} \int_{\mathbb{R}^n} g(\mathbf{y}_1) e^{-i \boldsymbol{\xi}^T Q \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}} d\mathbf{y}_1 d\mathbf{y}_2 = \frac{1}{\sqrt{2\pi}^n} \int_{\mathbb{R}^n} g(\mathbf{y}_1) e^{-i (Q_1^T \boldsymbol{\xi})^T \mathbf{y}_1 - i (Q_2^T \boldsymbol{\xi})^T \mathbf{y}_2} d\mathbf{y}_1 d\mathbf{y}_2$$

$$= \frac{1}{\sqrt{2\pi}^n} \int_{\mathbb{R}^k} g(\mathbf{y}_1) e^{-i (Q_1^T \boldsymbol{\xi})^T \mathbf{y}_1} d\mathbf{y}_1 \int_{\mathbb{R}^{n-k}} e^{-i (Q_2^T \boldsymbol{\xi})^T \mathbf{y}_2} d\mathbf{y}_2 = \sqrt{2\pi}^{n-k} \mathcal{F}[g](Q_1^T \boldsymbol{\xi}) \delta^{n-k}(Q_2^T \boldsymbol{\xi}),$$

where $\delta^{n-k}(s_1, \cdots, s_{n-k}) = \prod_{i=1}^{n-k} \delta(s_i)$. Here we used the equality $\int_{\mathbb{R}^{n-k}} e^{-i \mathbf{z}^T \mathbf{y}_2} d\mathbf{y}_2 = (2\pi)^{n-k} \delta^{n-k}(\mathbf{z})$. Thus, we obtain the needed representation.

($\Leftarrow$) Suppose that:

$$\mathcal{F}[l] = r(\mathbf{a}_1^T \mathbf{x}, \cdots, \mathbf{a}_{k'}^T \mathbf{x}) \prod_{i=k'+1}^{n} \delta(\mathbf{a}_i^T \mathbf{x}).$$

Using the inverse Fourier transform we get:

$$l(\boldsymbol{\xi}) = \mathcal{F}^{-1}\left[\mathcal{F}[l]\right](\boldsymbol{\xi}) = \frac{1}{\sqrt{2\pi}^n} \int_{\mathbb{R}^n} r(\mathbf{a}_1^T \mathbf{x}, \cdots, \mathbf{a}_{k'}^T \mathbf{x}) \prod_{i=k'+1}^{n} \delta(\mathbf{a}_i^T \mathbf{x}) e^{i \mathbf{x}^T \boldsymbol{\xi}} d\mathbf{x}.$$

After the change of variables $\mathbf{x} = O\mathbf{y}$, where $O = \left[\mathbf{a}_1, \cdots, \mathbf{a}_n\right]$, we get:

$$l(\boldsymbol{\xi}) = \frac{1}{\sqrt{2\pi}^n} \int_{\mathbb{R}^n} r(y_{1:k'}) \prod_{i=k'+1}^{n} \delta(y_i) e^{i \sum_{i=1}^n y_i \mathbf{a}_i^T \boldsymbol{\xi}} dy_{1:n}$$

$$= \frac{1}{\sqrt{2\pi}^n} \int_{\mathbb{R}^n} r(y_{1:k'}) e^{i \sum_{i=1}^{k'} y_i \mathbf{a}_i^T \boldsymbol{\xi}} dy_{1:k'} = \frac{1}{\sqrt{2\pi}^{n-k}} \tilde{g}(\mathbf{a}_1^T \boldsymbol{\xi}, \cdots, \mathbf{a}_k^T \boldsymbol{\xi}),$$

where $\tilde{g} = \mathcal{F}^{-1}[r]$. □

Substantively, the theorem claims that if the function's value depends only on the projection of an argument $\mathbf{x}$ onto the span$(\mathbf{w}_1, \cdots, \mathbf{w}_k)$, then frequencies of the spectrum of such function are all in the span$(\mathbf{w}_1, \cdots, \mathbf{w}_k)$.

Definition 1. Let $\mathfrak{C}(\mathbb{R}^n)$ be a set of symmetric Borel complex measures on $\mathbb{R}^n$, i.e., any $\mu \in \mathfrak{C}(\mathbb{R}^n)$ is a sigma-additive complex-valued function on the sigma-algebra of Borel subsets of $\mathbb{R}^n$ and $\mu(-A) = \mu(A)^*$ (where x^* denotes the complex conjugate of x). We will call a measure $\mu \in \mathfrak{C}(\mathbb{R}^n)$ *a k-dimensional measure* if there is a k'-dimensional linear subspace $S \subseteq \mathbb{R}^n$, $k' \leq k$ such that $\mu(A \cap S) = \mu(A)$ for any Borel set A. The minimal linear space S with the last property is called the support of μ and is denoted by $\operatorname{supp} \mu$. The set of all k-dimensional Borel complex measures is denoted by $\mathcal{G}_k$. The set of all pairs (μ_1, μ_2), where $\mu_1, \mu_2 \in \mathcal{G}_k$ and $\operatorname{supp} \mu_1 = \operatorname{supp} \mu_2$, is denoted by $\mathcal{G}_k^2$.

Thus, problem (3.1) is equivalent to the followinq equality:

$$\inf_{(\mu,\mathcal{F}^{-1}(f))\in\mathfrak{D}}(1-\rho)I(\mu,\mu_{\text{emp}})+\rho J(\mathcal{F}^{-1}(f))$$
$$=\inf_{(\mu,\mu')\in\mathcal{G}_k^2,\mu\in\mathfrak{B}(\mathbb{R}^n)}(1-\rho)I(\mu,\mu_{\text{emp}})+\rho J(\mathcal{F}^{-1}(\mu')).$$

Instead of minimization over tempered distributions, we will relax the property that the common support of the pair (μ,μ') is k-dimensional, reducing the problem to the following one:

$$(1-\rho)I(\mu,\mu_{\text{emp}})+\rho J(\mathcal{F}^{-1}[\mu'])\to\min_{(\mu,\mu')\in\mathfrak{B}(\mathbb{R}^n)\times\mathfrak{C}(\mathbb{R}^n),R(\mu,\mu')\le\epsilon},\tag{3.4}$$

where $R(\mu,\mu')$ is a penalty term that penalizes (μ,μ') if the dimension of their common support is greater than k. In the next section we describe one natural approach to construct such a penalty term R.

4 Penalty function

We need the following theorem.

Theorem 4.1. *Let $\gamma:\mathbb{R}^n\to\mathbb{R}$ be a positive definite continuous function and $(\mu,\mu')\in\mathfrak{B}(\mathbb{R}^n)\times\mathfrak{C}(\mathbb{R}^n)$ be such that*

$$\forall i,j\ \int_{\mathbb{R}^n\times\mathbb{R}^n}x_i\gamma(\mathbf{x}-\mathbf{y})y_j d\mu(\mathbf{x})d\mu(\mathbf{y})<\infty,$$
$$\int_{\mathbb{R}^n\times\mathbb{R}^n}x_i\gamma(\mathbf{x}-\mathbf{y})y_j d\mu'(\mathbf{x})^*d\mu'(\mathbf{y})<\infty.$$

The pair (μ,μ') is in $\mathcal{G}_k^2$ if and only if

$$\operatorname{rank}(\mathcal{M})\le k,$$

where

$$\mathcal{M}=a\int_{\mathbb{R}^n\times\mathbb{R}^n}\mathbf{x}\gamma(\mathbf{x}-\mathbf{y})\mathbf{y}^T d\mu(\mathbf{x})d\mu(\mathbf{y})$$
$$+b\int_{\mathbb{R}^n\times\mathbb{R}^n}\mathbf{x}\gamma(\mathbf{x}-\mathbf{y})\mathbf{y}^T d\mu'(\mathbf{x})^*d\mu'(\mathbf{y}),a>0,b>0.$$

Proof. ($\Rightarrow$) If $(\mu,\mu')\in\mathcal{G}_k^2$, then there is a k-dimensional linear subspace $S\subseteq\mathbb{R}^n$ such that $\mu(A\cap S)=\mu(A)$, $\mu'(A\cap S)=\mu'(A)$. Let $\{\mathbf{v}_i\}_{i=1}^n$ be an orthonormal basis in $\mathbb{R}^n$ such that $\mathbf{v}_i\perp S, i>k$. Then:

$$\int_{\mathbb{R}^n\times\mathbb{R}^n}\mathbf{x}\gamma(\mathbf{x}-\mathbf{y})\mathbf{y}^T d\mu(\mathbf{x})d\mu(\mathbf{y})=\int_{\mathbb{R}^n\times\mathbb{R}^n}\sum_{i=1}^n\mathbf{v}_i\mathbf{v}_i^T\mathbf{x}\gamma(\mathbf{x}-\mathbf{y})\mathbf{y}^T d\mu(\mathbf{x})d\mu(\mathbf{y}).$$

