DIMENSION REDUCTION AS AN OPTIMIZATION PROBLEM OVER A SET OF GENERALIZED FUNCTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We reformulate unsupervised dimension reduction problem (UDR) in the language of tempered distributions, i.e. as a problem of approximating an empirical probability density function \( p_{\text{emp}}(x) \) by another tempered distribution \( q(x) \) whose support is in a \( k \)-dimensional subspace. Thus, our problem is reduced to the minimization of the distance between \( q \) and \( p_{\text{emp}} \), \( D(q, p_{\text{emp}}) \), over a pertinent set of generalized functions.

This infinite-dimensional formulation allows to establish a connection with another classical problem of data science — the sufficient dimension reduction problem (SDR). Thus, an algorithm for the first problem induces an algorithm for the second and vice versa. In order to reduce an optimization problem over distributions to an optimization problem over ordinary functions we introduce a nonnegative penalty function \( R(f) \) that “forces” the support of \( f \) to be \( k \)-dimensional. Then we present an algorithm for minimization of \( I(f) + \lambda R(f) \), based on the idea of two-step iterative computation, briefly described as a) an adaptation to real data and to fake data sampled around a \( k \)-dimensional subspace found at a previous iteration, b) calculation of a new \( k \)-dimensional subspace. We demonstrate the method on 4 examples (3 UDR and 1 SDR) using synthetic data and standard datasets.

1 INTRODUCTION

Linear dimension reduction (LDR) is a family of problems in data science that includes principal component analysis, factor analysis, linear multidimensional scaling, Fisher’s linear discriminant analysis, canonical correlations analysis, sufficient dimensionality reduction (SDR), maximum autocorrelation factors, slow feature analysis and more. In unsupervised dimension reduction (UDR) we are given a finite number of points in \( \mathbb{R}^n \) (sampled according to some unknown distribution) and the goal is to find a “low-dimensional” affine (or linear) subspace that approximates “the support” of the distribution. The study field currently achieved a saturation level at which unifying frameworks for the problem become of special interest [Cunningham & Ghahramani (2015)]. An approach that we present in that paper is based on the theory of generalized functions, or tempered distributions [Soboleff (1936); Schwartz (1949)]. An important generalized function that cannot be represented as an ordinary function is the Dirac delta function, denoted \( \delta \), and \( \delta^n \) denotes its \( n \)-dimensional version.

Any dataset \( \{x_i\}_{i=1}^N \subseteq \mathbb{R}^n \) naturally corresponds to the distribution \( p_{\text{emp}}(x) = \frac{1}{N} \sum_{i=1}^N \delta^n(x - x_i) \) which, with some abuse of terminology, can be called the empirical probability density function. Based on that, UDR can be understood as a task whose goal is to approximate \( p_{\text{emp}}(x) \) by \( q(x) \), where \( q(x) \) is a distribution whose density is supported in a \( k \)-dimensional affine subspace \( A \subseteq \mathbb{R}^n \). Note that a function whose density is supported in some low-dimensional subset of \( \mathbb{R}^n \) is not an ordinary function. Exact definitions of such distributions can be found in Section 3. To formulate an optimization task we additionally need a loss \( D(p_{\text{emp}}, q) \) that measures the distance between the ground truth \( p_{\text{emp}} \) and a distribution \( q \), that we search for. Thus, in our approach, the UDR problem is defined as:

\[
I(q) = D(p_{\text{emp}}, q) \rightarrow \min_q
\]

under the condition that \( q(x) \) has a \( k \)-dimensional support.
The SDR problem is tightly connected with the UDR problem. In SDR, given supervised data, the goal is to find the so-called effective subspace, defined by its basis vectors \( \{ w_1, \ldots, w_k \} \subset \mathbb{R}^n \), such that the regression function can be searched in the form \( g(w_1^T x, \ldots, w_k^T x) \). In \cite{Wang2010} it was shown how a method originally developed for SDR can be turned into an UDR method, i.e. applied to unsupervised data, by simply setting an output to be equal to an input. The key observation of our analysis, stated in Theorem 2, is that a class of functions of the form \( g(w_1^T x, \ldots, w_k^T x) \) can be characterized as functions whose Fourier transform is supported in the corresponding effective subspace. In Section 4 we give 3 examples of UDR problems that we cast as and in the fourth example we formulate SDR as an optimization task with the search space dual to that of UDR. Thus, all 4 examples can be studied within the same optimization framework.

The structure of the paper is as follows: in Section 3 we formally define the search space in Problem 1 denoted \( \mathcal{G}_k \), and an image of \( \mathcal{G}_k \) under the Fourier transform, denoted \( \mathcal{F} \). Instead of searching directly in a set of generalized functions, \( \mathcal{G}_k \), in Section 5 we describe how we substitute an ordinary function for a distribution in the optimization task at the expense of adding a new penalty term to its objective, \( \lambda R(f) \). Using a gaussian kernel \( M(x, y) \), Theorem 4 characterizes generalized \( g \in \mathcal{G}_k \) as such \( g \) for which the matrix of properly defined integrals \( M_g = \text{Re} \int \int_{\mathbb{R}^n \times \mathbb{R}^n} x_i y_j g(x)^T M(x, y) g(y) dxdy \) is of rank \( k \). We define \( R(f) \) as a squared Frobenius distance from \( \sqrt{M_f} \) to the closest matrix of rank \( k \). In Section 6 we suggest a method for solving \( \min_{\mathcal{U}} I(\phi) + \lambda R(\phi) \) which we call the alternating scheme. Section 7 is dedicated to experiments with the alternating scheme on synthetic data and standard datasets.

## 2 Preliminaries and Notations

Throughout this paper we use standard terminology and notation from functional analysis. For exact definitions one can address the textbook on the theory of distributions \cite{Friedlander1998}. The Schwartz space of functions and its dual space are denoted by \( S(\mathbb{R}^n) \) and \( S'(\mathbb{R}^n) \) correspondingly. For a tempered distribution \( T \in S'(\mathbb{R}^n) \) and \( \phi \in S(\mathbb{R}^n) \), \( \langle T, \phi \rangle \) denotes \( T(\phi) \).

The Fourier and inverse Fourier transforms are denoted by \( \mathcal{F}, \mathcal{F}^{-1} : S'(\mathbb{R}^n) \to S'(\mathbb{R}^n) \). For brevity, we denote \( \mathcal{F}[f] \) by \( \hat{f} \). If all required conditions are satisfied, an integrable \( f : \mathbb{R}^n \to \mathbb{C} \) (or, a Borel measure \( \mu \) on \( \mathbb{R}^n \)) is used as the tempered distribution \( T_f \) (or, \( T_\mu \)) where \( \langle T_f, \phi \rangle = \int_{\mathbb{R}^n} f(x) \phi(x) dx \) (or, \( \langle T_\mu, \phi \rangle = \int_{\mathbb{R}^n} \phi(x) d\mu \)). For \( \Omega \subset S'(\mathbb{R}^n) \), \( \overline{\Omega} \) denotes the sequential closure of \( \Omega \) with respect to weak topology of \( S'(\mathbb{R}^n) \). By \( L_2(\mathbb{R}^n) \) we denote the \( L_2 \)-space with the inner product: \( \langle u, v \rangle_{L_2} = \int u(x)^* v(x) dx \). For \( \phi \in S(\mathbb{R}^n), \psi \in S'(\mathbb{R}^n) \), their convolution and multiplication are denoted by \( \phi \ast \psi \) and \( \phi \psi \) correspondingly. For \( g_1 \in S'(\mathbb{R}^k) \) and \( g_2 \in S'(\mathbb{R}^{n-k}) \), \( g_1 \otimes g_2 \in S'(\mathbb{R}^n) \) denotes their tensor product. For a square matrix \( A \), \( \text{Tr}(A) \) denotes its trace and for arbitrary matrix, \( \| A \|_F \overset{\text{def}}{=} \sqrt{\text{Tr}(A^T A)} \). Identity matrix of size \( n \) is denoted by \( I_n \).

## 3 Basic Function Classes

An example of a generalized function, whose density is concentrated in a \( k \)-dimensional subspace, is any distribution that can be represented as \( g \otimes \delta^{n-k} \overset{\text{def}}{=} g \otimes \delta \otimes \cdots \otimes \delta \) where \( g \in S'(\mathbb{R}^k) \).

If \( g = T_f \), where \( f : \mathbb{R}^k \to \mathbb{R} \) is an ordinary function, then \( g \otimes \delta^{n-k} \) can be understood as a generalized function whose density is concentrated in a subspace \( \{ x \in \mathbb{R}^n | x_i = 0, i > k \} \) and equals \( f(x_{1:k}) \). It can be shown that the distribution acts on \( \phi \in S(\mathbb{R}^n) \) in the following way:

\[
\langle T_f \otimes \delta^{n-k}, \phi \rangle = \int_{\mathbb{R}^k} f(x_{1:k}) \phi(x_{1:k}, 0_{n-k}) dx_{1:k}
\]

Now to generalize the latter definition to any \( k \)-dimensional subspace we have to introduce a change of variables in tempered distributions.

Let \( g \in S'(\mathbb{R}^n) \) and \( U \in \mathbb{R}^{n \times n} \) be an orthogonal matrix, i.e. \( U^T U = I_n \). Then, \( g_U \in S'(\mathbb{R}^n) \) is defined by the rule: \( \langle g_U, \phi \rangle = \langle g, \psi \rangle \) where \( \psi(x) = \phi(U^T x) \). If \( g = T_f \), the latter definition gives \( g_U = T_{fU} \) where \( f^U(x) = f(Ux) \). Now, we define classes of tempered distributions:

\[
\mathcal{G}_k = \{ (f \otimes \delta^{n-k})_U | f \in S'(\mathbb{R}^k), U \in \mathbb{R}^{n \times n}, U^T U = I_n \} \]
\[
G_k = \{(T_y \otimes \delta^{n-k})_U | f \in S(\mathbb{R}^k), U \in \mathbb{R}^{n \times n}, U^TU = I_n \}
\]
(3)
\[
F_k = \{T_x | r(x) = f(Ux), f \in S(\mathbb{R}^k), U \in \mathbb{R}^{k \times n}, \text{rank}(U) = k \}
\]
(4)
The first two classes are related as:

Theorem 1. \( G'_k = \overline{G}_k \).

The last two classes are isomorphic under the Fourier transform.

Theorem 2. \( F[G_k] = F_k \) and \( F^{-1}[F_k] = G_k \).

For any collection \( f_1, \ldots, f_s \in S'(\mathbb{R}^n) \), \( \text{span}_\mathbb{R}\{f_i\}_i^s \) denotes \( \{\sum_{i=1}^s \lambda_i f_i | \lambda_i \in \mathbb{R} \} \subseteq S'(\mathbb{R}^n) \), which is a linear space over \( \mathbb{R} \). The set \( G'_k \) has the following simple characterization:

Theorem 3. For any \( T \in S'(\mathbb{R}^n) \), \( T \in G'_k \) if and only if \( \dim \text{span}_\mathbb{R}\{x_1 T, x_2 T, \ldots, x_n T \} \leq k \).

Informally, the theorem holds because any linear dependency \( \alpha_1 x_1 T + \cdots + \alpha_n x_n T = 0 \) over \( \mathbb{R} \) implies that if \( \alpha_1 x_1 + \cdots + \alpha_n x_n \neq 0 \), then \( T = 0 \). This is equivalent to a statement that the support of \( T \) is concentrated on a subspace \( \alpha_1 x_1 + \cdots + \alpha_n x_n = 0 \). If \( \dim \text{span}_\mathbb{R}\{x_1 T, x_2 T, \ldots, x_n T \} \leq k \), then one can find \( n-k \) such dependencies, which means that the support of \( T \) is \( k \)-dimensional.

Let \( B(\mathbb{R}^n) \) denote the Borel sigma-algebra on \( \mathbb{R}^n \) and \( \mathcal{P} \) denote a set of all Borel probability measures on \( \mathbb{R}^n \). Let us now define

\[
P_k = \{ \mu \in \mathcal{P} | \exists v_1, \ldots, v_k \in \mathbb{R}^n, \forall A \in B(\mathbb{R}^n) : \mu(A) = \mu(A \cap \text{span}(v_1, \ldots, v_k)) \}
\]
(5)
i.e. \( P_k \) is a set of probability measures with all probability concentrated in some subspace \( \text{span}(v_1, \ldots, v_k) \) whose dimension is not greater than \( k \). It is easy to see that \( T_\mu \in G'_k \) for any \( \mu \in P_k \).

4 Examples of LDR formulations

UDR: Maximum mean discrepancy (MMD) Let \( k(x) = \frac{1}{(2\pi)^{n/2}} \exp(-\frac{|x|^2}{2}) \) be the radial gaussian kernel on \( \mathbb{R}^n \). The kernel \( k(x) \) defines the so-called kernel embedding of probability measures \( \phi \) [Muandet et al. (2017)]:

\[
\mu \in \mathcal{P} \rightarrow \phi(\mu) = k * \mu = \mathbb{E}_{y \sim \mu} k(x - y) = \int k(x - y) d\mu(y)
\]
The Maximum Mean Discrepancy (MMD) distance [Gretton et al. (2012)] is defined as the distance induced by metrics on \( L_2(\mathbb{R}^n) \), i.e. for two probability measures \( \mu, \nu \in \mathcal{P} \):

\[
d_{\text{MMD}}(\mu, \nu) = ||\phi(\mu) - \phi(\nu)||_{L_2(\mathbb{R}^n)}
\]

Let \( x_1, \ldots, x_N \in \mathbb{R}^n \) be the dataset of points. This dataset defines the empirical probabilistic measure \( \mu_{\text{data}} \) that corresponds to the tempered distribution \( T_{\mu_{\text{data}}} = \frac{1}{N} \sum_{i=1}^N \delta^n(x - x_i) \). We shall study a method concurrent to PCA that is based on solving the following problem:

\[
I(\nu) = d_{\text{MMD}}(\mu_{\text{data}}, \nu) = ||\phi(\mu_{\text{data}}) - \phi(\nu)||_{L_2(\mathbb{R}^n)} \rightarrow \min_{\nu \in P_k}
\]
(6)
i.e. we shall attempt to approximate the empirical probabilistic measure \( \mu_{\text{data}} \) with another probabilistic measure \( \nu \) which is supported in some \( k \)-dimensional subspace of \( \mathbb{R}^n \).

UDR: Distance based on higher moments (HM) It is well-known that maximum mean discrepancy measures the similarity between characteristic functions of two probability distributions in the \( O(\frac{1}{n}) \)-neighbourhood of the origin. Another approach to measure the similarity of two distributions is based on the difference between moments:

\[
d_{\text{HM}}(\mu, \nu)^2 = \sum_{s=1}^k \frac{\lambda_s}{n^s} \sum_{1 \leq i_1, \ldots, i_s \leq n} (m_{i_1 \ldots i_s} - n_{i_1 \ldots i_s})^2
\]
where \( m_{i_1 \ldots i_s} = \mathbb{E}_{X \sim \mu} [X[i_1] \cdots X[i_s]] \) and \( n_{i_1 \ldots i_s} = \mathbb{E}_{X \sim \nu} [X[i_1] \cdots X[i_s]] \) are corresponding moments. The positive parameters \( \lambda_1, \lambda_2, \lambda_3, \lambda_4 \) are chosen to fix the relative importance of the mean, the co-variance, the co-skewness and the co-kurtosis.
Thus, we will be interested in the following optimization task (analogous to (6)):
\[ d_{HM}(\mu_{\text{data}}, \nu) \rightarrow \min_{\nu \in \mathcal{P}_n} \]  

(7)

**UDR: Wasserstein distance (WD)**

Another important distance between probability measures that has the origins in the transport theory is the Wasserstein distance [Villani (2008)].