Since $\int_{\mathbb{R}^n}(\mathbf{v}_i^T\mathbf{x})^2\gamma(\mathbf{x}-\mathbf{y})d\mu(\mathbf{x})=0, i>k$, then:

$$\int_{\mathbb{R}^n\times\mathbb{R}^n}\sum_{i=1}^k\mathbf{v}_i\mathbf{v}_i^T\mathbf{x}\gamma(\mathbf{x}-\mathbf{y})\mathbf{y}^T d\mu(\mathbf{x})d\mu(\mathbf{y})=\sum_{i=1}^k\mathbf{v}_i\mathbf{v}_i^T\int_{\mathbb{R}^n\times\mathbb{R}^n}\mathbf{x}\gamma(\mathbf{x}-\mathbf{y})\mathbf{y}^T d\mu(\mathbf{x})d\mu(\mathbf{y}).$$

The same can be proven for μ', therefore:

$$\mathcal{M}=\sum_{i=1}^k\mathbf{v}_i\mathbf{v}_i^T\mathcal{M}$$

and we see that $\operatorname{rank}(\mathcal{M}) \leq k$.

($\Leftarrow$) If $\operatorname{rank}(\mathcal{M}) \leq k$, then there exist linearly independent vectors $\{\mathbf{u}_i\}_{i=k+1}^{n}$ such that

$$\mathbf{u}_i^T \mathcal{M} \mathbf{u}_i = a \int_{\mathbb{R}^n \times \mathbb{R}^n} \mathbf{u}_i^T \mathbf{x} \gamma(\mathbf{x}-\mathbf{y}) \mathbf{u}_i^T \mathbf{y} d\mu(\mathbf{x}) d\mu(\mathbf{y})$$
$$+ b \int_{\mathbb{R}^n \times \mathbb{R}^n} \mathbf{u}_i^T \mathbf{x} \gamma(\mathbf{x}-\mathbf{y}) \mathbf{u}_i^T \mathbf{y} d\mu'(\mathbf{x})^* d\mu'(\mathbf{y}) = 0.$$

From positive definiteness of γ we conclude that $\mu(\{\mathbf{x} | \mathbf{u}_i^T \mathbf{x} \neq 0\}) = 0$ and $\mu'(\{\mathbf{x} | \mathbf{u}_i^T \mathbf{x} \neq 0\}) = 0$. Therefore, $\mu(A) = \mu(A \cap \{\mathbf{x} | \mathbf{u}_i^T \mathbf{x}, i = \overline{k+1, n}\}) = 0$ and $\mu'(A) = \mu'(A \cap \{\mathbf{x} | \mathbf{u}_i^T \mathbf{x} = 0, i = \overline{k+1, n}\}) = 0$. Thus, $\operatorname{supp} \mu \subseteq \{\mathbf{x} | \mathbf{u}_i^T \mathbf{x} = 0, i = \overline{k+1, n}\}$ and $\operatorname{supp} \mu' \subseteq \{\mathbf{x} | \mathbf{u}_i^T \mathbf{x} = 0, i = \overline{k+1, n}\}$. □

Let us define

$$\mathcal{M}_\nu = \int_{\mathbb{R}^n \times \mathbb{R}^n} \mathbf{x} \gamma(\mathbf{x}-\mathbf{y}) \mathbf{y}^T d\nu(\mathbf{x})^* d\nu(\mathbf{y})$$

and

$$\mathcal{M}_{(\mu,\mu')} = (1-\rho)\mathcal{M}_\mu + \rho \mathcal{M}_{\mu'}.$$

Note that $\mathcal{M}_{(\mu,\mu')}$ is a positive semidefinite matrix, and therefore, the square root $\mathcal{M}_{(\mu,\mu')}^{1/2}$ is defined. Our definition for the penalty function R is as follows:

$$R(\mu, \mu') = \min_{\mathcal{M} \in \mathbb{R}^{n \times n} : \operatorname{rank}(\mathcal{M}) \leq k} \| \mathcal{M}_{(\mu,\mu')}^{1/2} - \mathcal{M} \|_F^2. \tag{4.1}$$

It is natural to expect that if $R(\mu, \mu') \leq \epsilon$ and $\epsilon > 0$ is small, i.e., $\mathcal{M}_{(\mu,\mu')}^{1/2}$ (together with $\mathcal{M}_{(\mu,\mu')}$) is close to some rank k matrix, then the common support of (μ, μ') is approximable by a k-dimensional linear subspace. Now, our goal is to develop an algorithm for the following problem:

$$(1-\rho) I(\mu, \mu_{\text{emp}}) + \rho J(\mathcal{F}^{-1}[\mu']) + \lambda R(\mu, \mu') \to \min_{(\mu,\mu') \in \mathfrak{B}(\mathbb{R}^n) \times \mathfrak{C}(\mathbb{R}^n)}, \tag{4.2}$$

where λ is a penalty parameter that can be chosen sufficiently large to force $R(\mu, \mu')$ to be small.

4.1 Another description of the penalty

Let us now give an alternative description of the penalty $R(\mu, \mu')$ that suits better to the goal of designing an algorithm for problem (3.4).

Let γ and s be smooth real-valued positive definite functions such that:

$$\gamma(\mathbf{x} - \mathbf{x}'') = \int_{\mathbb{R}^n} s(\mathbf{x} - \mathbf{x}') s(\mathbf{x}' - \mathbf{x}'') d\mathbf{x}'.$$

For example, $\gamma(\mathbf{x}) = \sqrt{\frac{\pi}{2}}^n e^{-\|\mathbf{x}\|^2/2}$ and $s(\mathbf{x}) = e^{-\|\mathbf{x}\|^2}$. Note that $\gamma(\mathbf{x}) = \gamma(-\mathbf{x})$ and $s(\mathbf{x}) = s(-\mathbf{x})$.

Let us define

$$S_{(\mu,\mu')} = [\int_{\mathbb{R}^n} s(\mathbf{x} - \mathbf{x}') \mathbf{x}' d\mu(\mathbf{x}') \int_{\mathbb{R}^n} s(\mathbf{x} - \mathbf{x}') \mathbf{x}' d\mu'(\mathbf{x}')] E,$$

where

$$E = \begin{bmatrix} \sqrt{1-\rho} & 0 \\ 0 & \sqrt{\rho} \end{bmatrix}.$$

This object is an $n \times 2$ matrix whose entries are functions.

By $L_2^{n \times 2}(\mathbb{R}^n)$ we denote the space of all matrices $\left[b_{ij}(\mathbf{x})\right]_{1 \leq i \leq n, j=1,2}$, where $b_{ij} \in L_2(\mathbb{R}^n)$. Also, $L_2^2(\mathbb{R}^n) \equiv L_2^{2 \times 1}(\mathbb{R}^n)$. Note that $L_2^2(\mathbb{R}^n)$ is a linear space over complex numbers. Let us also denote

by $\tilde{L}_2^2(\mathbb{R}^n)$ the real linear space that is the set $L_2^2(\mathbb{R}^n)$ considered over real numbers only, equipped with the inner product

$$\langle [\phi_1, \phi_2]^T, [\psi_1, \psi_2]^T \rangle_{\tilde{L}_2^2(\mathbb{R}^n)} = \mathrm{Re}\left\{ \langle \phi_1, \psi_1 \rangle_{L_2(\mathbb{R}^n)} + \langle \phi_2, \psi_2 \rangle_{L_2(\mathbb{R}^n)} \right\}.$$

The set of all bounded linear operators from $\tilde{L}_2^2(\mathbb{R}^n)$ to $\mathbb{R}^n$ is denoted by $\mathcal{B}^{n\times 2}$.

It is easy to see that any $A \in L_2^{n\times 2}(\mathbb{R}^n)$ defines a bounded linear operator O_A from $\tilde{L}_2^2(\mathbb{R}^n)$ to $\mathbb{R}^n$ by the following rule:

$$\begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} \in \tilde{L}_2^2(\mathbb{R}^n) \to^{O_A} \mathrm{Re} \int_{\mathbb{R}^n} A(\mathbf{x})^* \begin{bmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \end{bmatrix} d\mathbf{x}.$$

Moreover, it is easy to see that all bounded linear operators from $\tilde{L}_2^2(\mathbb{R}^n)$ to $\mathbb{R}^n$ can be represented in this way. Obviously, $\mathcal{B}^{n\times 2}$ is a Hilbert space, where the inner product is defined as:

$$\langle O_{A_1}, O_{A_2} \rangle_{\mathcal{B}^{n\times 2}} = \mathrm{Re} \int_{\mathbb{R}^n} \mathrm{Trace}\,(A_1(\mathbf{x})^\dagger A_2(\mathbf{x})) d\mathbf{x},$$

where for any $A(\mathbf{x}) = [A_{ij}(\mathbf{x})]_{1\le i\le n, 1\le j\le 2} \in L_2^{n\times 2}(\mathbb{R}^n)$, $A(\mathbf{x})^\dagger$ denotes the matrix $[A_{ji}(\mathbf{x})^*]_{1\le i\le 2, 1\le j\le n} \in L_2^{2\times n}(\mathbb{R}^n)$. Thus, the corresponding norm coincides with the trace norm. Recall that, for a bounded linear operator $O : \mathcal{H}_1 \to \mathcal{H}_2$ between Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$, the rank of O is defined as $\dim \mathrm{Im}(O)$, where $\mathrm{Im}(O) = \{O[\phi] \big| \phi \in \mathcal{H}_1\}$.

Theorem 4.2. *If* $\mathcal{M}_{(\mu,\mu')} < \infty$, *then* $S_{(\mu,\mu')} \in L_2^{n\times 2}(\mathbb{R}^n)$, $O_{S_{(\mu,\mu')}} O^\dagger_{S_{(\mu,\mu')}} = \mathcal{M}_{(\mu,\mu')}$, *and* $\mathrm{rank}(O_{S_{(\mu,\mu')}}) = \mathrm{rank}(\mathcal{M}_{(\mu,\mu')})$.

Sketch of the proof. Since

$$\langle O^\dagger_{S_{(\mu,\mu')}} \mathbf{y}, \begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} \rangle_{\tilde{L}_2^2(\mathbb{R}^n)} = \mathbf{y}^T (O_{S_{(\mu,\mu')}} \begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix})$$
$$= \mathbf{y}^T \mathrm{Re} \int_{\mathbb{R}^n} [\int_{\mathbb{R}^n} s(\mathbf{x}-\mathbf{x}')\mathbf{x}' d\mu(\mathbf{x}') \int_{\mathbb{R}^n} s(\mathbf{x}-\mathbf{x}')\mathbf{x}' d\mu'(\mathbf{x}')^*] E \begin{bmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \end{bmatrix} d\mathbf{x},$$

we obtain

$$O^\dagger_{S_{(\mu,\mu')}} \mathbf{y} = E \begin{bmatrix} \int_{\mathbb{R}^n} s(\mathbf{x}-\mathbf{x}')\mathbf{x}'^T \mathbf{y} d\mu(\mathbf{x}') \\ \int_{\mathbb{R}^n} s(\mathbf{x}-\mathbf{x}')\mathbf{x}'^T \mathbf{y} d\mu'(\mathbf{x}') \end{bmatrix}.$$

Let us now check that $O_{S_{(\mu,\mu')}} O^\dagger_{S_{(\mu,\mu')}} = \mathcal{M}_{(\mu,\mu')}$ by direct calculation:

$$\mathbf{y} \in \mathbb{R}^n \overset{O^\dagger_{S_{(\mu,\mu')}}}{\longrightarrow} E \begin{bmatrix} \int_{\mathbb{R}^n} s(\mathbf{x}-\mathbf{x}')\mathbf{x}'^T \mathbf{y} d\mu(\mathbf{x}') \\ \int_{\mathbb{R}^n} s(\mathbf{x}-\mathbf{x}')\mathbf{x}'^T \mathbf{y} d\mu'(\mathbf{x}') \end{bmatrix}$$
$$\overset{O_{S_{(\mu,\mu')}}}{\longrightarrow} \mathrm{Re} \int_{\mathbb{R}^n} [\int_{\mathbb{R}^n} s(\mathbf{x}-\mathbf{x}')\mathbf{x}' d\mu(\mathbf{x}') \int_{\mathbb{R}^n} s(\mathbf{x}-\mathbf{x}')\mathbf{x}' d\mu'(\mathbf{x}')^*]$$
$$EE \begin{bmatrix} \int_{\mathbb{R}^n} s(\mathbf{x}-\mathbf{x}')\mathbf{x}'^T \mathbf{y} d\mu(\mathbf{x}') \\ \int_{\mathbb{R}^n} s(\mathbf{x}-\mathbf{x}')\mathbf{x}'^T \mathbf{y} d\mu'(\mathbf{x}') \end{bmatrix} d\mathbf{x}$$
$$= (1-\rho) \int_{\mathbb{R}^n} \gamma(\mathbf{x}-\mathbf{x}')\mathbf{x}\mathbf{x}'^T \mathbf{y} d\mu(\mathbf{x}) d\mu(\mathbf{x}') + \rho \int_{\mathbb{R}^n} \gamma(\mathbf{x}-\mathbf{x}')\mathbf{x}\mathbf{x}'^T \mathbf{y} d\mu'(\mathbf{x})^* d\mu'(\mathbf{x}')$$
$$= \mathcal{M}_{(\mu,\mu')} \mathbf{y}.$$

From the equality $O_{S_{(\mu,\mu')}}O^{\dagger}_{S_{(\mu,\mu')}} = \mathcal{M}_{(\mu,\mu')}$ we conclude that $\operatorname{rank}(\mathcal{M}_{(\mu,\mu')}) \leq \operatorname{rank}(O_{S_{(\mu,\mu')}})$. Conversely, if $\mathbf{x}_1, \cdots, \mathbf{x}_r$ is a basis of $\operatorname{Im}(\mathcal{M}_{(\mu,\mu')})^{\perp}$, then $0 = \mathbf{x}_i^T \mathcal{M}_{(\mu,\mu')}\mathbf{x}_i = \|O^{\dagger}_{S_{(\mu,\mu')}}\mathbf{x}_i\|^2_{\tilde{L}^2_2(\mathbb{R}^n)}$, i.e. $O^{\dagger}_{S_{(\mu,\mu')}}\mathbf{x}_i = \mathbf{0}$. Therefore, for $i \in [r]$, we have

$$\langle O_{S_{(\mu,\mu')}}\phi, \mathbf{x}_i\rangle_{\tilde{L}^2_2(\mathbb{R}^n)} = \langle \phi, O^{\dagger}_{S_{(\mu,\mu')}}\mathbf{x}_i\rangle_{\tilde{L}^2_2(\mathbb{R}^n)} = 0,$$

for any ϕ, i.e., $\operatorname{Im}(O_{S_{(\mu,\mu')}}) \subseteq \{\mathbf{x}_1, \cdots, \mathbf{x}_r\}^{\perp}$. Thus, $\operatorname{rank}(O_{S_{(\mu,\mu')}}) \leq n - r = \operatorname{rank}(\mathcal{M}_{(\mu,\mu')})$. □

The Eckart-Young-Mirsky theorem [19, Theorem 4.4.7] in the theory of Singular Value Decomposition (SVD) gives us that

$$R(\mu,\mu') = \min_{\mathcal{M}\in\mathbb{R}^{n\times n}:\operatorname{rank}(\mathcal{M})\leq k} \|\mathcal{M}^{1/2}_{(\mu,\mu')} - \mathcal{M}\|^2_F = \sum_{i=k+1}^{n} \lambda_i,$$

where $\lambda_1 \geq \cdots \geq \lambda_n$ are eigenvalues of $\mathcal{M}_{(\mu,\mu')} = \mathcal{M}^{\frac{1}{2}T}_{(\mu,\mu')}\mathcal{M}^{\frac{1}{2}}_{(\mu,\mu')}$. Due to the relationship $O_{S_{(\mu,\mu')}}O^{\dagger}_{S_{(\mu,\mu')}} = \mathcal{M}_{(\mu,\mu')}$ the following statement is true.