Let \((\mathbb{R}^n, || \cdot ||)\) be a Banach space. Between any two Borel probability measures \(\mu, \nu\) on \(\mathbb{R}^n\) with \(\int ||x||d\mu < \infty\) and \(\int ||x||d\nu < \infty\) the Wasserstein distance is:

\[ W(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int ||x - y||d\pi \]

where \(\Pi(\mu, \nu)\) is a set of all couplings of \(\mu\) and \(\nu\). The Wasserstein distance defines another version of LDR problem:

\[ W(\mu_{\text{data}}, \nu) \rightarrow \min_{\nu \in \mathcal{P}_n} \]  

(8)

In the appendix [B] one can find proofs that in the case of \(L_1\) norm \(||x|| = \sum_i |x_i|\), the task 8 corresponds to the well-studied robust PCA problem [Candes et al. (2011)]. If, instead of the \(L_1\)-norm, we use the \(L_2\)-norm, this leads to another well-studied task, which is known as the outlier pursuit problem [Xu et al. (2010)].

**Sufficient dimension reduction (SDR)**

Given a labeled dataset \(\{(x_i, y_i)\}_{i=1}^N\) where \(x_i \in \mathbb{R}^n, y_i \in \mathcal{C}\) (\(\mathcal{C}\) is a finite set of classes for a classification, or \(\mathbb{R}\) for a regression problem), the sufficient dimension reduction problem can be informally described as a problem of finding vectors \(w_1, \cdots, w_k \in \mathbb{R}^n\) such that \(p(y|x_1, \cdots, w_k^T x) \approx p(y|x)\) (possibly, under some additional assumptions on the form of \(p(y|x)\)).

We formulate the SDR problem as an optimization task:

\[ \inf_{f \in \mathcal{F}_k} J(f) \]  

(9)

The object \(f : \mathbb{R}^n \rightarrow \mathbb{R}\) is a smooth real-valued function. 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 and for the binary classification case with 0-1 outputs we use the following cost functions correspondingly:

\[ J(f) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\epsilon \sim N(0, \sigma^2I_n)} |y_i - f(x_i + \epsilon)|^2 \]

\[ J(f) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\epsilon \sim N(0, \sigma^2I_n)} H \left( y_i, \frac{e^{f(x_i + \epsilon)}}{1 + e^{f(x_i + \epsilon)}} \right) \]

where \(H(y, p) = -y \log p - (1 - y) \log (1 - p)\) and \(\sigma > 0\) is a parameter.

By requiring \(f \in \mathcal{F}_k\), we assume that the regression function \(f\) satisfies (for \(k\) fixed in advance):

\[ f(x) = g(w_1^T x, \cdots, w_k^T x) \]

where \(w_1, \cdots, w_k \in \mathbb{R}^n\). Thus, given an input \(x\), an output of \(f\) depends on the projection of \(x\) onto \(\text{span}(w_1, \cdots, w_k)\). The set \(\text{span}(w_1, \cdots, w_k)\) is called the effective subspace. Note that the way we defined the SDR’s objective \(J(f)\) for the regression and the classification cases is not unique. There are definitions that has the same form [9] but deal with the conditional distribution \(p(y|x)\) as an argument, instead of the regression function.

5 Reduction of the Optimization Problem to Ordinary Functions

The central problem that our paper addresses is how to minimize an objective function over \(G_k^r\) (or \(\mathcal{P}_k\))? In this section we describe an approach based on penalty functions and kernels.

Let us assume for simplicity that \(M\) is the gaussian kernel, i.e. \(M(x, y) = G_\sigma^r(x - y)\) where \(G_\sigma^r(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{|x|^2}{2\sigma^2}}\). Besides the gaussian kernel our theory also captures many other kernels,
including cases of the Abel Kernel $\frac{1}{\sigma^2} e^{-\frac{|x-y|}{\sigma}}$ and the Fourier transform of the Abel Kernel, the Poisson kernel:

$$\frac{1}{\sigma^2+|x-y|^2}$$.

For $f, g \in S(\mathbb{R}^n)$ let us denote:

$$\langle f, M | g \rangle \overset{\text{def}}{=} \int_{\mathbb{R}^n \times \mathbb{R}^n} f(x)^* M(x, y) g(y) dx dy \leq \max_{x, y} M(x, y) \|f\|_1 \|g\|_1 < \infty$$

For general $f, g \in S'(\mathbb{R}^n)$ the expression $\langle f, M | g \rangle$ is defined if $\exists f_x, g_x \in S(\mathbb{R}^n)$ such that $T_{f_x} = f \ast G^n_x, T_{g_x} = g \ast G^n_x$ and $\langle f_x, M | g_x \rangle \to A$. Then, $\langle f, M | g \rangle \overset{\text{def}}{=} A$. For example, $\langle \delta^n | M | \delta^n \rangle = M(0, 0)$.

Theorem 3 concludes, from $f \in \mathcal{G}_k$, that $\dim \text{span}\{x_1 f, x_2 f, \cdots, x_n f\} \leq k$. Using the kernel $M$, one can build the Gram matrix from the collection of distributions, \([\langle x_i f | M | x_j f \rangle]_{1 \leq i, j \leq n}\). For any $f \in S'(\mathbb{R}^n)$ let us denote a real part of the Gram matrix $\{\text{Re} \langle x_i f | M | x_j f \rangle\}_{1 \leq i, j \leq n}$ by $M_f$ (if it is defined).

**Theorem 4.** If $f \in \mathcal{G}_k$, then $\langle x_i f | M | x_j f \rangle$ is defined and rank $M_f \leq k$.

**Definition 1.** Let $A \in \mathbb{R}^{n \times n}$ be a positive semidefinite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ (with counting multiplicities). Then, the Ky Fan $k$–anti-norm of $A$ is $\|A\|_k = \sum_{i=1}^k \lambda_{n+1-i, k}$.

Let $R(f) = \|M_f\|_{n-k}$. Theorem 4 tells us that for $f \in \mathcal{G}_k, R(f) = 0$. For ordinary $f$, the Eckart-Young-Mirsky theorem gives us $R(f) = \min_{A \in \mathbb{R}^{n \times n}, \text{rank } A \leq k} \|\sqrt{M_f} - A\|^2_\infty$. Thus, by penalizing the value of $R(f)$, we enforce $M_f$ to be close to some matrix of rank $k$. For $I : \mathcal{G}_k \cup S(\mathbb{R}^n) \to \mathbb{R}^+$, it is natural to reduce the optimization task over tempered distributions

$$I(f) \to \min_{f \in \mathcal{G}_k} \|f\|_k$$

**The Alternating Scheme**

We will concentrate on problem 11. It is known [Hiai, 2013] that the Ky Fan anti-norm is a concave function, i.e. $R(\phi) = \|M_\phi\|_{n-k}$ depends on $M_\phi$ in a concave way. It can be shown that the dependence of $R(\phi)$ on $\phi$ is both non-convex and non-concave, i.e. we deal with a non-convex optimization task.

The kernel $M(x, y) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{C}$ induces a linear operator from $L_2(\mathbb{R}^n)$ to $L_2(\mathbb{R}^n)$: $O(M)[f] = \int_{\mathbb{R}^n} M(x, y) f(y) dy$. For any operator $O$ between spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, we denote its range as $\mathcal{R}(O) = \{O(x) | x \in \mathcal{H}_1\}$. Let $B(\mathcal{H}_1, \mathcal{H}_2)$ denote a set of bounded linear operators between Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$. For $O \in B(\mathcal{H}_1, \mathcal{H}_2)$ the rank of $O$ is defined as $\dim \mathcal{R}(O)$. Let $L_2^r(\mathbb{R}^n)$ be the Hilbert space (over $\mathbb{R}$) of real-valued functions from $L_2(\mathbb{R}^n)$ and $L_2(\mathbb{R}^n) = L_2^r(\mathbb{R}^n) \times L_2^r(\mathbb{R}^n)$. The space $L_2^r(\mathbb{R}^n)$ is equivalent to $L_2(\mathbb{R}^n)$ treated as a linear space over $\mathbb{R}$. Below we do not distinguish $[\phi_1, \phi_2] \in L_2^r(\mathbb{R}^n)$ and $\phi_1 + i\phi_2 \in L_2^r(\mathbb{R}^n)$. It is easy to see that any $O \in B(L_2^r(\mathbb{R}^n), \mathbb{R}^n)$ can be given by formula:

$$O[\phi] = \text{Re} \langle O, \phi \rangle_{L_2^r(\mathbb{R}^n)}, O_i \in L_2(\mathbb{R}^n), i = 1, \ldots, n$$

i.e. $O \in B(L_2^r(\mathbb{R}^n), \mathbb{R}^n)$ can be identified with a vector of functions $O = [O_i]_{i=1}^{\infty}$, $O_i \in L_2(\mathbb{R}^n)$ and the Hilbert–Schmidt norm on $B(L_2^r(\mathbb{R}^n), \mathbb{R}^n)$ (i.e. $\sqrt{\text{Tr} O^\dagger O}$) is:

$$\|O\|_2 = \left( \sum_{i=1}^n \|O_i\|_{L_2^r(\mathbb{R}^n)}^2 \right)^{\frac{1}{2}}$$

(12)
Recall that for the kernel $M$, $O(M)$ is positive and self-adjoint. Since $O(M)$ is also bounded, then the square root $\sqrt{O(M)}$ can be correctly defined \cite{Rudin1991}. For any complex-valued function $f$ let us introduce a linear operator $S_f : L^2_2(\mathbb{R}^n) \to \mathbb{R}^2$ by the following rule:

$$S_f[\phi]_i = \Re \langle x_i f(x), \sqrt{O(M)}[\phi] \rangle_{L^2(\mathbb{R}^n)} \text{ i.e. } (S_f)_i = \sqrt{O(M)}|x_i f(x)|, i = 1, n$$

**Theorem 5.** If $\text{Tr} M_f < \infty$, then $S_f \in B(L^2_2(\mathbb{R}^n), \mathbb{R}^n)$ and $S_f S_f^\dagger = M_f$. Moreover,

$$R(f) = \min_{S \in B(L^2_2(\mathbb{R}^n), \mathbb{R}^n), \text{rank } S \leq k} \| S_f - S \|^2$$

and the minimum is attained at $S = P_f S_f$ where $P_f = \sum_{i=1}^k u_i u_i^T$ and $\{u_i\}^k_i$ are unit eigenvectors of $M_f$ corresponding to the $k$ largest eigenvalues (counting multiplicities).

Given the new representation $R(f) = \min_{S \in B(L^2_2(\mathbb{R}^n), \mathbb{R}^n), \text{rank } S \leq k} \| S_f - S \|^2$ it is natural to view the Task \cite{1} as a minimization of $I(\phi) + \lambda \| S_\phi - S \|^2$ over two objects: $\phi$ and $S \in B(L^2_2(\mathbb{R}^n), \mathbb{R}^n)$ : \text{rank } $S \leq k$. The simplest approach to minimize a function over two arguments is to optimize alternately, i.e. first over $\phi$, and then over $S : \text{rank } S \leq k$, and so on. Theorem \cite{5} gives that the minimization over $S$ is equivalent to the truncation of $\text{SVD}(S_\phi)$ at the $k$-th term. This idea, that we dub the alternating scheme, is described in Algorithm \cite{1}.

**Algorithm 1** The alternating scheme for \cite{1}

$P_0 \leftarrow 0, S_{\phi_0} \leftarrow 0$

for $t = 1, \ldots, T$ do

$\phi_t \leftarrow \arg\min_{\phi} I(\phi) + \lambda \| S_\phi - P_{t-1} S_{\phi_{t-1}} \|^2$ (minimizing over $\phi$)

Calculate $M_{\phi_t}$ and find $\{v_i\}_i^n$ s.t. $M_{\phi_t} v_1 = \lambda_1 v_1, \lambda_1 \geq \cdots \geq \lambda_n$

$P_t \leftarrow \sum_{i=1}^k v_i v_i^T$ (Truncated SVD($S_{\phi_t}$) is $P_t S_{\phi_t}$)

Output: $v_1, \ldots, v_k$

The alternating scheme \cite{1} allows for a reformulation in the dual space. By this we mean that in Scheme \cite{1} we substitute $\phi_t$ for the original $\phi_t$. If the primal Scheme \cite{1} deals with operators $S_\phi, S_{\phi_{t-1}}$, the dual version deals with vectors of functions $\sqrt{G_\sigma} \frac{\partial \hat{G}_{\sigma}}{\partial \hat{u}_1} \sqrt{G_\sigma} \frac{\partial \hat{G}_{\sigma-1}}{\partial \hat{u}_1}$. Details of the dual algorithm can be found in the appendix \cite{1}.

**7 Experiments**

The alternating scheme \cite{1} is a general optimization method which needs to be specified for every optimization task. We designed numerical specifications of the alternating scheme \cite{1} for all 4 optimization tasks: \cite{6, 7, 8, 9} and made experiments with all of them. Details of the algorithms, i.e. numerical methods to minimize over $\phi$ and calculate $M_{\phi_t}$, can be found in the appendix (C, \cite{1} and \cite{1}). Note that for Wasserstein distance minimization \cite{8} we exploit the alternating scheme in the initial form (i.e. \cite{1}), and for MMD \cite{6, 7} and SDR \cite{9} we use the dual version of the scheme.

**Behaviour of MMD for small $h$.** We studied the difference in the behaviour of PCA and a solution to \cite{8} obtained by the alternating scheme \cite{1} (MMD), for the case when $h$ is small compared to the standard deviation of features. Experiments show that they are sharply different when data points are sampled along a low-dimensional manifold \$2\$, which is bent globally, goes through the origin \$O$ and has a large curvature at \$O$. Because PCA is a global method and points do not lie on an affine subspace, interpreting principal directions is not straightforward.

We select a smooth function $f : \mathbb{R}^{n-1} \to \mathbb{R}$, such that $f(0) = 0$ and generate points in the following way: points $x_1, x_2, \ldots, x_N \sim [-10, 10]^{n-1}$ are sampled uniformly, after calculation of $y_i = f(x_i)$ we add some noise: $z_i = (x_i, y_i) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, 0.01I_n)$. Both PCA and MMD are applied to the dataset (first 3 pictures on Figure \cite{13}). As we see, MMD, unlike PCA, tries to catch ideal alignments of points rather that searching for a global alignment of points (which can be non-existent). This property of MMD makes it a promising tool for the calculation of the tangent space.
Experiments with outlier detection (MMD, HM, Wasserstein distance). Following the experiment setup of [Xu et al. (2010)], we choose parameters \( N = n = 400, \delta = 0.05(0.1), k = 10 \) and generate random matrices \( A \in \mathbb{R}^{N(1-\delta) \times k}, B \in \mathbb{R}^{n \times k} \) whose entries are i.i.d as \( \mathcal{N}(0,1) \). Then, to the columns of the matrix \( BA^T \in \mathbb{R}^{n \times N(1-\delta)} \) (whose rank is \( \leq k \)) are concatenated with the columns of the matrix \( C \in \mathbb{R}^{n \times N\delta} \): \( X = \text{concat}(BA^T,C) \in \mathbb{R}^{n \times N} \). The entries in \( C \) are either iid as \( \mathcal{N}(0,1) \) (case I) or \( \mathcal{N}(0,1) \) (case II). Let \( X = [x_1, \ldots, x_M] \), i.e. columns of \( X \) are the data points. Thus, \( N(1-\delta) \) columns of \( BA^T \) lie in a \( k \)-dimensional subspace of \( \mathbb{R}^n \) and \( N\delta \) columns of \( C \) are outliers, and solutions of tasks 6, 7, or 8 for this dataset are expected to be supported in a column space of \( BA^T \).