Theorem 4.3. *We have*

$$R(\mu,\mu') = \min_{S\in L^{n\times 2}_2(\mathbb{R}^n):\operatorname{rank}(O_S)\leq k} \|S_{(\mu,\mu')} - S\|^2_{L^{n\times 2}_2(\mathbb{R}^n)},$$

and the latter minimum is attained at $P_{(\mu,\mu')}S_{(\mu,\mu')}$, where $P_{(\mu,\mu')} = \sum_{i=1}^{k}\mathbf{u}_i\mathbf{u}_i^T$ is the projection operator to first k principal components of $\mathcal{M}_{(\mu,\mu')}$.

Sketch of the proof. First, we check that all arguments of the Eckart-Young-Mirsky theorem for matrices maintain in the case of bounded linear operators from $\tilde{L}^2_2(\mathbb{R}^n)$ to $\mathbb{R}^n$. Indeed, all arguments survive, because such operators are compact and can have only a finite spectrum, since $\mathbb{R}^n$ is finite-dimensional. Let us only describe an optimal S on which $\min\limits_{S\in L^{n\times 2}_2(\mathbb{R}^n):\operatorname{rank}(O_S)\leq k} \|S_{(\mu,\mu')} - S\|^2_{L^{n\times 2}_2(\mathbb{R}^n)}$ is attained.

Let $\mathbf{u}_1, \cdots, \mathbf{u}_n$ be orthonormal eigenvectors of $\mathcal{M}_{(\mu,\mu')} = O_{S_{(\mu,\mu')}}O^{\dagger}_{S_{(\mu,\mu')}}$ and $\lambda_1 \geq \cdots \geq \lambda_{n'} > 0$ be the corresponding non-zero eigenvalues. For $\sigma_i = \sqrt{\lambda_i}$ let us define $\mathbf{v}_i = \frac{O^{\dagger}_{S_{(\mu,\mu')}}[\mathbf{u}_i]}{\sigma_i}$. Here, $\mathbf{v}_i$ is equal to the function

$$\mathbf{v}_i(\mathbf{x}) = \frac{1}{\sigma_i}\begin{bmatrix}\sqrt{1-\rho}\int_{\mathbb{R}^n} s(\mathbf{x}-\mathbf{x}')\mathbf{u}_i^T\mathbf{x}'d\mu(\mathbf{x}') \\ \sqrt{\rho}\int_{\mathbb{R}^n} s(\mathbf{x}-\mathbf{x}')\mathbf{u}_i^T\mathbf{x}'d\mu'(\mathbf{x}')\end{bmatrix}.$$

It is easy to see that $\mathbf{v}_1, \cdots \mathbf{v}_{n'}$ is an orthonormal basis in $\operatorname{Im} O^{\dagger}_{S_{(\mu,\mu')}}$, and SVD for $O_{S_{(\mu,\mu')}}$ is:

$$O_{S_{(\mu,\mu')}} = \sum_{i=1}^{n'} \sigma_i \mathbf{u}_i\mathbf{v}_i^{\dagger}.$$

By the Eckart-Young-Mirsky theorem, an optimal $S = O_F$ in $\min\limits_{S\in\mathcal{B}^{n\times 2},\operatorname{rank}S\leq k} \|O_{S_{(\mu,\mu')}} - S\|^2_{\mathcal{B}^{n\times 2}}$ is defined by a truncation of SVD for $O_{S_{(\mu,\mu')}}$ at kth term, i.e.,

$$\begin{aligned} F = \Bigg[\sqrt{1-\rho}\int_{\mathbb{R}^n} s(\mathbf{x}-\mathbf{x}')\sum_{i=1}^{k}\mathbf{u}_i\mathbf{u}_i^T\mathbf{x}'d\mu(\mathbf{x}'), \\ \sqrt{\rho}\int_{\mathbb{R}^n} s(\mathbf{x}-\mathbf{x}')\sum_{i=1}^{k}\mathbf{u}_i\mathbf{u}_i^T\mathbf{x}'d\mu'(\mathbf{x}')\Bigg] = P_{(\mu,\mu')}S_{(\mu,\mu')}, \end{aligned} \tag{4.3}$$

where $P_{(\mu,\mu')} = \sum_{i=1}^{k}\mathbf{u}_i\mathbf{u}_i^T$ is the projection operator to first k principal components of $\mathcal{M}_{(\mu,\mu')}$. From the equality $\|O_A\|_{\mathcal{B}^{n\times 2}} = \|A\|_{L^{n\times 2}_2(\mathbb{R}^n)}$ we obtain the statement of theorem. □

5 The alternating scheme

Complex measures are in the weak closure of the main functional classes, L_2 and Sobolev spaces, continuous functions, single-layer neural networks, etc. The major gain from penalty formulation (4.2) is that, instead of optimizing over complex measures whose supports have empty interior, we can vary the argument over a space of ordinary functions $\mathfrak{F}$, where $\mathfrak{F} \subset C(\mathbb{R}^n)$ can be chosen as any class of functions which is dense (with respect to the weak weak topology) in $\mathfrak{D}' = \mathfrak{B}(\mathbb{R}^n) \times \mathfrak{C}(\mathbb{R}^n)$, for example:

$$\mathfrak{F} = \{\mathbf{g} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} | g_1, g_2 \in C(\mathbb{R}^n), \forall \mathbf{x}\, g_1(\mathbf{x}) \geq 0, \int_{\mathbb{R}^n} g_1(\mathbf{x}) d\mathbf{x} = 1, \int_{\mathbb{R}^n} |g_2(\mathbf{x})| d\mathbf{x} < \infty\}. \tag{5.1}$$

Thus, we need to solve the following optimization problem:

$$(1-\rho) I(g_1, \mu_{\text{emp}}) + \rho J(\mathcal{F}^{-1}[g_2]) + \lambda R(g_1, g_2) \to \min_{(g_1,g_2)\in\mathfrak{F}}. \tag{5.2}$$

Using Theorem 4.3 we can represent problem (5.2) in the following form:

$$\Phi(\mathbf{g} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}, S) = (1-\rho) I(g_1, \mu_{\text{emp}}) + \rho J(\mathcal{F}^{-1}[g_2]) + \lambda \|S_{(g_1,g_2)} - S\|^2_{L_2^{n\times 2}(\mathbb{R}^n)} \to \min_{\mathbf{g}\in\mathfrak{F}, S\in L_2^{n\times 2}(\mathbb{R}^n):\operatorname{rank}(O_S)\leq k}.$$

The simplest idea for an optimization is to minimize over $\mathbf{g} = (g_1, g_2) \in \mathfrak{F}$ and over $S \in L_2^{n\times 2}(\mathbb{R}^n) : \operatorname{rank}(O_S) \leq k$ alternatingly. The first part would be an optimization over an infinite-dimensional object, which cannot be implemented in practice. To avoid infiniteness, we will fix a proper parameterized set of functions $\mathfrak{M} \subseteq \mathfrak{F}$ (e.g., deep neural networks [36, 1, 2, 39, 40, 24]) and optimize over $\mathfrak{M}$. A general scheme of optimization is given in Algorithm 1.

Algorithm 1 Alternating scheme

1: **procedure**
2: $S_0 \leftarrow 0$
3: **for** $t = 1, \cdots, N$ **do**
4: $\mathbf{g}_t \leftarrow \arg\min_{\mathbf{g}\in\mathfrak{M}} \Phi(\mathbf{g}, S_{t-1})$
5: $S_t \leftarrow \arg \min_{S\in L_2^{n\times 2}(\mathbb{R}^n):\operatorname{rank}(O_S)\leq k} \|S_{\mathbf{g}_t} - S\|^2_{L_2^{n\times 2}(\mathbb{R}^n)}$

Note that Step 5 of the algorithm is equivalent to minimizing $\|S_{\mathbf{g}_t} - S\|^2_{L_2^{n\times 2}(\mathbb{R}^n)}$ over $S \in L_2^{n\times 2}(\mathbb{R}^n) : \operatorname{rank}(O_S) \leq k$. In the previous section we have already described an optimal solution for that task (equation (4.3)):

$$S_t = P_{\mathbf{g}_t} S_{\mathbf{g}_t},$$
$$S_{\mathbf{g}_t} = \int_{\mathbb{R}^n} s(\mathbf{x}-\mathbf{x}')\mathbf{x}'\mathbf{g}_t^T(\mathbf{x}')E d\mathbf{x}'.$$

Here $P_{\mathbf{g}_t} \in \mathbb{R}^{n\times n}$ is the projection operator to the first k principal components of

$$\mathcal{M}_{\mathbf{g}_t} = \int_{\mathbb{R}^n\times\mathbb{R}^n} \gamma(\mathbf{x}-\mathbf{y})\mathbf{x}\mathbf{y}^T \mathbf{g}_t(\mathbf{x})^\dagger E^2 \mathbf{g}_t(\mathbf{y}) d\mathbf{x} d\mathbf{y}.$$

The hardest part of that step is to estimate the matrix $\mathcal{M}_{\mathbf{g}_t}$ for a given $\mathbf{g}_t \in \mathfrak{M}$. Thus, a practical implementation of our algorithm would require $\mathfrak{M}$ to be defined in such a way that the latter integral can be calculated either analytically or numerically. At the same time, $\mathfrak{M}$ should be rich enough to approximate functions from $\mathfrak{D}$ in terms of weak topology. Thus, to summarize, $\mathfrak{M}$ should be:

- dense in $\mathfrak{D}'$.
- $\mathcal{M}_{\mathbf{g}}$ is efficiently computable.

An example of $\mathfrak{M}$ that satisfies the latter 2 conditions will be given in the next section.

5.1 Return to initial coordinates

In applications, it is desirable that an algorithm for the problem deals with the function $\mathcal{F}^{-1}[\mathbf{g}] = [\mathcal{F}^{-1}[g_1], \mathcal{F}^{-1}[g_2]]$, rather than with $[g_1, g_2]$.

The pair $[\mathcal{F}^{-1}[g_1],\ \mathcal{F}^{-1}[g_2]]$ has the following interpretation. Since g_1 is the probability density function that approximates the empirical distribution μ_{emp}, then $G_1 = \mathcal{F}^{-1}[g_1]$ is the characteristic function of g_1 (up to a constant factor). By Bochner's theorem [5] we know that the function can serve as the characteristic function of a probability distribution if and only if it is continuous, positive definite and

$$G_1(\mathbf{0}) = 1.$$

The function $G_2 = \mathcal{F}^{-1}[g_2]$ is simply the regression function in the term $\rho J(\mathcal{F}^{-1}[g_2]) = \rho J(G_2)$, i.e., G_2 is a real-valued function.

Thus, in the dual reformulation we search over pairs $[G_1,\ G_2]$ where G_1 is complex-valued, continuous, positive definite and G_2 is real-valued.

The specifics of scheme 1 is that it allows such a reformulation.

Indeed, at step 4 of the algorithm we minimize the expression:

$$\Phi(\mathbf{g}, S) = (1-\rho)I(g_1, \mu_{\text{emp}}) + \rho J(\mathcal{F}^{-1}[g_2]) + \lambda \|S_{\mathbf{g}} - S_{t-1}\|^2_{L_2^{n\times 2}(\mathbb{R}^n)},$$

where $S_{t-1} = P_{\mathbf{g}_{t-1}} S_{\mathbf{g}_{t-1}}$ has been calculated on the previous iteration. According to Theorem 4.3 and formula (4.3) we can rewrite the penalty term $\lambda \|S_{\mathbf{g}} - S_{t-1}\|^2_{L_2^{n\times 2}(\mathbb{R}^n)}$ as:

$$\lambda \int_{\mathbb{R}^n \times \mathbb{R}^n} \gamma(\mathbf{x}-\mathbf{y}) E(\mathbf{g}(\mathbf{x})\mathbf{x}^T - \mathbf{g}_{t-1}(\mathbf{x})\mathbf{x}^T P^T_{\mathbf{g}_{t-1}})(\mathbf{y}\mathbf{g}(\mathbf{y})^T - P_{\mathbf{g}_{t-1}}\mathbf{y}\mathbf{g}_{t-1}(\mathbf{y})^T) E d\mathbf{x} d\mathbf{y}.$$

Let us denote $\mathbf{G} = \mathcal{F}^{-1}[\mathbf{g}]$ and $\mathbf{G}_{t-1} = \mathcal{F}^{-1}[\mathbf{g}_{t-1}]$. Using the well-known duality between x_i and $-\mathrm{i}\partial_{x_i}$, we see that

$$\mathcal{F}^{-1}[\mathbf{x}\mathbf{g}(\mathbf{x})^T] = -\mathrm{i}\partial_{\mathbf{x}} \mathbf{G}^T,$$
$$\mathcal{F}^{-1}[P_{\mathbf{g}_{t-1}}\mathbf{x}\mathbf{g}_{t-1}(\mathbf{x})^T] = -\mathrm{i} P_{\mathbf{g}_{t-1}} \partial_{\mathbf{x}} \mathbf{G}^T_{t-1}.$$

The unitarity of the inverse Fourier transform and the convolution theorem gives us that the penalty term equals

$$\lambda' \int_{\mathbb{R}^n} p(\mathbf{x}) \left\| E\frac{\partial \mathbf{G}}{\partial \mathbf{x}} - E\frac{\partial \mathbf{G}_{t-1}}{\partial \mathbf{x}} P_{t-1} \right\|^2_F d\mathbf{x}, \tag{5.3}$$

where $\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \left[\frac{\partial f_i}{\partial x_j}\right]_{1\le i\le 2, 1\le j\le n}$ denotes the Jacobian of $\mathbf{f}: \mathbb{R}^n \to \mathbb{R}^2$ and $p = \mathcal{F}^{-1}[\gamma]$, $P_{t-1} = P_{\mathbf{g}_{t-1}}$.

The set dual to $\mathfrak{M}$ is defined as $\mathfrak{M}' = \{\mathcal{F}^{-1}[\mathbf{h}] | \mathbf{h} \in \mathfrak{M}\}$. The matrix $\mathcal{M}_t = \mathcal{M}_{\mathbf{g}_t}$ can also be calculated using $\mathbf{G}_t$:

$$\mathcal{M}_t = \left[\int_{\mathbb{R}^n \times \mathbb{R}^n} \gamma(\mathbf{x}-\mathbf{y}) x_i \mathbf{g}_t^\dagger(\mathbf{x}) y_j E^2 \mathbf{g}_t(\mathbf{y}) d\mathbf{x} d\mathbf{y}\right]_{n\times n}$$
$$\propto \left[\int_{\mathbb{R}^n} p(\mathbf{x}) \frac{\partial E\mathbf{G}_t}{\partial x_i}^\dagger \frac{\partial E\mathbf{G}_t}{\partial x_j} d\mathbf{x}\right]_{n\times n} = \int_{\mathbb{R}^n} p(\mathbf{x}) \frac{\partial E\mathbf{G}_t}{\partial \mathbf{x}}^\dagger \frac{\partial E\mathbf{G}_t}{\partial \mathbf{x}} d\mathbf{x}.$$

For $\mathbf{G} = [G_1,\ G_2]^T$ we define:

$$\Phi'_t(\mathbf{G}) = (1-\rho)I(\mathcal{F}[G_1], \mu_{\text{emp}}) + \rho J(G_2) + \lambda' \int_{\mathbb{R}^n} p(\mathbf{x}) \left|\left| \frac{\partial E\mathbf{G}}{\partial \mathbf{x}} - \frac{\partial E\mathbf{G}_{t-1}}{\partial \mathbf{x}} P_{t-1} \right|\right|_F^2 d\mathbf{x}.$$

Note that we can assume that G_2 and $\mathbf{G}_{t2}$ are real-valued, because the objective $\Phi'_t(\mathbf{G})$ always attains its minimum on such a pair $[G_1^*\ G_2^*]$ that G_2^* is a real-valued function.

Thus, algorithm 2 is dual to algorithm 1.

Algorithm 2 Alternating scheme with initial coordinates

```
procedure
    P_0 ← 0
    for t = 1, ··· , N do
        G_t ← arg min_{G∈𝔐'} Φ'_{t−1}(G)
        M_t ← ∫_{ℝ^n} p(x) (∂EG_t/∂x)^† (∂EG_t/∂x) dx
        P_t ← projection to first k principal components of M_t
```

5.2 The description of $\mathfrak{M}'$

Let us now give an example of a set $\mathfrak{M}$ that satisfies both conditions that we imposed in Section 5. Instead of defining $\mathfrak{M}$ we will define its dual

$$\mathfrak{M}' = \left\{ \mathbf{H} = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \mid h_i \in FF_i \right\},$$

where FF_2 is a set of functions defined by the standard single layer neural network:

$$\sum_{i=1}^{M} w_i \psi(\mathbf{a}_i^T \mathbf{x} + b_i),$$

where $\mathbf{a}_i \in \mathbb{R}^n, w_i, b_i \in \mathbb{R}$ are parameters and M is a hyperparameter. For ψ we only assume that it is some non-constant function whose first derivatives are continuous and bounded. In practice we use the hyperbolic tangent function for ψ.