After every iteration (step \( t \) of the alternating scheme [1]) we calculate the Frobenius distance between the projection operator \( P_t \) of [1] and the projection operator \( P \) to the column space of \( BA^T \), i.e. \( \|P_t - P\|_F \). For the task 8, the dependence of \( \|P_t - P\|_F \) on \( t \) for different values of parameters \( \delta \) and \( \lambda \) is shown in Figure 1b. For tasks 6, 7 the behaviour of the alternating scheme is similar, 7 iterations are enough to approach the optimal subspace.

Besides the speed of convergence we were also interested in how \( \|P^* - P\|_F \), where \( P^* = \lim_{t \to \infty} P_t \) is the final projection operator (e.g. \( P_{20} \) in practice), depends on the parameter \( \sigma \) of the kernel \( M = G_{\sigma} \). It is natural to expect the quality of the solution \( P^* \) to degrade as \( \sigma \to +\infty \) (this corresponds to \( M(x,y) \to 0 \)), and, less trivially, as \( \sigma \to 0 \) (this corresponds to \( M(x,y) \to \delta^\infty(x-y) \)).

Experiments with the sufficient dimension reduction. We made experiments on the standard datasets, [Heart, Breast Cancer, Ionosphere, Diabetes, Boston house prices and Wine quality]. First we applied Sliced Inverse Regression algorithm (SIR) [Li (1991)] to the training set and calculated the effective subspace for \( k = 2, 3 \). All points were projected onto that space and we obtained two- or three-dimensional representations of input points. In the last step we applied 10 nearest neighbours algorithm (KNN) to predict outputs (based on reduced inputs) on the test set (for the regression case, the 10-KNN regression was used). The same scheme was repeated with PCA, Kernel Dimensionality Reduction (KDR) algorithm [Fukumizu et al. (2004)] and the alternating scheme [1] adapted for SDR.
We experimented with the dual version of algorithm setting (after the data was standardized) the kernel’s parameter $\sigma = 0.81$ and $\lambda = 10.0$. Details of its numerical implementation can be found in the appendix. In the table one can see the obtained test set accuracy on the classification tasks and $R^2$ on the regression tasks. As we see from the table, after reducing the dimension of an input to $k = 2$, we are still able to obtain good accuracy of prediction on a test set.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Dimension $k$</th>
<th>2</th>
<th>3</th>
<th>2</th>
<th>3</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart (acc)</td>
<td>PCA</td>
<td>79.80</td>
<td>92.49</td>
<td>86.33</td>
<td>88.77</td>
<td>81.48</td>
<td>83.50</td>
<td></td>
</tr>
<tr>
<td>Breast (acc)</td>
<td>SIR</td>
<td>93.46</td>
<td>97.30</td>
<td>93.13</td>
<td>95.95</td>
<td>97.88</td>
<td>97.69</td>
<td></td>
</tr>
<tr>
<td>Ionosphere</td>
<td>KDR</td>
<td>80.29</td>
<td>89.14</td>
<td>43.47</td>
<td>43.61</td>
<td>41.82</td>
<td>44.30</td>
<td></td>
</tr>
<tr>
<td>Diabetes ($R^2$)</td>
<td>AS 1</td>
<td>25.34</td>
<td>76.03</td>
<td>77.88</td>
<td>79.97</td>
<td>73.21</td>
<td>77.88</td>
<td></td>
</tr>
<tr>
<td>Boston ($R^2$)</td>
<td></td>
<td>56.42</td>
<td>98.68</td>
<td>98.30</td>
<td>96.02</td>
<td>97.10</td>
<td>96.93</td>
<td></td>
</tr>
<tr>
<td>Wine ($R^2$)</td>
<td></td>
<td>93.91</td>
<td>98.29</td>
<td>98.29</td>
<td>44.48</td>
<td>44.48</td>
<td>44.48</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: The cross-validated accuracies/$R^2$ of KNN on 2 or 3-dimensional input representations.

The code is available on [github](https://github) to facilitate the reproducibility of our results.

8 RELATED WORK

We present an optimization framework in which the search space is $G'_k$, or $P_k$. Another unifying framework for LDR tasks is suggested by Cunningham & Ghahramani (2015) in which the basic search space is the Stiefel manifold $S(n, k)$. The main disadvantage of using $G'_k$, instead of the Stiefel manifold, is that its infinite number of dimensions requires a special procedure to turn an optimization into a finite-dimensional task. Both an optimization over $G'_k$ and over $S(n, k)$ is typically hard: for a final point, at best one can guarantee that it is a local extremum. Promising aspects of $G'_k$ are: a) $G'_k$ allows to formulate naturally a new class of objectives on it, b) local extrema on $G'_k$ substantially differ from local extrema on $S(n, k)$, because a local search over $G'_k$ uses more degrees of freedom.

Using Ky-Fan $k$-antinorm as a regularizer for the matrix completion problem has been suggested by Hu et al. (2013) and further developed in Oh et al. (2016), Liu et al. (2016), Hong et al. (2016). Unlike this chain of works, we formulate an infinite-dimensional task. Also, our regularizer $R(f) = \|Mf\|_{n-k}$ is a sum of smallest $n-k$ squared singular values of the operator $S_f$ where $S_f$ depends on $f$ linearly. The idea of alternating two basic stages, the convex optimization and SVD, is ubiquitous in low-rank optimization, see e.g. Mazumder et al. (2010), Hastie et al. (2015).


9 CONCLUSIONS

We develope a new optimization framework for LDR problems. The alternating scheme for the optimization task demonstrates both the computational efficiency and the applicability to real-world data. The algorithm performs quite stably when we vary most of the hyperparameters, though it crucially depends on two parameters, the bandwidth of the “smoothing” kernel $M$, $\sigma$, and the penalty parameter $\lambda$. We believe that the MMD/HM/WD methods for UDR could be used as an alternative to PCA in study fields in which data demonstrate “heavy-tailed” and “non-gaussian” behaviour, such as financial applications. Also, our formulation of SDR is free from any assumptions on the distribution of input-output pairs, which makes it an alternative to other methods of the efficient subspace estimation. More detailed report on these topics is a subject of future research.

Since the role of the parameter $\sigma$ is similar to that of the bandwidth in the kernel density estimation, we use Silverman’s rule of thumb to set $\sigma = N^{-1/(n+4)}$. 


REFERENCES


A PROOFS FOR SECTION 3

A.1 PROOF OF THEOREM 1 GIVEN FOR COMPLETENESS

Proof. The inclusion $G'_k \subseteq \overline{G_k}^*$ follows from a well-known fact that $S(\mathbb{R}^k)$ is dense in $S'(\mathbb{R}^k)$. I.e. for any $f \in S'(\mathbb{R}^k)$ one can always find a sequence $\{f_i\} \subseteq S(\mathbb{R}^k)$ such that $T f_i \to f$. Therefore, for any $(f \otimes \delta^{n-k})_U \in G'_k$ there is a sequence $\{(T f_i \otimes \delta^{n-k})_U\} \subseteq G_k$ such that $(T f_i \otimes \delta^{n-k})_U \to (f \otimes \delta^{n-k})_U$. Thus, $G'_k \subseteq \overline{G_k}^*$.  

Since $G_k \subseteq G_k'$, to prove $G_k' = \overline{G_k'}^\sigma$ it is enough to show that $G_k'$ is sequentially closed.

We need a simple fact from a theory of distributions.

**Lemma 1.** If $T_i \to^\ast T$ and $\phi_i \to \phi$, then $\langle T_i, \phi_i \rangle \to \langle T, \phi \rangle$.

**Proof of Lemma.** Schwartz space $S(\mathbb{R}^n)$ is a Fréchet space, therefore the Banach-Steinhaus theorem applies to $S'(\mathbb{R}^n)$. Since $T_i \to^\ast T$, we have $\sup_i |\langle T_i, \phi \rangle| < \infty$ for any $\phi \in \mathcal{F}(\mathbb{R}^n)$. From the Banach-Steinhaus theorem, applied to a set $\{T_i\}_i$, we obtain for any $\epsilon > 0$, there is a neighbourhood $U$ of $0 \in S(\mathbb{R}^n)$ such that $|\langle T_i, \phi \rangle| < \epsilon$ whenever $\phi \in U$. Thus, $|\langle T_i, \phi_i - \phi \rangle| < \epsilon$ for a large enough $i$. From that we conclude that $\langle T_i, \phi_i \rangle \to \langle T, \phi \rangle$.

For any $T \in S'(\mathbb{R}^n)$ and $\psi \in \mathcal{S}(\mathbb{R}^{n-k})$, let us define $T^\psi \in S'(\mathbb{R}^k)$ as $\langle T^\psi, \phi \rangle = \langle T, \phi \otimes \psi \rangle$.

Suppose that $\{f_i\}_i \subseteq S'(\mathbb{R}^k)$, $\{U_i\}_i$ are such that $(f_i \otimes \delta^{n-k})_{U_i} \to^* f$. We need to prove that $f \in G_k'$. Since a set of orthogonal matrices is compact, then one can always find a subsequence $\{U_{n_i}\}$ such that $U_{n_i} \to U$. Since $(f_{n_i} \otimes \delta^{n-k})_{U_{n_i}} \to^* f$ and $\phi(U_{n_i}) \to \phi(U)\phi(U)$ (for any fixed $\phi \in \mathcal{S}(\mathbb{R}^n)$), using Lemma 1 we obtain:

$$\langle f_{n_i} \otimes \delta^{n-k}, \phi \rangle = \langle f_{n_i} \otimes \delta^{n-k}, \phi(U_{n_i}) \rangle \to \langle f, \phi(U)\phi(U) \rangle = \langle f_U, \phi(U) \rangle$$

Thus, we have:

$$f_{n_i} \otimes \delta^{n-k} \to^* f_U$$

From the last we see that $f_{n_i} \to^* f_U$ where $\psi$ is such that $\psi(0) = 1$. Therefore, $f_U = f_U \otimes \delta^{n-k}$ and $f = (f_U \otimes \delta^{n-k})_U \in G_k'$.

**A.2 PROOF OF THEOREM**

**Proof.** Let us prove first that if $g = T_f \otimes \delta^{n-k}$, then

$$F[g] = T_r$$

where $r(x) = \hat{f}(x_1:k), x \in \mathbb{R}^n$. For that we have to prove that $\langle F[g], \phi \rangle = \langle T_r, \phi \rangle$ for any $\phi \in \mathcal{S}(\mathbb{R}^n)$. Indeed,

$$\langle F[g], \phi \rangle = \langle g, F[\phi] \rangle = \langle T_f \otimes \delta^{n-k}, \int_{\mathbb{R}^n} \phi(y)e^{-ix^Ty}dy \rangle =$$

$$\langle T_f, \int_{\mathbb{R}^n} \phi(y)e^{-ix_1:k:Y_1:k}dy \rangle = \int_{\mathbb{R}^n} \hat{f}(r(x))\phi(y)e^{-ix_1:k:Y_1:k}dy dx_1:k =$$

$$\int_{\mathbb{R}^n} \hat{f}(y)\phi(y)dy = \langle T_r, \phi \rangle$$

Let us calculate the image of $G_k$ under the Fourier transform. It is easy to see that for any $g \in S'(\mathbb{R}^n)$, $\phi \in \mathcal{S}(\mathbb{R}^n)$ and orthogonal $U \in \mathbb{R}^{n \times n}$ we have:

$$\langle F[gu], \phi(x) \rangle = \langle gu, F[\phi](x) \rangle = \langle g, F[\phi](U^Tx) \rangle =$$

$$= \langle g, F[\phi(U^Tx)] \rangle = \langle F[g], \phi(U^Tx) \rangle = \langle (F[g])_U, \phi(x) \rangle$$

Therefore, $F[gu] = (F[g])_U$. Thus, if $g = T_f \otimes \delta^{n-k}$, then

$$(F[g])_U = (T_r)_U = T_r'$$

where $r'(x) = r(Ux) = \hat{f}(U_kx)$ and $U_k \in \mathbb{R}^{k \times n}$ is a matrix consisting of first $k$ rows of $U$. Thus, $T_{r'} \in \mathcal{F}_k$.

Let us show that by varying $f \in S(\mathbb{R}^k)$ and $U$ in the expression $\hat{f}(U_kx)$ we can obtain any function from $\mathcal{F}_k$. For this it is enough to show that $\mathcal{F}_k$ is equivalent to the following set of functions:

$$Q = \{ g(U_kx) | g \in \mathcal{S}(\mathbb{R}^k), U_k \in \mathbb{R}^{k \times n}, U_kU_k^T = I_k \}$$

The fact $Q \subseteq \mathcal{F}_k$ is obvious. Let us now prove that $Q \supseteq \{ g(Px) | g \in \mathcal{S}(\mathbb{R}^k), P \in \mathbb{R}^{k \times n}, \text{rank } P = k \} = \mathcal{F}_k$. Indeed, if $f(x) = g(Px)$, then $\hat{f}(x) = g'(U_kx)$ where $U_k = (PP^T)^{-1/2}P$ and $g'(y) = g((PP^T)^{-1/2}y)$. By construction, $U_kU_k^T = I_k$ and $g' \in \mathcal{S}(\mathbb{R}^k)$. Thus, $Q = \mathcal{F}_k$.

Therefore, $F[G_k] = \mathcal{F}_k$, and from the bijectivity of the Fourier transform we obtain $F^{-1}[\mathcal{F}_k] = G_k$.
A.3 Proof of Theorem 3

Proof of Theorem 3 (⇒). Let us prove that from $T = (f \otimes \delta^{n-k})_U, f \in S'(\mathbb{R}^k), U^TU = I_n$ it follows that dim $\text{span}_R \{x_1T, x_2T, \ldots, x_nT\} \leq k$.

It is easy to see that $x_i[f \otimes \delta^{n-k}] = 0$ if $i > k$. If $U = [u_1, \ldots, u_n]^T$, then for $i > k$ we have $0 = (x_i[f \otimes \delta^{n-k}])_U = u_i^T f(x \otimes \delta^{n-k})_U = u_i^T x$. Thus, we have $n - k$ orthogonal vectors, $u_{k+1}, \ldots, u_n$, such that $[x_1T, \ldots, x_nT]^T u_i = 0$. Using standard linear algebra we obtain there are at most $k'$ distributions $x_i, T, \ldots, x_{i+k'}, k' \leq k$ that form a basis of $\text{span}_R \{x_iT\}^n$.

To prove the second part of the theorem we need the following lemma.

Lemma 2. If $T \in S'(\mathbb{R}^n)$ is such that $y_iT = 0$ for any $i > k$, then $T \in \mathcal{G}_k'$.