We define FF_1 as a set of functions given in the following parameterized form:

$$\sum_{i=1}^{M'} \alpha_i e^{\mathrm{i}\omega_i^T \mathbf{x}}, \tag{5.4}$$

where $\alpha_i > 0$ and $\sum_{i=1}^{M'} \alpha_i = 1$. It is easy to see that such functions are always positive definite. In practice we use the expression $\sum_{i=1}^{M'} \alpha_i \cos(\omega_i^T \mathbf{x})$, because the empirical distribution is made symmetrical before we apply the algorithm. The number of neurons in the single-layer neural network with the cosine activation function, M', is a hyperparameter.

Let us show that $\mathfrak{M}'$ is dense in $\mathcal{F}^{-1}[\mathfrak{D}']$. It is a well-known fact that for any compact set $\Omega \subseteq \mathbb{R}^n$, single-layer neural networks can approximate any function in $C(\mathbb{R}^n)$ with an arbitrary accuracy [4]. Thus, the weak closure of FF_2 contains $\mathcal{F}^{-1}[\mathfrak{C}(\mathbb{R}^n)]$ (all functions in the last class are real-valued), i.e. FF_2 covers all interesting functions that can serve as candidates for the regression function.

Using Theorem 2 from [4], it can be shown that the conical hull of $\{e^{\mathrm{i}\omega_i^T \mathbf{x}} | \omega \in \mathbb{R}^n\}$ is dense in the cone of continuous positive definite functions, and therefore, are dense with respect to weak topology in $\mathcal{F}^{-1}[\mathfrak{B}(\mathbb{R}^n)]$. Thus, by the expression (5.4) we are able to approximate all characteristic functions.

5.3 Maximum mean discrepancy metric

All our experiments were done for the following well-known distance function [16]:

$$I(\mu, \mu_{\text{emp}}) = \int_{\mathbb{R}^n} q(\mathbf{x})|\phi_\mu(\mathbf{x}) - \phi_{\text{emp}}(\mathbf{x})|^2 d\mathbf{x}, \tag{5.5}$$

where $q(\mathbf{x}) > 0$ is a continuous function such that $\int_{\mathbb{R}^n} q(\mathbf{x})d\mathbf{x} = 1$ and ϕ_μ, ϕ_{emp} are the characteristic functions of the distributions μ, μ_{emp} correspondingly. It is easy to see that

$$\phi_{\text{emp}}(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\xi}\sim\mu_{\text{emp}}} e^{\mathrm{i}\mathbf{x}^T\boldsymbol{\xi}} = \frac{1}{N}\sum_{i=1}^{N} e^{\mathrm{i}\mathbf{x}_i^T\mathbf{x}}.$$

The Maclaurin series of the characteristic function has the form:

$$\phi_\mu(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\xi}\sim\mu} e^{\mathrm{i}\mathbf{x}^T\boldsymbol{\xi}} = \sum_{j=0}^{\infty} \frac{\mathbb{E}_{\boldsymbol{\xi}\sim\mu}(\mathrm{i}\mathbf{x}^T\boldsymbol{\xi})^j}{j!}.$$

Thus, by approximating $\phi_{\text{emp}}(\mathbf{x})$ in a neighborhood of the origin, defined by $q(\mathbf{x})$, our method tries to approximate all moments of μ_{emp} simultaneously.

Our experiments show that algorithm 2 converges to a better solution if we set $q(\mathbf{x}) \propto p(\mathbf{x}) = \mathcal{F}^{-1}[\gamma]$.

5.4 Practical algorithm with initial coordinates

Let us set $p(\mathbf{x}) = \frac{e^{-\frac{\|\mathbf{x}\|^2}{2\delta^2}}}{\sqrt{2\pi\delta^2}^n}$. Step 4 of algorithm 2:

$$\theta_t = \arg\min_\theta \Phi'_{t-1}(\mathbf{G}_\theta)$$

is done via the gradient descent-type algorithm (we use the Adam optimizer [20], a popular tool in AI [31, 21, 30, 33, 37, 32, 41, 25]) that needs as an oracle an unbiased estimator of the gradient at a given point θ.

In practice it is natural to estimate the gradient of $\Phi'_{t-1}(\mathbf{G}_\theta)$ as follows:

$$\nabla_\theta\Bigg(\frac{1-\rho}{n_{\text{batch}}}\sum_{i=1}^{n_{\text{batch}}}\|\mathbf{G}_{\theta 1}(\mathbf{z}_i) - \phi_{\text{emp}}(\mathbf{z}_i)\|^2 + \frac{\rho}{n_{\text{batch}}}\sum_{i=1}^{n_{\text{batch}}}\|y_i - \mathbf{G}_{\theta 2}(\mathbf{x}_i + \boldsymbol{\epsilon}_i)\|^2 + \frac{\tilde{\lambda}}{n_{\text{batch}}}\sum_{i=1}^{n_{\text{batch}}}\|\frac{\partial E\mathbf{G}_\theta}{\partial\mathbf{x}}(\mathbf{z}'_i) - \frac{\partial E\mathbf{G}_{\theta_{t-1}}}{\partial\mathbf{x}}(\mathbf{z}'_i)P_{t-1}\|^2\Bigg),$$

where $\mathbf{x}_1, \cdots, \mathbf{x}_{n_{\text{batch}}} \sim^{iid} \mu_{\text{emp}}$ ($y_1, \cdots, y_{n_{\text{batch}}}$ are the corresponding outputs), $\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_{n_{\text{batch}}} \sim^{iid} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}}$ and $\mathbf{z}_1, \cdots, \mathbf{z}_{n_{\text{batch}}}, \mathbf{z}'_1, \cdots, \mathbf{z}'_{n_{\text{batch}}} \sim^{iid} p\ (= q)$.

Since

$$\mathcal{M}_t = \int_{\mathbb{R}^n} p(\mathbf{x})\frac{\partial E\mathbf{G}_{\theta_t}}{\partial\mathbf{x}}^\dagger\frac{\partial E\mathbf{G}_{\theta_t}}{\partial\mathbf{x}}d\mathbf{x},$$

it is natural to estimate it as:

$$\hat{\mathcal{M}}_t = \frac{1}{N_{\text{batch}}}\sum_{i=1}^{N_{\text{batch}}}\frac{\partial E\mathbf{G}_{\theta_t}}{\partial\mathbf{x}}^\dagger(\mathbf{t}_i)\frac{\partial E\mathbf{G}_{\theta_t}}{\partial\mathbf{x}}(\mathbf{t}_i),$$

where $\mathbf{t}_1, \cdots, \mathbf{t}_{N_{\text{batch}}}$ are independent identically distributed with distribution p. Thus, the pseudocode of the algorithm can be found below.

Algorithm 3 Practical algorithm with initial coordinates

$P_0 \longleftarrow \mathbf{0}, \theta_0 \longleftarrow \mathbf{0}$
for $t = 1, \cdots, T$ **do**
  while θ has not converged **do**
    Sample $\mathbf{x}_1, \cdots, \mathbf{x}_{n_{\text{batch}}} \sim^{iid} \mu_{\text{emp}}$ with the corresponding outputs $y_1, \cdots, y_{n_{\text{batch}}}$
    Sample $\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_{n_{\text{batch}}} \sim^{iid} \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}}$
    Sample $\mathbf{z}_1, \cdots, \mathbf{z}_{n_{\text{batch}}}, \mathbf{z}'_1, \cdots, \mathbf{z}'_{n_{\text{batch}}} \sim^{iid} p$
    $L \longleftarrow \left(\frac{1-\rho}{n_{\text{batch}}} \sum_{i=1}^{n_{\text{batch}}} \|\mathbf{G}_{\theta 1}(\mathbf{z}_i) - \phi_{\text{emp}}(\mathbf{z}_i)\|^2 + \frac{\rho}{n_{\text{batch}}} \sum_{i=1}^{n_{\text{batch}}} \|y_i - \mathbf{G}_{\theta 2}(\mathbf{x}_i + \boldsymbol{\epsilon}_i)\|^2 + \frac{\tilde{\lambda}}{n_{\text{batch}}} \sum_{i=1}^{n_{\text{batch}}} \|\frac{\partial E \mathbf{G}_\theta}{\partial \mathbf{x}}(\mathbf{z}'_i) - \frac{\partial E \mathbf{G}_{\theta_{t-1}}}{\partial \mathbf{x}}(\mathbf{z}'_i) P_{t-1}\|^2 \right)$
    $\theta \longleftarrow \text{Adam}(\nabla_\theta L, \theta, \alpha, \beta_1, \beta_2)$
  $\theta_t \longleftarrow \theta$
  Sample $\mathbf{t}_1, \cdots, \mathbf{t}_{N_{\text{batch}}} \sim^{iid} p$
  $\hat{\mathcal{M}}_t \longleftarrow \frac{1}{N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} \frac{\partial E \mathbf{G}_{\theta_t}}{\partial \mathbf{x}}(\mathbf{t}_i)^\dagger \frac{\partial E \mathbf{G}_{\theta_t}}{\partial \mathbf{x}}(\mathbf{t}_i)$
  Find $\{\mathbf{v}_i\}_1^n$ s.t. $\hat{\mathcal{M}}_t \mathbf{v}_i = \lambda_i \mathbf{v}_i$, $\lambda_1 \geq \cdots \geq \lambda_n$
  $P_t \longleftarrow \sum_{i=1}^k \mathbf{v}_i \mathbf{v}_i^T$

Output: $\mathbf{v}_1, \cdots, \mathbf{v}_k$

5.5 Experiments

We made experiments on the standard datasets: Heart, Breast Cancer, Diabetes, Boston house prices and Wine quality. First, we applied Principal Component Analysis (PCA) and Sliced Inverse Regression (SIR) algorithms to the training set and calculated the effective subspace for $k = 2$. All points were projected to that space and we obtained two-dimensional representations of input points. In the last step we applied the 10 nearest neighbors algorithm (KNN) to predict outputs of the test set (for the regression case, the KNN regression was used). The same scheme was repeated with the Kernel Dimensionality Reduction (KDR) algorithm [14] and the alternating scheme with $\rho = 1, \frac{1}{2}, 0$. We verified that $\rho = 1$ corresponds to the pure sufficient dimension reduction case, because the best prediction was achieved for this value.

We experimented with algorithm 3 setting the key parameters as[1] (the data was standardized):

$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^n} e^{-\frac{\|\mathbf{x}\|^2}{2}},$$
$$T = 50, \tilde{\lambda} = 10.0, \sigma = 0.8,$$
$$\alpha = 10^{-4}, \beta_1 = 0.5, \beta_2 = 0.9.$$

Hyperparameters θ of the neural network model $\mathbf{G}_\theta$ were set as: $M' = N$ and M equals either 50 or 100. The parameters of the characteristic function were initialized as $\alpha_i^0 = \frac{1}{N}$, $\omega_i^0 = \mathbf{x}_i$ so that $\sum_{i=1}^{M'} \alpha_i^0 \cos(\omega_i^0 \cdot \mathbf{x}) = \phi_{\text{emp}}(\mathbf{x})$. Depending on the dataset, $n_{\text{batch}} \approx \frac{N}{10}$ and $N_{\text{batch}} \approx 10000$. In table 1 one can see the obtained test set accuracy on the classification tasks and R^2 on the regression tasks.

From Table 1 we see that the alternating scheme for $\rho = 1$ always outperforms SIR. As expected, the worst prediction accuracy is shown for $\rho = 0$, when the sufficiency term $J(f)$ is absent. Surprisingly, results for $\rho = \frac{1}{2}$ are also comparable with SIR's, which indicates that the two basic requirements, the proximity to the empirical distribution and the sufficiency to predict an output are often not mutually exclusive. The conclusion holds only if the distance to the empirical distribution

[1] Since the role of the parameter σ is similar to that of the bandwidth in the kernel density estimation, we use Silverman's rule of thumb to set σ.

Dataset	PCA %	SIR	KDR	$\rho = 1$	$\rho = \frac{1}{2}$	$\rho = 0$
Heart (acc)	79.80	81.1	84.5	83.2	80.4	81.8
Breast (acc)	93.46	96.5	95.1	97.1	95.6	94.0
Diabetes (R^2)	25.34	43.8	38.4	44.2	39.9	23.4
Boston (R^2)	56.42	76.2	70.4	76.7	74.2	51.3
Wine (R^2)	93.91	81.7	89.9	95.2	92.7	94.1

Table 1: The cross-validated accuracies and R^2 of KNN on 2-dimensional input representations.

is the maximum mean discrepancy distance. Figure 1 shows how the 2D scatter plots of points on the effective subspace look for the values of $\rho = 1, \frac{1}{2}, 0$.

Codes that simplify the reproducibility of our results can be downloaded from GitHub repository $https://github.com/k-nic/LR\_SDR$.

6 Conclusions

Unsupervised dimension reduction and sufficient dimension reduction tasks are usually treated as optimization tasks with substantially different objectives. An approach suggested in the paper deals with the hybrid setting, i.e. the objective that we study is the sum of the term that measures the consistency with an unsupervised part of data and the term that measures correlations between a projected input and an output. We develop a new approach to the minimization of such objectives that we call the alternating scheme. The results demonstrate that it is often possible to find a low-dimensional subspace that approximates well the support of unsupervised data, and at the same time can serve as an efficient subspace for the regression.

Acknowledgments

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP27510283), PI R. Takhanov.

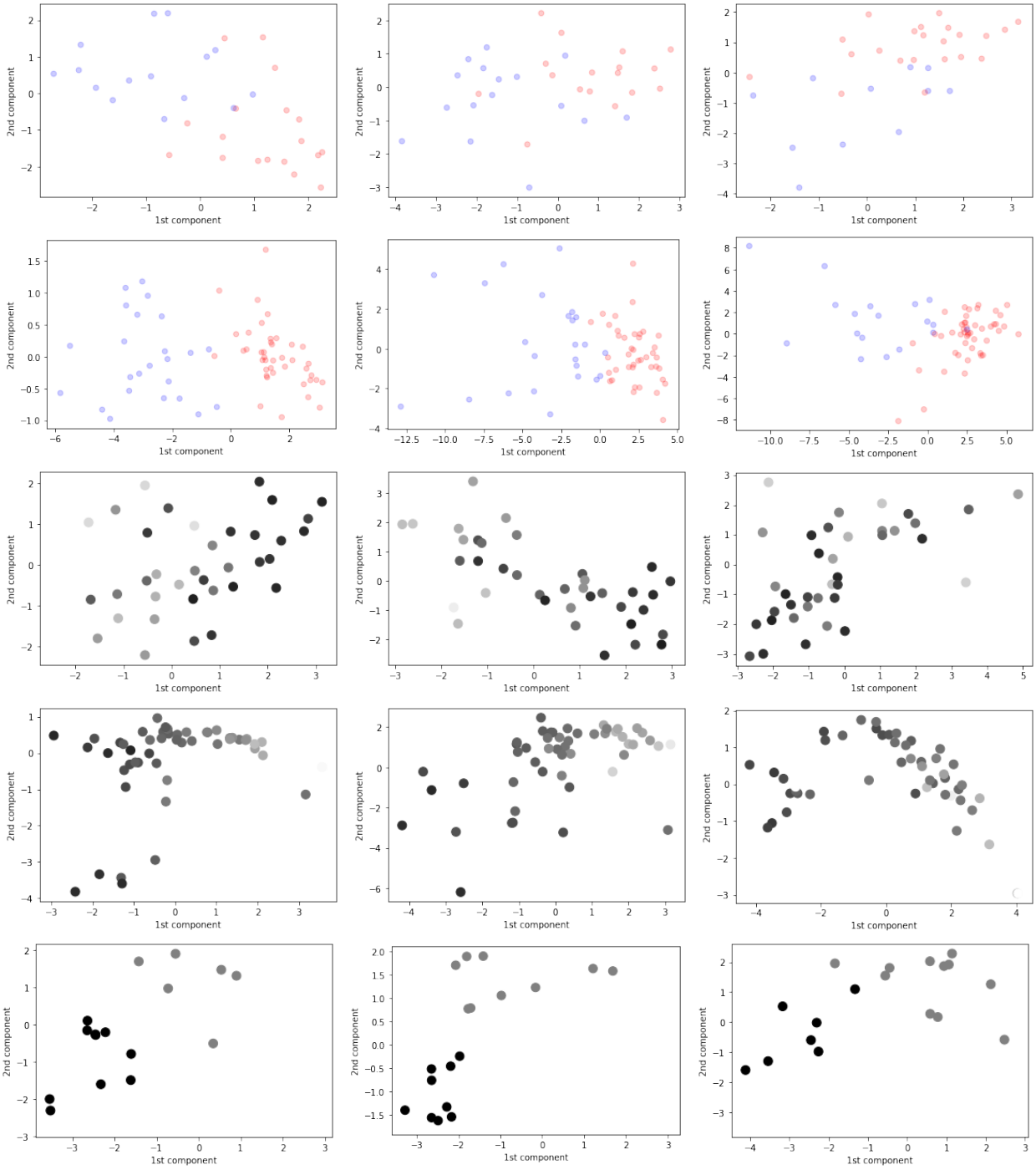

Figure 1: 2d-scatter plots of test sets for Heart, Breast Cancer, Diabetes, Boston and Wine datasets (rows) and for $\rho = 1, \frac{1}{2}, 0$ (columns). For regression tasks, the blackness of a point is proportional to a target variable's value.