Proof of Lemma. Recall from functional analysis, for $f \in S'(\mathbb{R}^n)$, the tempered distribution $\frac{\partial f}{\partial x_i}$ is defined by the condition $\langle \frac{\partial f}{\partial x_i}, \phi \rangle = -\langle f, \frac{\partial \phi}{\partial x_i} \rangle$. Once the Fourier transform is applied, our lemma’s dual version is equivalent to the following formulation: if $\frac{\partial f}{\partial x_i} = 0, i > k$, then $f \in \mathcal{T}'_k$. Let us prove it in this formulation.

A set of infinitely differentiable functions with a compact support is denoted by $C_c^\infty(\mathbb{R})$. Suppose $\phi \in S(\mathbb{R}^n)$ and $p \in C_c^\infty(\mathbb{R})$ are chosen in such a way that $\int_{-\infty}^{\infty} p(y_i)dy_i = 1, \text{supp} p \subset [A, B]$. Let us define:

$$r(x) = \int_{-\infty}^{x_i} \phi(x, y_i)dy_i - \int_{-\infty}^{x_i} p(y_i)dy_i \int_{-\infty}^{\infty} \phi(x, y_i)dy_i$$

It is easy to see that for any $\alpha \in \mathbb{N}^{n-1}, \alpha' \in \mathbb{N}, \beta \in \mathbb{N}^{n-1}, \beta' \in \mathbb{N}$ we have (at least one derivative over $x_i$ is present):

$$x_\alpha x_\alpha' \int_{\mathbb{R}^n} x_{i+1} \frac{\partial^{\beta+1+\beta'}}{\partial x_{i+1}^{\beta} \partial x_{i+1}^{\beta'}} = x_\alpha x_\alpha' \int_{\mathbb{R}^n} x_{i+1} \frac{\partial^{\beta+1+\beta'}}{\partial x_{i+1}^{\beta} \partial x_{i+1}^{\beta'}}$$

The terms $x_\alpha x_\alpha' \int_{\mathbb{R}^n} x_{i+1} \frac{\partial^{\beta+1+\beta'}}{\partial x_{i+1}^{\beta} \partial x_{i+1}^{\beta'}}$ are bounded by the definition of $S(\mathbb{R}^n), C_c^\infty(\mathbb{R})$. The boundedness of $\int_{-\infty}^{\infty} x_\alpha x_\alpha' \int_{\mathbb{R}^n} x_{i+1} \frac{\partial^{\beta+1+\beta'}}{\partial x_{i+1}^{\beta} \partial x_{i+1}^{\beta'}} dy_i$ is a consequence of the inequality (which holds because $\phi \in S(\mathbb{R}^n)$): $|x_\alpha \int_{\mathbb{R}^n} \partial^{\beta} \phi(x, y_i)dy_i| \leq C'_{1+y_i^{\alpha+1}}$.

Analogously (not a single derivative over $x_i$ is present):

$$x_\alpha x_\alpha' \int_{\mathbb{R}^n} x_{i+1} \frac{\partial^{\beta+1+\beta'}}{\partial x_{i+1}^{\beta} \partial x_{i+1}^{\beta'}} = x_\alpha x_\alpha' \int_{-\infty}^{\infty} x_{i+1} \frac{\partial^{\beta+1+\beta'}}{\partial x_{i+1}^{\beta} \partial x_{i+1}^{\beta'}}$$

The second term is 0 when $x_i \leq A$. It is also bounded when $x_i > A$ because $|x_\alpha \int_{\mathbb{R}^n} \partial^{\beta} \phi(x, y_i)dy_i| \leq C'_{1+y_i^{\alpha+1}}$ and:

$$|x_\alpha \int_{x_i}^{\infty} x_{i+1} \frac{\partial^{\beta} \phi(x, y_i)dy_i}{}| \leq |x_i|^{\alpha'} \int_{x_i}^{\infty} C'_{1+y_i^{\alpha+1}} dy_i$$

The latter is bounded because $\lim_{x_i \to +\infty} |x_i|^{\alpha'} \int_{x_i}^{\infty} C'_{1+y_i^{\alpha+1}} dy_i = 0$. 

Under review as a conference paper at ICLR 2021
The first term is 0 when $x_i \geq B$ and it is bounded for $x_i < B$:

$$\left|x_i^{1/2} \int_{-\infty}^{x_i} \frac{\partial^2 \phi(x_i, y_i)}{\partial y^2_i} dy_i \right| \leq |x_i|^{1/2} \int_{-\infty}^{x_i} \frac{C'}{(1 + y_i^2)^{\alpha + 1}} dy_i$$

The latter is also bounded, since $\lim_{x_i \to -\infty} |x_i|^{1/2} \int_{-\infty}^{x_i} \frac{C'}{(1 + y_i^2)^{\alpha + 1}} dy_i = 0$.

Thus, $x^\alpha \frac{\partial^2 \phi}{\partial y^2}$ is bounded and $r \in S(\mathbb{R}^n)$. Therefore $\frac{\partial f}{\partial x_i} = 0$ implies:

$$\langle f, \frac{\partial r}{\partial x_i} \rangle = 0 \Rightarrow f[\phi] = f[p(x_i)] \int_{-\infty}^{\infty} \phi(x_i, y_i) dy_i$$

Since this sequence of arguments works for any $i > k$, we can apply them sequentially to initial $\phi \in S(\mathbb{R}^n)$ w.r.t. $x_{k+1}, \ldots, x_n$ and obtain for any $p_{k+1}, \ldots, p_n \in C_c(\mathbb{R})$ such that $\int_{-\infty}^{\infty} p_i(y_i) dy_i = 1$:

$$f[\phi] = f[p_{k+1}(x_{k+1}) \cdots p_n(x_n)] \int_{\mathbb{R}^{n-k}} \phi(x_{1:k}, x_{k+1:n}) dx_{k+1:n}$$

Moreover, since $C_c^\infty(\mathbb{R})$ is dense in $S(\mathbb{R})$, we can assume that $p_{k+1}, \ldots, p_n \in S(\mathbb{R})$. For the inverse Fourier transform $T = F^{-1}[f]$ the latter condition becomes equivalent to:

$$\langle T, \phi \rangle = \langle T, p_{k+1}(x_{k+1}) \cdots p_n(x_n) \rangle \phi(x_{1:k}, 0_{k+1:n})$$

for any $p_{k+1}, \ldots, p_n \in S(\mathbb{R})$ such that $p_i(0) = 1$. Let us define $p_i(x_i) = e^{-x_i^2}$. It is easy to check that $T = g \otimes \delta_{n-k}$ where $g \in S(\mathbb{R}^k), \langle g, \psi \rangle = \langle T, e^{-|x_{k+1:n}|^2} \psi(x_{1:k}) \rangle$ for $\psi \in S(\mathbb{R}^k)$. I.e. $T \in G'_k$ and lemma is proved. □

**Proof of Theorem 3.** If $\dim \text{span}_n \{x_1 T, x_2 T, \ldots, x_n T\} \leq k$, then

$$\dim \{v \in \mathbb{R}^n | [x_1 T, \ldots, x_n T] v = 0 \} \geq n - k$$

Thus, there exist at least $n - k$ orthonormal vectors $v_{k+1}, \ldots, v_n$, such that $[x_1 T, \ldots, x_n T] v_i = 0$. Therefore, $[x_1 T, \ldots, x_n T] v_i = (v_i^T x) T = 0$.

Let us complete $v_{k+1}, \ldots, v_n$ to form an orthonormal basis of $\mathbb{R}^n$: $v_1, \ldots, v_n$. Let us define a matrix $V = [v_1, \ldots, v_n]$. It is easy to see that:

$$((v_i^T x) V) V = (v_i^T V x) V = x_i T V$$

Since for $i > k$ we have $(v_i^T x) T = 0$, then $x_i T V = 0$. Using lemma 2 we obtain $T V \in G'_k$. Therefore, $(TV) V^T = T \in G'_k$. Theorem proved. □

**B STRUCTURE OF WD**

Recall that $(\mathbb{R}^n, || \cdot ||)$ is a Banach space. Now, let us consider an optimization problem: for a given $X \in \mathbb{R}^{n \times N}$ solve

$$||X - L|| \rightarrow \min_{\text{rank}(L) \leq k}$$

where $|| \cdot ||$ is extended to $\mathbb{R}^{n \times N}$ by $||[s_1, \ldots, s_N]|| \overset{\text{def}}{=} \sum_i ||s_i||$.

The following simple theorem shows that the two tasks are connected, so that the solution of one directly leads to the solution of another.

**Theorem 6.** Given data points $\{x_1, \ldots, x_N\}$, let $X = [x_1, \ldots, x_N] \in \mathbb{R}^{n \times N}$. Then,

$$\min_{\nu \in P_n} W(\mu_{\text{data}}, \nu) = \frac{1}{N} \min_{\nu \in P_n} \max_{Y \in \mathbb{R}^{n \times N}, \text{rank}(Y) \leq k} ||X - Y||$$

Moreover, $\min_{\nu \in P_n} W(\mu_{\text{data}}, \nu)$ is attained on $\nu^*$, where $\nu^*$ is a uniform distribution over $\{y_i\}_{i=1}^N$ and $[y_1, \ldots, y_N] \in \arg\min_{Y \in \mathbb{R}^{n \times N}, \text{rank}(Y) \leq k} ||X - Y||$. 

13
Proof. Let us prove first that $\inf_{\mu \in \mathcal{P}_k} W(\mu_{\text{data}}, \mu) \leq \frac{1}{N} ||X - Y^*||$ where $Y^* = [y_1, \cdots, y_N] \in \arg \min_{Y \in \mathbb{R}^{n \times N}, \text{rank}(Y) \leq k} ||X - Y||$

Let $\pi$ be a uniform distribution over $\{(x_i, y_i)\}_{i=1}^N$ and $\mu^*$ be a uniform distribution over $\{y_i\}_{i=1}^N$. Since $\pi \in \Pi(\mu_{\text{data}}, \mu^*)$, we obtain $W(\mu_{\text{data}}, \mu^*) \leq \frac{1}{N} \sum_{i=1}^N ||x_i - y_i|| = \frac{1}{N} ||X - Y^*||$. The support of $\mu^*$ is $k$-dimensional, because $\text{rank}(Y^*) \leq k$. Thus, we have $\mu^* \in \mathcal{P}_k$ and $\inf_{\mu \in \mathcal{P}_k} W(\mu_{\text{data}}, \mu) \leq W(\mu_{\text{data}}, \mu^*) \leq \frac{1}{N} ||X - Y^*||$. Now, if we prove the inverse inequality, i.e. $\inf_{\mu \in \mathcal{P}_k} W(\mu_{\text{data}}, \mu) \geq \frac{1}{N} ||X - Y^*||$, this will imply that $\inf_{\mu \in \mathcal{P}_k} W(\mu_{\text{data}}, \mu) = \frac{1}{N} ||X - Y^*||$ and therefore, $\inf_{\mu \in \mathcal{P}_k} W(\mu_{\text{data}}, \mu) = W(\mu_{\text{data}}, \mu^*)$. This will in the end give us $\mu^* \in \arg \inf_{\mu \in \mathcal{P}_k} W(\mu_{\text{data}}, \mu)$.

Let $\{\mu_t\}_{t=1}^\infty$ be such that $\mu_t \in \mathcal{P}_k$ and $W(\mu_{\text{data}}, \mu_t) - \inf_{\mu \in \mathcal{P}_k} W(\mu_{\text{data}}, \mu) \to 0$. Let $L_t$ denote a $k$-dimensional support of $\mu_t$ and $P_t$ is a projection operator onto $L_t$.

Let $\mu_t^*$ be a uniform distribution over $\{P_t x_1, \cdots, P_t x_N\}$, i.e. $\mu_t^*(A) = \frac{1}{N} \sum_{i=1}^N |P_t x_i \in A|$. It is easy to see that $W(\mu_t^*, \mu_{\text{data}}) \leq W(\mu_t, \mu_{\text{data}})$, because $\mu_t^*$ and $\mu_t$ share the same $k$-dimensional support $L_t$, but the “transportation of a mass” concentrated in point $x_i$ of the empirical distribution $\mu_{\text{emp}}$ can be most optimally done by just moving it to $P_t x_i$ (i.e. to the closest point on $L_t$). Thus, we have $\inf_{\mu \in \mathcal{P}_k} W(\mu_{\text{data}}, \mu) \leq W(\mu_{\text{data}}, \mu_t^*) \leq W(\mu_{\text{data}}, \mu_t)$, and therefore, $W(\mu_{\text{data}}, \mu_t^*) - \inf_{\mu \in \mathcal{P}_k} W(\mu_{\text{data}}, \mu) \to 0$.

Since a set of projection operators is compact, one can always extract a subsequence $\{P_{t_s}\}_{s=1}^\infty$ such that $P_{t_s} \to P$. It is easy to see that $\mu_t^* \to \mu^*$ (i.e. $W(\mu_t^*, \mu^*) \to 0$) where $\mu^*$ is a uniform distribution over $\{P x_1, \cdots, P x_N\}$. For that distribution we have $W(\mu_{\text{data}}, \mu^*) = \lim_{s \to \infty} W(\mu_{\text{data}}, \mu_{t_s}^*) = \inf_{\mu \in \mathcal{P}_k} W(\mu_{\text{data}}, \mu)$. Thus, the infimum is attained on $\mu^*$.

It is easy to see that $W(\mu_{\text{data}}, \mu^*) = W(\mu^*, \mu_{\text{data}}) = \frac{1}{N} ||X - PX||$. Since $\text{rank}(PX) \leq k$ we obtain $W(\mu_{\text{data}}, \mu^*) \geq \frac{1}{N} \min_{Y \in \mathbb{R}^{n \times N}, \text{rank}(Y) \leq k} ||X - Y||$. This completes the proof. \hfill \Box

Note that the case of $L_1$ norm $||x|| = \sum_i |x_i|$ in the task \cite{Candes2011} corresponds to the well-studied \textit{robust PCA} problem \cite{Candes2011}. If, instead of the $L_1$-norm, we use the $L_2$-norm, this leads to another task:

$$||X - L||_{1,2} \to \min_{\text{rank}(L) \leq k}$$

(14)

where $||S||_{1,2} = \sum_j \sqrt{\sum_i s_{ij}^2}$, which known as the \textit{outlier pursuit problem} \cite{Xu2010}.

C \hspace{1em} \textbf{PROPER KERNELS AND PROOF OF THEOREM 4}

C.1 \hspace{1em} \textbf{PROPER KERNELS}

In the main part of the paper we assume $M$ to be a gaussian kernel, though the theory can be applied to a more general case of the so called \textit{proper kernels}.

Recall that for any operator $O$ between spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ we denote its range as $\mathcal{R}[O] = \{O(x) | x \in \mathcal{H}_1\}$. For $\Omega \subseteq \mathcal{S}(\mathbb{R}^n)$, $\overline{\Pi}$ denotes the sequential closure of $\Omega$ with respect to natural topology of $\mathcal{S}(\mathbb{R}^n)$. A set of continuous functions in $\mathbb{R}^n$ is denoted by $C(\mathbb{R}^n)$. A set of infinitely differentiable functions with compact support in $\mathbb{R}^n$ is denoted as $C_c^\infty(\mathbb{R}^n)$

\textbf{Definition 2.} The function $M(x, y): \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{C}$ is called the proper kernel if and only if

- $O(M)[f] = \int_{\mathbb{R}^n} M(x, y) f(y) dy$ is a linear operator from $L_2(\mathbb{R}^n)$ to $L_2(\mathbb{R}^n)$.
- $M(y, x) = M(x, y)^*$.
- $\langle f, O(M)[f] \rangle_{L_2(\mathbb{R}^n)} > 0$, $\forall f \in L_2(\mathbb{R}^n)$, $f \neq 0$.
- $\max_{x, y} |M(x, y)| < \infty$.
- $\mathcal{R}[O(M)] \cap \mathcal{S}(\mathbb{R}^n) = \mathcal{S}(\mathbb{R}^n)$.
The Gaussian kernel \( M(x, y) = G^n_x(x - y) \), which is of special interest from an application-oriented perspective, is captured by the following lemma:

**Lemma 3.** If \( \zeta, \hat{\zeta} \in C(\mathbb{R}^n) \) are bounded, \( \forall x \hat{\zeta}(x) > 0 \), then \( M(x, y) = \zeta(x - y) \) is a proper kernel.

**Proof.** Verification of the first four conditions is easy, so we only check the fifth condition. Let us denote linear operators \( C_f = \zeta * f \) and \( O_g[f](x) = g(x)f(x) \). Then we have \( \mathcal{F}[C_f[L_2(\mathbb{R}^n)]] = O_g[L_2(\mathbb{R}^n)] \supseteq C_{\mathbb{R}}^\infty(\mathbb{R}^n) \). Therefore, \( \mathcal{R}[O(M)] = C_c[L_2(\mathbb{R}^n)] \supseteq \mathcal{F}^{-1}[C_{\mathbb{R}}^\infty(\mathbb{R}^n)] \). Since \( C_{\mathbb{R}}^\infty(\mathbb{R}^n) \) is dense in \( S(\mathbb{R}^n) \), then \( \mathcal{F}^{-1}[C_{\mathbb{R}}^\infty(\mathbb{R}^n)] \) also has this property. I.e. \( \mathcal{R}[O(M)] \cap S(\mathbb{R}^n) = S(\mathbb{R}^n) \).

Besides the Gaussian kernel the lemma also captures a case of the Abel Kernel \( \zeta(x) = e^{-|x|} \). It is well-known that the Fourier transform of the Abel Kernel is the Poisson kernel: \( \hat{\zeta}(x) = \frac{c_n}{(1+|x|^2)^{\frac{n+1}{2}}} \)

(which is also proper).

**C.2 Proof of Theorem 7**

We will prove a more general statement:

**Theorem 7.** Let \( M(x, y) \) be a proper kernel and, additionally, a Lipschitz function. If \( f \in G_k \), then \( \langle x_i | f \rangle \) is defined and rank \( M_f \leq k \).

**Proof.** Let us first show that \( \langle f | M(g) \rangle \) is defined for all \( f, g \in G_k \). Note that for any \( f = (T_\Lambda \otimes T_\delta^{n-K}) \in G_k \) we have

\[
T_f = (T_\Lambda \otimes T_\delta^{n-K}) \ast G^n_x = ((T_\Lambda \ast G^n_x) \otimes T_\delta^{n-K})
\]

Let us denote \( a_\epsilon = a \ast G^n_x \) and \( b_\epsilon = b \ast G^n_x \). It is easy to see that

\[
f_\epsilon = (a_\epsilon(x_1:k)G^n_x(x_{k+1:n}))_U \in S(\mathbb{R}^n)
\]

From a well-known property of the Weierstrass transform we have:

\[
\|f_\epsilon\|_{L_1} = \|a_\epsilon\|_{L_1} \cdot \|G^n_x\|_{L_1} \leq \|a\|_{L_1}
\]

From this we obtain for any \( f = (T_\Lambda \otimes T_\delta^{n-K}) U, g = (T_\delta \otimes T_\delta^{n-k}) U \in G_k \):

\[
\|f_\epsilon\|_{L_1} \|g_\epsilon\|_{L_1} \leq \max_{x,y} \|M(x,y)\| \|a\|_{L_1} \|b\|_{L_1} < \infty
\]

Thus, \( \langle f | M(g) \rangle \) is defined and:

\[
\langle f_\epsilon | M(g_\epsilon) \rangle = \int_{\mathbb{R}^n \times \mathbb{R}^n} a^*_\epsilon(x_1:k)G^n_x(x_{k+1:n})M(U^T x, V^T y)b_\epsilon(y_1:k)G^n_x(y_{k+1:n}) dxdy =
\]

\[
= \int_{\mathbb{R}^k \times \mathbb{R}^k} a^*_\epsilon(x_1:k)M_e(x_1:k, y_1:k)b_\epsilon(y_1:k)dx_1:k dy_1:k
\]

where

\[
M_e(x_1:k, y_1:k) = \int_{\mathbb{R}^n-K \times \mathbb{R}^n-K} G^n_x(x_{k+1:n})M(U^T x, V^T y)G^n_x(y_{k+1:n}) dx_{k+1:n} dy_{k+1:n}
\]

Let \( U_k, V_k \in \mathbb{R}^{n \times n} \) be matrices that comprise the first \( k \) rows of \( U, V \) correspondingly and \( n-k \) zero rows below. Also, let \( L \) denote Lipschitz constant for \( M \) such that \( |M(x,y) - M(x',y')| \leq L(|x-x'| + |y-y'|) \). For the function \( M_e(x_1:k, y_1:k) \) we have:

\[
|M_e(x_1:k, y_1:k) - M(U_k^T x, V_k^T y)| =
\]

\[
\int_{\mathbb{R}^{2n-k}} G^n_x(x_{k+1:n}) (M(U^T x, V^T y) - M(U_k^T x, V_k^T y)) G^n_x(y_{k+1:n}) dx_{k+1:n} dy_{k+1:n}
\]
Obviously, \( \text{rank} f \) is the Gram matrix of the collection \( \{ (U - U_k)^T x \} \).

Under review as a conference paper at ICLR 2021

For a sequence \( \epsilon > 0 \) there exists bounded \( \tilde{M}(x_{1:k}, y_{1:k}) = M(U_1^T x, V_1^T y) \) such that:

\[
M(x_{1:k}, y_{1:k}) \xrightarrow{\epsilon \to 0} \tilde{M}(x_{1:k}, y_{1:k}) \text{ in } L_\infty(\mathbb{R}^{2k})
\]

Further we assume that \( \epsilon > 0 \) is small enough, so that \( M(x_{1:k}, y_{1:k}) \leq C = 2 \max |M| \). Now we have:

\[
| \langle f, M \rangle_{g_k} - \int_{\mathbb{R}^k \times \mathbb{R}^k} a^*(x_{1:k}) \tilde{M}(x_{1:k}, y_{1:k}) b(y_{1:k}) dx_{1:k} dy_{1:k} |
\]

\[
\leq \int_{\mathbb{R}^k \times \mathbb{R}^k} | (a^*(x_{1:k})) M(x_{1:k}, y_{1:k}) b(y_{1:k}) - a^*(x_{1:k}) \tilde{M}(x_{1:k}, y_{1:k}) b(y_{1:k}) | dx_{1:k} dy_{1:k} +
\]

\[
\int_{\mathbb{R}^k \times \mathbb{R}^k} M(x_{1:k}, y_{1:k}) a^*_e(x_{1:k}) (b_e(y_{1:k}) - b(y_{1:k})) dx_{1:k} dy_{1:k} +
\]

\[
\int_{\mathbb{R}^k \times \mathbb{R}^k} a^*(x_{1:k}) b(y_{1:k}) (M(x_{1:k}, y_{1:k}) - \tilde{M}(x_{1:k}, y_{1:k})) dx_{1:k} dy_{1:k} \leq
\]

\[
C \| a_e \|_{L_1} \| b_e - b \|_{L_1} + C \| b \|_{L_1} \| a_e - a \|_{L_1} + \| a^*(x_{1:k}) b(y_{1:k}) \|_{L_1} \| M - \tilde{M} \|_{L_\infty}
\]

It is well-known (e.g. see Theorem 2.25) that \( \| a_e - a \|_{L_p}, \| b_e - b \|_{L_p} \to 0, \| a_e \|_{L_1} \leq \| a \|_{L_1} \) and \( \| M - \tilde{M} \|_{L_\infty} \to 0 \). Thus, \( \epsilon \to 0 \) \( \langle f, M \rangle_{g_k} \) exists and \( \langle f, M \rangle_{g_k} \) is defined.

Let us now prove that rank \( M_f \) \( \leq k \). Since \( f \in G_k \), then \( f = (T_g \otimes \delta^{n-k}) U \) where \( U \) is an orthogonal matrix and \( U = [u_1, \ldots, u_n] \). It is easy to see that:

\[
\langle x_i, f \rangle M(x_j, f) = \langle (x_i)_{U^T} M(U^T x, U^T y) (x_j)_{U^T} \rangle =
\]

\[
\langle w_i^T x T_g \otimes \delta^{n-k} | M(U^T x, U^T y) | w_j^T x T_g \otimes \delta^{n-k} \rangle
\]

Let us now denote \( V = [u_1, \ldots, u_n] \in \mathbb{R}^{k \times n} \) a submatrix of \( U \) in which only first \( k \) rows of \( U \) are present. Then, the latter integral is equal to:

\[
\int_{\mathbb{R}^k \times \mathbb{R}^k} u_i^T x_{1:k} y_{1:k} u_j g(x_{1:k}) g(y_{1:k}) (V^T x_{1:k}, V^T y_{1:k}) dx_{1:k} dy_{1:k} = u_i^T B u_j
\]

where

\[
B = \{ [x_i | M' | x_j | g] \}_{1 \leq i, j \leq k}, \quad M'(x_{1:k}, y_{1:k}) = M(V^T x_{1:k}, V^T y_{1:k})
\]

is the Gram matrix of the collection \( \{ x_i g(x_{1:k}) \}_{i=1}^k \subseteq S(\mathbb{R}^k) \).

Obviously, \( \text{rank } M_f = \text{rank } [\text{Re } u_i^T B u_j]_{1 \leq i, j \leq n} = \text{rank } V^T (\text{Re } B) V \leq \text{rank } V = k \).

**D General Theory of the Reduction for Section 5**

For a sequence \( \{ f_s \}_{s \to \infty} \subseteq S'(\mathbb{R}^n) \), \( \text{Lim } f_s \) denotes a set of points \( f \in S'(\mathbb{R}^n) \), such that there exists a growing sequence \( \{ s_i \} \subseteq \mathbb{N} \) and \( \lim_{s \to \infty} f_{s_i} = f \).
D.1 Regular solutions and reduction theorems

For $I : G'_k \cup S(\mathbb{R}^n) \to \mathbb{R}^+$, it is natural to reduce the optimization task $\text{argmin}_{f \in G'_k \cup S(\mathbb{R}^n)} I$ to an optimization task over ordinary functions with a penalty term $\lambda R(f)$. To have an equivalence between $\text{argmin}_{f \in G'_k \cup S(\mathbb{R}^n)} I$ and $\text{argmin}_{f \in \mathbb{R}^n} I + \lambda R(f)$ we need to assume that $I$’s behaviour when approaching $f \in G'_k$ from a set $S(\mathbb{R}^n)$ is continuous, i.e. for any sequence $\{f_i\} \subseteq S(\mathbb{R}^n)$ such that $T_{f_i} \to f \in G'_k$, we have $\lim_{i \to \infty} I(T_{f_i}) = I(f)$. Let us introduce the notion of a regular solution both for $\text{argmin}_{f \in G'_k \cup S(\mathbb{R}^n)} I$ and $\text{argmin}_{f \in \mathbb{R}^n} I + \lambda R(f)$.

**Definition 4.** A sequence $\{f_i\} \subseteq S(\mathbb{R}^n)$ is said to solve $\text{argmin}_{f \in G'_k} I$ if:

$$I(f_i) + \lambda_i R(f_i) \leq \inf_{f \in S(\mathbb{R}^n)} I(f) + \lambda_i R(f) + \epsilon_i$$

where $\epsilon_i \to +0$ and $\lambda_i \to +\infty$, $i \to +\infty$. If, additionally, $\text{Tr}(M_{f_i})$ is bounded, then $\{f_i\} \subseteq S(\mathbb{R}^n)$ is said to solve $\text{argmin}_{f \in G'_k} I$ regularily.

Let us define $\text{rsol} (I(f), R(f)) = \bigcup_{\{f_i\} \text{ r. solves (11)}} \lim_{i \to \infty} T_{f_i}$.

**Theorem 8.** If $M$ is a proper kernel, then $\text{rsol} (I(f), R(f)) \subseteq \text{argmin}_{f \in G'_k} I(f)$.

**Theorem 9.** If $M$ is a proper kernel and $\text{rsol} (I(f), R(f)) \neq 0$, then $\text{argmin}_{f \in G'_k} I(f) \subseteq \text{rsol} (I(f), R(f))$.

**Theorem 10** (Reduction theorem). If $M$ is a proper kernel, $\text{argmin}_{f \in G'_k} I(f) \subseteq B_k$ and $\text{rsol} (I(f), R(f)) \neq 0$, then $\text{rsol} (I(f), R(f)) = \text{argmin}_{f \in G'_k} I(f)$.

Suppose that we now solve a sequence of problems $\text{argmin}_{f \in G'_k} I(f)$ and find $\{f_s\} \subseteq S(\mathbb{R}^n)$. According to Theorems 8 and 9, the following are potential scenarios:

1. $\text{Tr}(M_{f_s})$ blows up and the convergence is not guaranteed. This situation can be avoided by controlling $\text{Tr}(M_{f_s})$ in an optimization process. In practice, when $f$ has a parameterized form, this can be done by bounding parameters.

If $\text{Tr}(M_{f_s})$ does not blow up, we still have two subcases:

2.1 $\lim_{s \to \infty} T_{f_s} \neq 0$. This implies a positive outcome to approach $\text{argmin}_{f \in \mathbb{R}^n} I(f)$ to the optimization problem.

2.2 $\lim_{s \to \infty} T_{f_s} = 0$. This exotic situation can happen only if a sequence $T_{f_s}$ leaves any sequentially compact subset of $S'(\mathbb{R}^n)$. Bounding parameters also tackles this case.

D.2 Proofs of Theorem 8 and 9

For any $f = (T_l \otimes \delta^{n-k})_U \in G_k$ and $\sigma > 0$, let us define $f_\sigma$ as:

$$T_{f_\sigma} = (T_l \otimes \delta^{n-k})_U * G^\sigma_n = (T_{f_\sigma} \otimes T_{G^\sigma_{n-k}})_U$$

$$f_\sigma = (l_\sigma(x_{1:k})G^\sigma_{n-k}(x_{k+1:n}))_U$$
\[ l_\sigma = l \ast G_\sigma^k \]

We have \( T_{f_\sigma} \to^* (T_1 \otimes \delta^{n-k})_U \) as \( \sigma \to +0 \).

**Lemma 4.** For any \( f \in G_k \), \( \lim_{\sigma \to +0} \langle x_i f_\sigma | M | x_j f_\sigma \rangle = 0 \), for any \( (i, j) \notin \{1, \ldots, k\}^2 \), and \( \sup_{\sigma \in [0,1]} \langle x_i f_\sigma | M | x_j f_\sigma \rangle < \infty \), for any \( (i, j) \in \{1, \ldots, k\}^2 \).