References

[1] Z. Assylbekov, R. Takhanov, *Reusing weights in subword-aware neural language models*. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2018), 1413–1423.

[2] Z. Assylbekov, R. Takhanov, *Context vectors are reflections of word vectors in half the dimensions (extended abstract)*. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (2021).

[3] M.T. Augustine, P. Patil, M. Bhushan, S. Bhartiya, *Autoencoder with ordered variance for nonlinear model identification*. Arxiv (2024), arXiv:2402.14031.

[4] A.R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*. IEEE Transactions on Information Theory 39 (1993), no. 3, 930–945.

[5] S. Bochner, *Vorlesungen über Fouriersche Integrale*. Akad. Verl.-Ges., 1932.

[6] R.D. Cook, *SAVE: a method for dimension reduction and graphics in regression*. Communications in Statistics - Theory and Methods 29 (2000), no. 9-10, 2109–2121.

[7] R.D. Cook, L. Forzani, *Principal fitted components for dimension reduction in regression*. Statistical Science 23 (2008), no. 4, 485–501.

[8] R.D. Cook, L. Forzani, *Likelihood-based sufficient dimension reduction*. Journal of the American Statistical Association 104 (2009), no. 485, 197–208.

[9] R.D. Cook, S. Weisberg, *Sliced inverse regression for dimension reduction: comment*. Journal of the American Statistical Association 86 (1991), no. 414, 328–332.

[10] T.F. Cox, M. Cox, *Multidimensional scaling*. Chapman and Hall/CRC, 2000.

[11] J.P. Cunningham, Z. Ghahramani, *Linear dimensionality reduction: survey, insights, and generalizations*. Journal of Machine Learning Research 16 (2015), 2859–2900.

[12] R.A. Fisher, *The use of multiple measurements in taxonomic problems*. Annals of Eugenics 7 (1936), no. 2, 179–188.

[13] K. Pearson, *On lines and planes of closest fit to systems of points in space*. Philosophical Magazine 2 (1901), no. 11, 559–572.

[14] K. Fukumizu, F.R. Bach, M.I. Jordan, *Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces*. Journal of Machine Learning Research 5 (2004), 73–99.

[15] B. Ghojogh, A. Ghodsi, F. Karray, M. Crowley, *Sufficient dimension reduction for high-dimensional regression and low-dimensional embedding: tutorial and survey*. Arxiv (2021), arXiv:2110.09620.

[16] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, A. Smola, *A kernel two-sample test*. Journal of Machine Learning Research 13 (2012), 723–773.

[17] T. Hastie, R. Tibshirani, *Discriminant analysis by Gaussian mixtures*. Journal of the Royal Statistical Society Series B 58 (1996), no. 1, 155–176.

[18] H. Hotelling, *Relations between two sets of variates*. Biometrika 28 (1936), no. 3-4, 321–377.

[19] T. Hsing, R. Eubank, *Theoretical foundations of functional data analysis*. Wiley, 2015.

[20] D.P. Kingma, J. Ba, *Adam: a method for stochastic optimization*. Proceedings of the International Conference on Learning Representations (2015).

[21] V. Kolmogorov, M. Rolinek, R. Takhanov, *Effectiveness of structural restrictions for hybrid CSPs*. Proceedings of the 26th International Symposium on Algorithms and Computation (2015), 566–577.

[22] K.-C. Li, *Sliced inverse regression for dimension reduction*. Journal of the American Statistical Association 86 (1991), no. 414, 316–327.

[23] K.-C. Li, *On principal Hessian directions for data visualization and dimension reduction.* Journal of the American Statistical Association 87 (1992), no. 420, 1025–1039.

[24] K. Mynbaev, Zh. Assylbekov, *Convergence of the partition function in the static word embedding model.* Eurasian Mathematical Journal 13 (2020), no. 4, 70–81.

[25] K. Mynbaev, *Asymptotic distribution of the OLS estimator for a mixed spatial model.* Journal of Multivariate Analysis 101 (2010), no. 3, 733–748.

[26] S. Mukherjee, D.-X. Zhou, *Learning coordinate covariances via gradients.* Journal of Machine Learning Research 7 (2006), 519–549.

[27] B. Schölkopf, A. Smola, K.-R. Müller, *Nonlinear component analysis as a kernel eigenvalue problem.* Neural Computation 10 (1998), no. 5, 1299–1319.

[28] M. Sugiyama, *Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis.* Journal of Machine Learning Research 8 (2007), 1027–1061.

[29] P. Switzer, A. Green, *Min/max autocorrelation factors for multivariate spatial imagery.* Stanford University, 1984.

[30] R. Takhanov, *Extensions of the minimum cost homomorphism problem.* Proceedings of the International Computing and Combinatorics Conference (2010), 328–337.

[31] R. Takhanov, *Hybrid VCSPs with crisp and valued conservative templates.* Proceedings of the 28th International Symposium on Algorithms and Computation (2017).

[32] R. Takhanov, *The algebraic structure of the densification and the sparsification tasks for CSPs.* Constraints 28 (2023), no. 1, 13–44.

[33] R. Takhanov, *Computing a partition function of a generalized pattern-based energy over a semiring.* Theory of Computing Systems 67 (2023), no. 4, 760–784.

[34] R. Takhanov, *On the speed of uniform convergence in Mercer's theorem.* Journal of Mathematical Analysis and Applications 518 (2023), 126718.

[35] R. Takhanov, *Reducing the dimensionality of data using tempered distributions.* Digital Signal Processing 133 (2023), 103819.

[36] R. Takhanov, *Multi-layer random features and the approximation power of neural networks.* Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence (2024).

[37] R. Takhanov, *On the induced problem for fixed-template CSPs.* Proceedings of the 49th International Conference on Current Trends in Theory and Practice of Computer Science (2024), 485–499.

[38] R. Takhanov, Y.S. Abylkairov, M. Tezekbayev, *Autoencoders for a manifold learning problem with a Jacobian rank constraint.* Pattern Recognition 143 (2023), 109777.

[39] R. Takhanov, Z. Assylbekov, *Patterns versus characters in subword-aware neural language modeling.* Proceedings of the International Conference on Neural Information Processing (2017), 157–166.

[40] R. Takhanov, V. Kolmogorov, *Inference algorithms for pattern-based CRFs on sequence data.* Proceedings of the International Conference on Machine Learning (2013), 145–153.

[41] R. Takhanov, V. Kolmogorov, *Combining pattern-based CRFs and weighted context-free grammars.* Intelligent Data Analysis 26 (2022), 257–272.

[42] M. Wang, F. Sha, M.I. Jordan, *Unsupervised kernel dimension reduction.* Advances in Neural Information Processing Systems 23, 2010, 2379–2387.

[43] L. Wiskott, T.J. Sejnowski, *Slow feature analysis: unsupervised learning of invariances.* Neural Computation 14 (2002), no. 4, 715–770.

[44] Y. Xia, H. Tong, W.K. Li, L.-X. Zhu, *An adaptive estimation of dimension reduction space.* Journal of the Royal Statistical Society Series B 64 (2002), no. 3, 363–410.

Zhalgas Mukanov
Department of Fundamental Mathematics
L.N. Gumilyov Eurasian National University
13 Munaitpasov St
010008 Astana, Republic of Kazakhstan
E-mails: mukanovj@mail.ru

Anuar Sharafudinov
AILabs Technologies
159 West Broadway Ste 200
Salt Lake City, UTAH, USA
E-mail: AnuarSh@ailabs.kz

Rustem Takhanov, Arman Bekembayev
Mathematics Department
Nazarbayev University
53 Kabanbay Batyr Ave
010000 Astana, Republic of Kazakhstan
E-mails: rustem.takhanov@nu.edu.kz, arman.bekembayev@nu.edu.kz

Received: 02.08.2024