**Proof.** W.l.o.g. we can assume that \( f = T_1 \otimes \delta^{n-k} \), \( l \in S(\mathbb{R}^k) \). If \( i > k \), \( j \leq k \) we have:

\[
\langle x_i f_\sigma | M | x_j f_\sigma \rangle = \frac{1}{(2\pi \sigma^2)^{n-k}} \int_{\mathbb{R}^n \times \mathbb{R}^n} x_i y_j e^{-\frac{|x_{i+1}+1|^2}{2\sigma^2}} l_\sigma(x_{1:k}) M(x, y)e^{-\frac{|y_{k+1}|^2}{2\sigma^2}} l_\sigma(y_{1:k}) dx dy =
\int_{\mathbb{R}^n} \frac{1}{\sqrt{2\pi \sigma^2}} x_i e^{-\frac{|x_{i+1}|^2}{2\sigma^2}} l_\sigma(x_{1:k}) P(x) dx
\]

where \( P(x) = \int_{\mathbb{R}^n} \frac{1}{\sqrt{2\pi \sigma^2}} y_j e^{-\frac{|y_{k+1}|^2}{2\sigma^2}} l_\sigma(y_{1:k}) dy \). Using the Hölder inequality we obtain:

\[
|\langle x_i f_\sigma | M | x_j f_\sigma \rangle| \leq ||\frac{1}{\sqrt{2\pi \sigma^2}} x_i e^{-\frac{|x_{i+1}|^2}{2\sigma^2}} l_\sigma(x_{1:k})||_{L_1(\mathbb{R}^n)} ||P||_{L_\infty(\mathbb{R}^n)} =
\]

Thus,

\[
|\langle x_i f_\sigma | M | x_j f_\sigma \rangle| \leq ||\frac{1}{\sqrt{2\pi \sigma^2}} x_i e^{-\frac{|x_{i+1}|^2}{2\sigma^2}} ||_{L_1(\mathbb{R}^n-k)} ||l_\sigma||_{L_1(\mathbb{R}^k)} ||y_j||_{L_1(\mathbb{R}^k)} = ||\frac{1}{\sqrt{2\pi \sigma^2}} x_i e^{-\frac{|x_{i+1}|^2}{2\sigma^2}} ||_{L_1(\mathbb{R}^n-k)} ||y_j||_{L_1(\mathbb{R}^k)}
\]

Using \( ||l_\sigma||_{L_1(\mathbb{R}^k)} \to ||l||_{L_1(\mathbb{R}^k)} \sigma \to +0 \), \( ||y_j||_{L_1(\mathbb{R}^k)} \to +0 \), we see the boundedness of \( ||l_\sigma||_{L_1(\mathbb{R}^k)} ||y_j||_{L_1(\mathbb{R}^k)} \) and proceed:

\[
\leq C ||\frac{1}{\sqrt{2\pi \sigma^2}} x_i e^{-\frac{|x_{i+1}|^2}{2\sigma^2}} ||_{L_1(\mathbb{R}^n-k)}
\]

It is easy to see that \( ||\frac{1}{\sqrt{2\pi \sigma^2}} x_i e^{-\frac{|x_{i+1}|^2}{2\sigma^2}} ||_{L_1(\mathbb{R}^n-k)} \to 0 \) as \( \sigma \to 0 \), therefore \( \langle x_i f_\sigma | M | x_j f_\sigma \rangle \to 0 \).

Similarly, we can prove that \( \langle x_i f_\sigma | M | x_j f_\sigma \rangle \to 0 \) if \( i, j > k \).

The entries of the main \( k \times k \) minor \( \langle x_i f_\sigma | M | x_j f_\sigma \rangle \) for \( 1 \leq i, j \leq k \) are bounded, because:

\[
\text{Tr} M_{f_\sigma} = \frac{1}{(2\pi \sigma^2)^{n-k}} \int_{\mathbb{R}^n \times \mathbb{R}^n} x \cdot y e^{-\frac{|x_{i+1}|^2}{2\sigma^2}} l_\sigma(x_{1:k}) M(x, y) e^{-\frac{|y_{k+1}|^2}{2\sigma^2}} l_\sigma(y_{1:k}) dx dy \leq
\]

\[
(2\pi \sigma^2)^{n-k} \int_{\mathbb{R}^n \times \mathbb{R}^n} (|x_{1:k} \cdot y_{1:k}| + |x_{k+1,n} \cdot y_{k+1,n}|) e^{-\frac{|x_{i+1}|^2}{2\sigma^2}} l_\sigma(x_{1:k}) l_\sigma(y_{1:k}) dx dy \leq
\gamma \int_{\mathbb{R}^n \times \mathbb{R}^n} |x_{1:k} \cdot y_{1:k}| l_\sigma(x_{1:k}) l_\sigma(y_{1:k}) dx dy + \gamma ||l_\sigma||_{L_1(n-k)}^2 \sigma^2 \leq
\gamma \sum_{j=1}^k ||y_j||_{L_1(\mathbb{R}^k)}^2 ||l_\sigma||_{L_1(n-k)}^2 + \gamma ||l_\sigma||_{L_1(\mathbb{R}^k)}^2 \sigma^2
\]

Again, using \( ||l_\sigma||_{L_1(\mathbb{R}^k)} \to ||l||_{L_1(\mathbb{R}^k)} \sigma \to +0 \), \( ||y_j||_{L_1(\mathbb{R}^k)} \to +0 \), \( ||l_\sigma||_{L_1(\mathbb{R}^k)} \sigma \to +0 \), we obtain the boundedness of RHS. \(\square\)
Corollary 1. For any \( f \in G_k \), \( \lim_{\sigma \to 0} R(f_{\sigma}) = 0 \).

Proof. W.l.o.g., we can assume that \( f = T_l \odot \delta^{n-k} \), \( l \in S(\mathbb{R}^k) \). By lemma, all entries of \( M_{f_{\sigma}} \) except those of the main \( k \times k \) minor approach 0 as \( \sigma \to 0 \). This means that

\[
\lim_{\sigma \to +0} Q(f_{\sigma}) = 0
\]

where \( Q(f_{\sigma}) = \sum_{i=k+1}^{n} \langle x_i f_{\sigma}, |M| x_i f_{\sigma} \rangle \). Let \( v_1, \cdots, v_n \) be unit eigenvectors of \( M_{f_{\sigma}} \) corresponding to the eigenvalues \( \lambda_1 \geq \cdots \geq \lambda_n \), \( P = \sum_{i=k+1}^{n} e_i e_i^T \), then

\[
R(f_{\sigma}) = \sum_{i=k+1}^{n} \lambda_i = \min_{p_i \in [0,1], \sum_i p_i = n-k} \sum_{i=1}^{n} \lambda_i p_i \leq \sum_{i=1}^{n} \lambda_i \text{Tr} (P v_i v_i^T P) = \text{Tr} (PM_{f_{\sigma}} P) = Q(f_{\sigma})
\]

Since \( R(f_{\sigma}) \leq Q(f_{\sigma}) \), we obtain \( \lim_{\sigma \to 0} R(f_{\sigma}) = 0 \). \( \square \)

D.2.1 Proof of Theorem 8

Proof. Suppose that a sequence \( \{f_i\}_{i=1}^{\infty} \subseteq S(\mathbb{R}^n) \) regularly solves (7) and \( T \in \text{Lim} f_i \). W.l.o.g., we can assume that \( T_{f_i} \to T \) and \( \text{Tr}(M_{f_i}) \) is bounded and \( I(f_i) + \lambda_i R(f_i) \leq \inf_{f \in S(\mathbb{R}^n)} I(f) + \lambda_i R(f) + \epsilon_i, \epsilon_i \to 0 \). Below we use continuity of \( I \) and corollary 1.

\[
\inf_{f \in S(\mathbb{R}^n)} I(f) + \lambda_i R(f) \leq \inf_{f \in G_k} \inf_{\sigma > 0} I(f_{\sigma}) + \lambda_i R(f_{\sigma}) \leq \inf_{f \in G_k} \lim_{\sigma \to +0} I(f_{\sigma}) + \lambda_i R(f_{\sigma}) \leq \inf_{f \in G_k} I(f)
\]

from which we conclude that \( \lambda_i R(f_i) \leq \inf_{f \in G_k} I(f) + \epsilon_i \) and, therefore, \( R(f_i) \to 0 \).

For each \( l \), let us define \( P_l \) as the projection operator to a subspace spanned by first principal components of the matrix \( \sqrt{M_{f_i}} \), i.e.

\[
P_l = \sum_{i=1}^{k} v_i v_i^T
\]

where \( v_1, \cdots, v_k \) are orthonormal eigenvectors that correspond to \( k \) largest eigenvalues of \( \sqrt{M_{f_i}} \). From the Eckart-Young-Mirsky theorem we see that \( R(f_i) = \| \sqrt{M_{f_i}} - P_l \sqrt{M_{f_i}} \|_F^2 \). Since a set of all projection operators \( \{P \in \mathbb{R}^{n \times n} | P^2 = P, P^T = P \} \) is a compact subset of \( \mathbb{R}^{n^2} \), one can always find a projection operator \( P = \sum_{i=1}^{k} v_i v_i^T \) and a growing subsequence \( \{f_i \} \) such that \( \|P_{f_i} - P\|_F \to 0 \) as \( s \to \infty \). Thus, for the subsequence \( \{f_i \} \) we have:

\[
\| \sqrt{M_{f_{f_i}}} - P \sqrt{M_{f_{f_i}}} \|_F = \| \sqrt{M_{f_{f_i}}} - P_l \sqrt{M_{f_{f_i}}} + P_l \sqrt{M_{f_{f_i}}} - P \sqrt{M_{f_{f_i}}} \|_F \leq \| \sqrt{M_{f_{f_i}}} - P_l \sqrt{M_{f_{f_i}}} \|_F + \| P_l - P \|_F \| \sqrt{M_{f_{f_i}}} \|_F = \sqrt{R(f_i)} + \| P_l - P \|_F \sqrt{\text{Tr}(M_{f_i})}
\]

and using the boundedness of \( \text{Tr}(M_{f_i}) \) we obtain \( \| \sqrt{M_{f_{f_i}}} - P \sqrt{M_{f_{f_i}}} \|_F \to 0 \).

Since \( \| \sqrt{M_{f_{f_i}}} - P \sqrt{M_{f_{f_i}}} \|_F \to 0 \), let us complete \( v_1, \cdots, v_k \) to an orthonormal basis \( v_1, \cdots, v_n \) and make the change of variables \( y_j = v_j^T x \). Let us denote \( V = [v_1, \cdots, v_n] \) and let \( V^T = [w_1, \cdots, w_n] \). Then, after that change of variables any function \( f(x) \) corresponds to \( f'(y) = f(V y) \) and the kernel \( M \) corresponds to \( M'(y, y') = M(V y, V y') \). If we apply that change of variables in the integral expression of \( \langle x, f | M | x_j f \rangle \), we will obtain:

\[
\langle x, f | M | x_j f \rangle = \langle w_i^T y f' | M' | w_j^T y f' \rangle = w_i^T \langle |y_i f' | M' | y_j f' \rangle w_j \Rightarrow \Re \langle x, f | M | x_j f \rangle = w_i^T \Re \langle |y_i f' | M' | y_j f' \rangle w_j
\]

I.e. \( M_f = VM_f' V^T \), or \( M_f' = V^T M_f V \). Note that \( P = V I^k V^T \) where \( I^k \) is a diagonal matrix whose main \( k \times k \) minor is the identity matrix, and all other entries are zeros. Using the fact that the Frobenius norm of orthogonally similar matrices are equal and the identity \( V^T \sqrt{M_{f_{f_i}}} V = \sqrt{V^T M_{f_{f_i}} V} \), we obtain:

\[
\| \sqrt{M_{f_{f_i}}} - P \sqrt{M_{f_{f_i}}} \|_F = \| V^T \sqrt{M_{f_{f_i}}} V - V^T P \sqrt{M_{f_{f_i}}} V \|_F =
\]
the change of variables we still have $f = L$, and, therefore, $y = L(x)$. Let us define $\psi$ then there exists a sequence $\sigma > 0$. Therefore, for any $\lambda$, $\psi = O(M')[\phi]$. Since $O(M')$ is a strictly positive self-adjoint operator, by the Cauchy-Schwarz inequality, we obtain:

$$\left| \langle y_i f_i^*, O(M')[\phi] \rangle \right| \leq \sqrt{\langle y_i f_i^* | O(M') | y_i f_i^* \rangle} \sqrt{\langle \phi, O(M')[\phi] \rangle}$$

Therefore, for any $\psi \in \mathcal{R}(O(M')) \cap S(\mathbb{R}^n)$ and $i > k$ we have $\lim_{i \to \infty} \langle y_i f_i^*, \psi \rangle = \lim_{i \to \infty} \langle f_i^*, y_i \rangle = 0$. Since $f_i^* \to^* T_V$ we obtain $\langle T_V, \psi \rangle = \langle y_i T_V, \psi \rangle = 0$ for any $\psi \in \mathcal{R}(O(M')) \cap S(\mathbb{R}^n)$. But the density of $\mathcal{R}(O(M')) \cap S(\mathbb{R}^n)$ in $S(\mathbb{R}^n)$ implies that $y_i T_V = 0$.

Using lemma 2 and $(T_V)_{V'} = T$ we obtain $T \in G_k'$. Thus, we proved that $T_{f_i} \to T \in G_k'$. Thus, since $I(f_i) \leq I(f_i) + \lambda_i R(f_i) \leq \inf_{f \in G_k'} I(f) + \epsilon_i$ and $I$ is continuous, we finally get that $I(T) \leq \inf_{f \in G_k'} I(f)$, i.e. $T \in \text{Arg min}_{f \in G_k'} I(f)$. 

D.2.2 Proof of Theorem 2

Proof. Suppose $f^* \in \text{Arg min}_{f \in G_k'} I(f) \cap B_k$, i.e. $f^* \in B_k$ and $I(f^*) = \min_{f \in B_k} I(f)$. Since $f^* \in B_k$, then there exists a sequence $\{s_i^*\} \subseteq G_k$ such that $T_{s_i^*} \to^* f^*$ and $\text{Tr } M_{s_i^*} < \infty$.

Let us define $s_i^* \in S(\mathbb{R}^n)$ as $T_{s_i^*} = T_{s_i^*} * G_{\alpha_i}$. Since $\lim_{i \to \infty} R(s_i^*) = 0$ (lemma 4), there exists $\sigma_i > 0$, such that $R(s_i^*) < \frac{1}{\sigma}$, whenever $0 < \sigma \leq \sigma_i$. Also, by definition, $\text{Tr } M_{s_i^*} = \lim_{i \to \infty} \text{Tr } M_{s_i^*}$. Therefore, there exists $\sigma_i > 0$, such that $\text{Tr } M_{s_i^*} < \text{Tr } M_{s_i^*} + 1$ whenever $0 < \sigma \leq \sigma_i$.

If we set $\sigma_i^* = \min\{\sigma_i, \sigma_i', \frac{1}{\sigma_i}\}$, then a sequence $\{s_i^*\} \subseteq S(\mathbb{R}^n)$ satisfies:

$$\lim_{i \to \infty} R(s_i^*) = 0$$

$$\text{Tr } M_{s_i^*} < \infty$$

and (using lemma 1)

$$T_{s_i^*} \to^* f^*$$

Due to the continuity of $I$ we have

$$\lim_{i \to \infty} I(s_i^*) = I(f^*)$$

Now we set $f_i = s_i^*$, $\lambda_i = \frac{1}{\sqrt{R(f_i)}}$ and we obtain the needed sequence:

$$\lim_{i \to \infty} I(f_i) = \lim_{i \to \infty} I(f_i) + \lambda_i R(f_i) = I(f^*)$$

where $\text{Tr } M_{f_i}$ is bounded. It remains to check that our sequence regularly solves (7), i.e. $\lim_{i \to \infty} \inf_{f \in S(\mathbb{R}^n)} I(f) + \lambda_i R(f_i) = I(f^*)$ (this will imply $\lim_{i \to \infty} \inf_{f \in S(\mathbb{R}^n)} I(f) + \lambda_i R(f_i) = 0$). The inequality in one direction is obvious,

$$\inf_{f \in S(\mathbb{R}^n)} I(f) + \lambda_i R(f_i) \leq \inf_{f \in G_{\sigma}} I(f_{\sigma}) + \lambda_i R(f_{\sigma}) \leq \inf_{f \in G_{\sigma}} \lim_{\sigma \to 0} I(f_{\sigma}) + \lambda_i R(f_{\sigma}) =$$
\[ \inf_{f \in \mathcal{B}} I(f) = I(f^*) \]

Let us prove the inverse inequality.

Since \( \text{rsol} (I(f), R(f)) \neq \emptyset \), there exists \( \{ \hat{f}_i \} \subseteq \mathcal{S}(\mathbb{R}^n) \) such that:
\[
I(\hat{f}_i) + \lambda_i R(\hat{f}_i) \leq \inf_{f \in \mathcal{S}(\mathbb{R}^n)} I(f) + \lambda_i R(f) + \epsilon_i, \quad \lim_{i \to +\infty} \lambda_i = +\infty, \quad \lim_{i \to +\infty} \epsilon_i = 0, \quad \text{Tr} M_{\hat{f}_i} < \infty
\]
and \( a = \lim_{i \to +\infty} T_{\hat{f}_i} \). From Theorem 5 we obtain \( a = \text{Arg \ min}_{f \in \mathcal{B}} I(f) \).

One can always find a subset \( \{ \hat{\lambda}_d_i \} \subseteq \{ \hat{\lambda}_i \} \) such that \( \hat{\lambda}_d_i < \lambda_i, \hat{\lambda}_d_i \to \infty \) and obtain:
\[
\inf_{f \in \mathcal{S}(\mathbb{R}^n)} I(f) + \lambda_i R(f) \geq \inf_{f \in \mathcal{S}(\mathbb{R}^n)} I(f) + \hat{\lambda}_d_i R(f) \geq I(\hat{f}_d_i) + \hat{\lambda}_d_i R(\hat{f}_d_i) - \epsilon_d_i \geq I(\hat{f}_d_i) - \epsilon_d_i
\]

Therefore,
\[
\lim_{i \to \infty} \inf_{f \in \mathcal{S}(\mathbb{R}^n)} I(f) + \lambda_i R(f) \geq \lim_{i \to \infty} I(\hat{f}_d_i) - \epsilon_d_i = I(a) = \inf_{f \in \mathcal{B}} I(f) = I(f^*)
\]

This proves that \( \{ f_i \} \) regularly solves (7) and \( \lim_{i \to \infty} f_i = f^* \) i.e. \( f^* \in \text{rsol} (I(f), R(f)) \). \( \square \)

## E PROOF OF THEOREM [5]

Again we will prove a more general statement.

**Theorem 11.** If \( M \) is a proper and a real-valued kernel, \( O(M) \) is bounded and \( \text{Tr} M_f < \infty \), then \( S_f \in \mathcal{B}(L_2^2(\mathbb{R}^n), \mathbb{R}^n) \) and \( S_f S_f^\dagger = M_f \). Moreover,
\[
R(f) = \min_{S \in \mathcal{B}(L_2^2(\mathbb{R}^n), \mathbb{R}^n), \text{rank } S \leq k} \| S f - S \|_2^2
\]
and the minimum is attained at \( S = P_f S_f \) where \( P_f = \sum_{i=1}^k u_i u_i^\dagger \) and \( \{ u_i \}^\dagger \) are unit eigenvectors of \( M_f \) corresponding to the \( k \) largest eigenvalues (counting multiplicities).

*Proof.** The boundedness of \( S_f \) follows from the Cauchy-Schwarz inequality:
\[
| S_f [\phi] |^2 = | \text{Re} \langle \sqrt{O(M)}[x_1 f], \phi \rangle |^2 \leq \langle \sqrt{O(M)}[x_1 f], \sqrt{O(M)}[x_1 f] \rangle \langle \phi, \phi \rangle = \langle x_1 f, O(M)[x_1 f] \rangle \langle \phi, \phi \rangle
\]
and therefore:
\[
\| S_f [\phi] \|_2^2 = \sum_{i=1}^n | S_f [\phi] |^2 \leq \text{Tr } M_f \| \phi \|_{L_2(\mathbb{R}^n)}^2
\]
I.e., we have checked that \( S_f \) is bounded.

By definition, \( S_f^\dagger : \mathbb{R}^n \to L_2^2(\mathbb{R}^n) \times L_2^2(\mathbb{R}^n) \) and \( (u, S_f[\phi_1, \phi_2]) = (S_f^\dagger[u], [\phi_1, \phi_2]) \) and \( \phi_1, \phi_2 \in L_2^2(\mathbb{R}^n) \times L_2^2(\mathbb{R}^n) \). Let us denote \( f_1 = \text{Re } f, f_2 = \text{Im } f \). It is easy to see that the following operator satisfies the latter identity:
\[
O[u] = [\sqrt{O(M)}[f_1(x)x^T u], \sqrt{O(M)}[f_2(x)x^T u]]
\]
Since the adjoint is unique, then \( S_f^\dagger = O \). Let us calculate \( S_f S_f^\dagger \):
\[
u \xrightarrow{S_f^\dagger} [\sqrt{O(M)}[f_1(x)x^T u], \sqrt{O(M)}[f_2(x)x^T u]] \xrightarrow{S_f} \sum_{i=1}^n [\langle x_1 f_1(x), \sqrt{O(M)}[\sqrt{O(M)}[f_1(x)x^T u]] \rangle + \langle x_1 f_2(x), \sqrt{O(M)}[\sqrt{O(M)}[f_2(x)x^T u]] \rangle] + \cdots + [\langle x_n f_1(x), \sqrt{O(M)}[\sqrt{O(M)}[f_1(x)x^T u]] \rangle + \langle x_n f_2(x), \sqrt{O(M)}[\sqrt{O(M)}[f_2(x)x^T u]] \rangle]
\]
Proof. Let $\sigma = \sqrt{\lambda_i}$ be the $i$-th eigenvalue of $M_f$. Then, we have

$$
\sum_{j=1}^{n} \langle x_j f_j(x), O(M)[f_j(x)x^T u] \rangle = |\langle x_i f, M[x_j f] \rangle_{1 \leq i,j \leq n} u = M_f u |
$$

Thus, $S_f S_f^\dagger = M_f$. Since $\text{Tr} S_f S_f^\dagger < \infty$ and $\|S_f^\dagger u\|^2 \leq \langle u, M_f u \rangle$, we obtain $S_f^\dagger$ is a bounded operator.

Let $u_1, \cdots, u_n$ be orthonormal eigenvectors of $M_f = S_f S_f^\dagger$ and $\lambda_1 \geq \cdots \geq \lambda_n > 0$ be corresponding nonzero eigenvalues. For $\sigma_i = \sqrt{\lambda_i}$ let us define $v_i = \frac{S_f^\dagger u_i}{\sigma_i}$. Vector $v_i$ corresponds to a pair of functions:

$$
v_i = \frac{1}{\sigma_i} \left[ \sqrt{O(M)[f_1(x)x^T u_i]}, \sqrt{O(M)[f_2(x)x^T u_i]} \right] \in L_2^0(\mathbb{R}^n) \times L_2^0(\mathbb{R}^n)
$$

It is easy to see that $v_1, \cdots, v_n$ is an orthonormal basis in $\text{Im} S_f^\dagger$, and $S_f^\dagger$ can be expanded in the following way:

$$
S_f^\dagger = \sum_{i=1}^{n'} \sigma_i v_i u_i^\dagger
$$

and therefore, SVD for $S_f$ is:

$$
S_f = \sum_{i=1}^{n'} \sigma_i u_i v_i^\dagger
$$

By the Eckart-Young-Mirsky theorem (see Theorem 4.4.7 from [Hsing & Eubank 2015]), an optimal $S$ in $\min_{S \in B(L_2^1(\mathbb{R}^n))} \|S - S\|_2^2$ is defined by a truncation of SVD for $S_f$ at $k$th term, i.e.:

$$
S = \sum_{i=1}^{k} \sigma_i u_i v_i^\dagger = P_f S_f
$$

where $P_f = \sum_{i=1}^{k} u_i u_i^\dagger$ is a projection operator to first $k$ principal components of $M_f$. Moreover, $\|S_f - P_f S_f\|^2 = \sum_{i=k+1}^{n'} \sigma_i^2 = \|M_f\|_{n-k} = R(f)$.

\hfill \Box

F THE ALTERNATING SCHEME IN THE DUAL SPACE FOR $M(x, y) = \zeta(x - y)$

When $M(x, y) = \zeta(x - y)$, the alternating scheme allows for a reformulation in the dual space. By this we mean that in Scheme 1 we substitute $\phi$ for the original $\phi_t$. If the primal Scheme 1 deals with operators $S_\phi, S_{\phi_{t-1}}$, the dual version deals with vectors of functions $\sqrt{\zeta \frac{\partial \hat{\phi}}{\partial x}} : \sqrt{\zeta \frac{\partial \hat{\phi}_{t-1}}{\partial x}}$. The substitution is based on the following simple fact:

**Theorem 12.** If $M(x, y) = \zeta(x - y)$, $\zeta, \hat{\zeta} \in C(\mathbb{R}^n)$ and $\forall \zeta i, \hat{\zeta} > 0$, then there exist constants $c_1$ and $c_2$ such that $\|S_\phi - P_{t-1} S_{\phi_{t-1}}\|_2 = c_1 \| \frac{\partial \hat{\phi}}{\partial x} - P_{t-1} \frac{\partial \hat{\phi}_{t-1}}{\partial x} \|_2 \| \zeta \|_{L_2,\zeta(\mathbb{R}^n)}$ and $\langle x_i f | M | x_j f \rangle = c_2 \frac{\partial f}{\partial x_i} \frac{\partial \hat{f}}{\partial x_j} \| \zeta \|_{L_2,\zeta(\mathbb{R}^n)}$

**Proof.** Let $f: \mathbb{R}^n \to \mathbb{C}$ be such that $\|x_i f\|_{L_2(\mathbb{R}^n)} < \infty$.

$$
O(M)[\psi] = \zeta * \psi \Rightarrow \mathcal{F} \{O(M)[\psi] \} \sim \sqrt{\hat{\zeta}} \hat{\psi} \Rightarrow \mathcal{F} \left\{ \sqrt{O(M)[\psi]} \right\} \sim \sqrt{\hat{\zeta}} \hat{\psi} \Rightarrow
$$

$$
S_f[\psi] = \text{Re} \langle x_i f, \sqrt{O(M)[\psi]} \rangle \sim \text{Re} \langle \mathcal{F} \{ x_i f \}, \mathcal{F} \left\{ \sqrt{O(M)[\psi]} \right\} \rangle \sim \text{Re} \left( i \frac{\partial \hat{f}}{\partial x_i} \sqrt{\hat{\zeta}} \hat{\psi} \right) = \text{Re} \left( i \sqrt{\hat{\zeta}} \frac{\partial \hat{f}}{\partial x_i} \hat{\psi} \right)
$$

22
Since $S_f[\psi]_i = \text{Re} \langle (S_f)_i, \psi \rangle \sim \text{Re} \langle \widehat{(S_f)}_i, \hat{\psi} \rangle$, we obtain

$$\widehat{(S_f)}_i = \kappa \sqrt{\xi} \frac{\partial \hat{\phi}}{\partial x_i}$$  \hfill (17)

where $\kappa$ is a constant.

Let us now introduce a vector of functions $V_f = [(S_f)_1, \cdots, (S_f)_n]^T \in L^2_\mathbb{R}^n$. Using (17) we obtain $\widehat{(S_f)}_i = \kappa \sqrt{\xi} \frac{\partial \hat{\phi}}{\partial x_i}$, and therefore:

$$\hat{V}_f = \kappa \sqrt{\xi} \frac{\partial \hat{\phi}}{\partial x}$$

Thus, the expression $||S_\phi - P_{t-1} S_{\phi_{t-1}}||_2^2$ in the alternating scheme can be rewritten as:

$$||V_\phi - P_{t-1} V_{\phi_{t-1}}||_{L^2(\mathbb{R}^n)}^2 \sim ||\kappa \sqrt{\xi} \frac{\partial \hat{\phi}}{\partial x} - P_{t-1} \kappa \sqrt{\xi} \frac{\partial \hat{\phi}_{t-1}}{\partial x}||_{L^2(\mathbb{R}^n)}^2 \sim

|| \frac{\partial \hat{\phi}}{\partial x} - \frac{\partial \hat{\phi}_{t-1}}{\partial x}||_2^2 ||_{L^2(\mathbb{R}^n)}^2$$

The matrix $M_f$ can also be calculated from $\hat{f}$ using the following identity:

$$\langle x_i f, M [x_j f] \rangle = \langle x_i f, \zeta * (x_j f) \rangle \sim \langle \frac{\partial \hat{f}}{\partial x_i}, \zeta \frac{\partial \hat{f}}{\partial x_j} \rangle = \langle \frac{\partial \hat{f}}{\partial x_i}, \frac{\partial \hat{f}}{\partial x_j} \rangle_{L^2,\zeta(\mathbb{R}^n)}$$

Let us introduce a function $\tilde{f}$ such that $\tilde{f}(f) = I(\hat{f})$. Then, we see that all steps in Scheme 1 can be performed with $\phi_t$ rather than with $\phi_\lambda$, using the algorithm 2.

Informally, the dual algorithm works as follows: at each iteration $t$ we compute a function $\hat{\phi}_t$ adapting it to data (the term $\hat{I}(\hat{\phi})$) and adapting its gradient field to the rank reduced gradient field of the previous $\phi_{t-1}$. For a sufficiently large $T$, it will converge and $\hat{\phi}_T \approx \hat{\phi}_{T-1}$. Then, the second term in the last step will be approximately equal to $\lambda \cdot ||\frac{\partial \hat{\phi}_T}{\partial x} - P_{T-1} \frac{\partial \hat{\phi}_{T-1}}{\partial x}||_2^2 ||_{L^2,\zeta(\mathbb{R}^n)}$, enforcing $\frac{\partial \hat{\phi}_T}{\partial x} \approx P_{T-1} \frac{\partial \hat{\phi}_{T-1}}{\partial x}$ for random $x \sim \frac{\zeta}{||\zeta||_{L^1}}$. Thus, gradients $\frac{\partial \hat{\phi}_T}{\partial x}$ lie in a $k$-dimensional subspace $\text{col}P_{T-1}$. This last property is a characteristic property of functions from $\mathcal{F}_k$.

Algorithm 2 The alternating scheme in the dual space.

$P_0 \leftarrow 0, \hat{\phi}_0 \leftarrow 0$

for $t = 1, \cdots, T$ do

$\hat{\phi}_t \leftarrow \text{arg min} \hat{I}(\hat{\phi}) + \lambda \cdot ||\frac{\partial \hat{\phi}}{\partial x} - P_{t-1} \frac{\partial \hat{\phi}_{t-1}}{\partial x}||_2^2 ||_{L^2,\zeta(\mathbb{R}^n)}$

Calculate $M_t = \left[ \text{Re} \langle \frac{\partial \hat{\phi}_t}{\partial x}, \frac{\partial \hat{\phi}_t}{\partial x} \rangle_{L^2,\zeta(\mathbb{R}^n)} \right]$

Find $\{v_i\}_1^n$ s.t. $M_t v_i = \lambda_i v_i, \lambda_1 \geq \cdots \geq \lambda_n$

$P_t \leftarrow \sum_{i=1}^k v_i v_i^T$

Output: $v_1, \cdots, v_k$

G A NUMERICAL ALTERNATING SCHEME FOR MMD

G.1 STRUCTURE OF $\mathcal{F}[P_k]$

From theorems 1 and 2, $\mathcal{F}[P_k] \subset \mathcal{F}_k^*$. In fact, a famous theorem of Bochner (1932) gives us that the Fourier transform of any positive finite Borel measure is a continuous positive definite function. That is, if $f \in \mathcal{F}[P]$, then for any distinct $y_1, \cdots, y_s \in \mathbb{R}^n$ the matrix $[f(y_i - y_j)]_{i,j=1}^s$ is
positive semidefinite. Since $\mu(\mathbb{R}^n) = 1$, we additionally have $f(0) = 1$. Let PDF denote the set of all continuous positive definite functions on $\mathbb{R}^n$ and

$$
\mathcal{M}_k = \{ f \in \text{PDF} | \exists \mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n, g : \mathbb{R}^k \rightarrow \mathbb{C} \text{ s.t. } f(\mathbf{x}) = g(\mathbf{v}_1^T \mathbf{x}, \ldots, \mathbf{v}_k^T \mathbf{x}), f(0) = 1 \} \quad (18)
$$

Thus, from Theorem 13 we obtain that the task 6 is equivalent to:

**Theorem 13.** $\mathcal{F}[\mathcal{P}_k] = \mathcal{M}_k$.

### G.2 The dual form of MMD

Let us define another gaussian kernel $\gamma(x) = e^{-\frac{\|x\|_2^2}{2}} = \mathcal{F}[k]$. Let $p_{\text{data}}(x)$ denote the characteristic function of the random vector $\mathbf{X}_{\text{data}} \sim \mu_{\text{data}}$. By definition, $p_{\text{data}}(x) = \mathbb{E}[e^{i \mathbf{X}_{\text{data}}^T \mathbf{x}}] = \frac{1}{N} \sum_{i=1}^{N} e^{i \mathbf{x}^T \mathbf{x}}$. Thus, $p_{\text{data}} \propto \mathcal{F}^{-1}[\mu_{\text{data}}]$ and $\mu_{\text{data}} \propto \mathcal{F}[p_{\text{data}}]$.

Using the isometry property of the Fourier transform for $L_2(\mathbb{R}^n)$ and the convolution theorem, we see that:

$$
d_{\text{MMD}}(\mu, \nu) = \|k * \mu - k * \nu\|_{L_2(\mathbb{R}^n)} \propto \|\gamma(x)(\mathcal{F}[\mu](x) - \mathcal{F}[\nu](x))\|_{L_2(\mathbb{R}^n)}
$$

Thus, from Theorem 13 we obtain that the task 6 is equivalent to:

$$
\|p_{\text{data}} - q\|_{L_2, \gamma(\mathbb{R}^n)} \rightarrow \min_{q \in \mathcal{M}_k} \quad (19)
$$

### G.3 Algorithms for MMD

Let $\Pi_k : \mathcal{G}_k \rightarrow \{1, +\infty\}$ and $\mathcal{M}_k : \mathcal{F}_k \rightarrow \{1, +\infty\}$ be simple penalty functions:

$$
\Pi_k(\phi) = 1, \text{ if } \phi \in \mathcal{P}_k \text{ and } \Pi_k(\phi) = \infty, \text{ otherwise}
$$

$$
\mathcal{M}_k(\phi) = 1, \text{ if } \phi \in \mathcal{M}_k \text{ and } \mathcal{M}_k(\phi) = \infty, \text{ otherwise}
$$

Then, the task 6 is equivalent to:

$$
I(\phi) = d_{\text{MMD}}^2(\mu_{\text{data}}, \phi) \Pi_k(\phi) \rightarrow \inf_{\phi \in \mathcal{G}_k}
$$

From the result of the previous section we see that if $I(\phi) = \tilde{I}(\hat{\phi})$, then:

$$
\tilde{I}(\hat{\phi}) = \|p_{\text{data}} - \hat{\phi}\|_{L_2, \gamma(\mathbb{R}^n)}^2 \mathcal{M}_k(\hat{\phi})
$$

Thus, the Algorithm 3 is an adaptation of Algorithm 2 to MMD.

**Algorithm 3** The alternating scheme in the dual space for MMD

\[
\begin{align*}
\hat{P}_0 & \leftarrow 0, q_0 \leftarrow 0 \\
\text{for } t = 1, \cdots, T \text{ do} & \\
1. q_t \leftarrow \arg \min_{q \in \mathcal{M}_k} \int_{\mathbb{R}^n} \gamma(x)^2|p_{\text{data}}(x) - q(x)|^2 dx + \lambda \int_{\mathbb{R}^n} \hat{\zeta}(x)||\frac{\partial q}{\partial x} - P_{t-1} \frac{\partial q_{t-1}}{\partial x}||^2 dx \\
2. \text{Calculate } M_t = \left[ \frac{\partial q}{\partial x_1}, \cdots, \frac{\partial q}{\partial x}\right]_{L_2, \gamma(\mathbb{R}^n)} \\
3. \text{Find } \{v_i\}_{i=1}^n \text{ s.t. } M_t v_i = \lambda_i v_i, \lambda_1 \geq \cdots \geq \lambda_n \\
4. P_t \leftarrow \sum_{i=1}^k v_i v_i^T \\
\text{Output: } \mathcal{L} = \text{span}(v_1, \cdots, v_k)
\end{align*}
\]

If the function $p_{\text{data}}$ is real-valued, then only real-valued functions can appear in the Algorithm 3. This assumption can be satisfied by adding reflections of initial points to the dataset (after it was centered).

At step 1, we search over $q$ given in the following parameterized form:

$$
q_\phi(x) = \sum_{i=1}^n \alpha_i \cos(\omega_i^T x) \quad (20)
$$
where \( \alpha_i > 0 \) and \( \sum_{i=1}^{m} \alpha_i = 1 \). In our implementation, we set \([\alpha_i]_{i=1}^{m} = \text{softmax}([u_i]_{i=1}^{m})\) and \( u_i \)'s are unconstrained. The number of neurons in a single layer neural network with a cosine activation function, \( nn \), is a hyperparameter. Let us denote parameters \( \{\omega_i, u_i\}_{i=1}^{m} \) by \( \theta \). It is easy to see the function \( q_0 \) is positive definite. Moreover, using Theorem 2 from Barron\cite{barron1993universal}, it can be shown that a set of all such functions, i.e. the convex hull of \( \{\cos(\omega^T x) | \omega \in \mathbb{R}^n\} \), is dense in a set of real-valued functions from \( M_k \). Though this parameterization is quite natural, finding architectures with more expressive power in a space of real-valued positive definite functions is an open problem.

Now, to minimize

\[
\Psi(\theta) = \int_{\mathbb{R}^n} \gamma(x)^2 |p_{\text{data}}(x) - q_0(x)|^2 d\mathbf{x} + \lambda \int_{\mathbb{R}^n} \hat{\zeta}(x)\{ \frac{\partial q_0}{\partial \mathbf{x}} \}^2 d\mathbf{x}
\]

with stochastic gradient descent methods (in our case, the Adam optimizer) we need to have an unbiased estimator of

\[
\nabla_\theta \Psi(\theta) \propto \mathbb{E}_{x \sim f} \nabla_\theta |p_{\text{data}}(x) - q_0(x)|^2 + \lambda \mathbb{E}_{z' \sim \hat{\zeta}} \nabla_\theta \| \frac{\partial q_0}{\partial \mathbf{x}}(z') - P_{t-1} \frac{\partial q_{\theta_{t-1}}}{\partial \mathbf{x}}(z') \|^2_2
\]

where \( z \sim f \) denotes that the random vector \( z \) is sampled according to the probability density function \( \frac{f(x)}{\int_{\mathbb{R}^n} f(x) dx} \). Thus, a natural estimator of the gradient is:

\[
\frac{1}{m} \sum_{i=1}^{m} \nabla_\theta |p_{\text{data}}(z_i) - q_0(z_i)|^2 + \frac{\lambda}{m} \sum_{i=1}^{m} \nabla_\theta \| \frac{\partial q_0}{\partial \mathbf{x}}(\xi_i) - P_{t-1} \frac{\partial q_{\theta_{t-1}}}{\partial \mathbf{x}}(\xi_i) \|^2_2
\]

where \( \{z_i\}_{i=1}^{m} \sim \text{iid} \gamma^2 \) and \( \{\xi_i\}_{i=1}^{m} \sim \text{iid} \hat{\zeta} \).

The last important issue with the practical numerical algorithm is the calculation of \( M_t \) at step 2. It is easy to see that:

\[
M_t = \mathbb{E}_{x \sim \hat{\zeta}} \left( \frac{\partial q_t}{\partial \mathbf{x}}(x) \frac{\partial q_t}{\partial \mathbf{x}}(x)^T \right)
\]

In practice we sample \( \chi_1, \cdots, \chi_l \sim \hat{\zeta} \) and estimate \( M_t \) as follows:

\[
M_t = \frac{1}{l} \sum_{i=1}^{l} \left( \frac{\partial q_t}{\partial \mathbf{x}}(\chi_i) \frac{\partial q_t}{\partial \mathbf{x}}(\chi_i)^T \right)
\]

The details of the numerical algorithm are given below\cite{4}. In all our experiments with MMD we set \( \hat{\zeta} = \gamma^2 \).

\textbf{Algorithm 4} The numerical algorithm for MMD. Hyperparameters: \( \hat{\lambda}, h, \sigma, m, t, \alpha, \beta_1, \beta_2, \text{nn} \).

\[
P_0 \leftarrow 0, \theta_0 \leftarrow 0 \\
\text{for } t = 1, \cdots, T \text{ do} \\
\quad \text{while } \theta \text{ has not converged do} \\
\quad \quad \text{Sample } \{z_i\}_{i=1}^{m} \sim \text{iid } \gamma^2 \\
\quad \quad \text{Sample } \{\xi_i\}_{i=1}^{m} \sim \text{iid } \hat{\zeta} \\
\quad \quad L \leftarrow \frac{1}{m} \sum_{i=1}^{m} \|p_{\text{data}}(z_i) - q_0(z_i)\|^2 + \frac{\hat{\lambda}}{m} \sum_{i=1}^{m} \| \frac{\partial q_0(\xi_i)}{\partial \mathbf{x}} - P_{t-1} \frac{\partial q_{\theta_{t-1}}(\xi_i)}{\partial \mathbf{x}} \|^2_2 \\
\quad \quad \theta_t \leftarrow \text{Adam}(\nabla_\theta L, \theta, \alpha, \beta_1, \beta_2) \\
\quad \text{Sample } \{\chi_i\}_{i=1}^{l} \sim \text{iid } \hat{\zeta} \\
\quad \text{Calculate } M_t = \left( \frac{\partial q_t}{\partial \mathbf{x}}(\chi_i) \frac{\partial q_t}{\partial \mathbf{x}}(\chi_i)^T \right) \\
\quad \text{Find } \{v_i\}_{i=1}^{l} \text{ s.t. } M_t v_i = \lambda_i v_i, \lambda_1 \geq \cdots \geq \lambda_n \\
\quad P_t \leftarrow \sum_{i=1}^{l} v_i v_i^T \\
\quad \text{Output: } v_1, \cdots, v_k
\]
H A NUMERICAL ALTERNATING SCHEME FOR HM

H.1 THE DUAL FORM OF HM

Due to a well-known relationship between moments of the probability measure μ and its characteristic function ρ, i.e. \( i^s \mu_{i_1 \ldots i_s} = \frac{\partial^s \rho(0)}{\partial x_{i_1} \ldots \partial x_{i_s}} \), the task 7 is equivalent to:

\[
\sum_{s=1}^{4} \frac{\lambda_s}{n^s} \sum_{1 \leq i_1, \ldots, i_s \leq n} |\frac{\partial^s p_{\text{data}}(0)}{\partial x_{i_1} \ldots \partial x_{i_s}} - \frac{\partial^s q(0)}{\partial x_{i_1} \ldots \partial x_{i_s}}|^2 \rightarrow \min_{q \in \mathcal{M}_k} \tag{21}
\]

Note that the maximum mean discrepancy distance and the distance based on higher moments are substantially different. Indeed, even if we set \( h \) as a large value (which makes \( \frac{\lambda}{h} \approx 0 \)), the MMD distance, unlike the HM distance, neglects higher order derivatives of the characteristic functions in the neighbourhood of the origin. Moreover, from the dual form \( \tilde{I} \), it is clear that \( d_{\text{HM}}(\mu_{\text{data}}, \nu) \) is a degenerate case of a weighted Sobolev norm between characteristic functions of \( \mu_{\text{data}} \) and \( \nu \).

H.2 ALGORITHMS FOR HM

Analogously to the case of MMD we see that the task 7 is equivalent to:

\[
I(\phi) = d_{\text{HM}}(\mu_{\text{data}}, \phi)^2 \Pi_k(\phi) \rightarrow \inf_{\phi \in \mathcal{G}_k} I(\phi)
\]

and

\[
\tilde{I}(\phi) = \sum_{s=1}^{4} \frac{\lambda_s}{n^s} \sum_{1 \leq i_1, \ldots, i_s \leq n} |\frac{\partial^s p_{\text{data}}(0)}{\partial x_{i_1} \ldots \partial x_{i_s}} - \frac{\partial^s \phi(0)}{\partial x_{i_1} \ldots \partial x_{i_s}}|^2 \mathcal{M}_k(\phi)
\]

Thus, the Algorithm 5 is an adaptation of Algorithm 2 to HM.

**Algorithm 5** The alternating scheme in the dual space for HM

\[
\begin{align*}
P_0 & \leftarrow 0, \ q_0 \leftarrow 0 \\
& \text{for } t = 1, \ldots, T \text{ do} \\
& \quad q_t \leftarrow \arg \min_{q \in \mathcal{M}_k} \sum_{s=1}^{4} \frac{\lambda_s}{n^s} \sum_{1 \leq i_1, \ldots, i_s \leq n} |\frac{\partial^s p_{\text{data}}(0)}{\partial x_{i_1} \ldots \partial x_{i_s}} - \frac{\partial^s q(0)}{\partial x_{i_1} \ldots \partial x_{i_s}}|^2 + \lambda \int_\mathbb{R}^n |\phi(x)||\frac{\partial \phi}{\partial x} - P_{t-1} \frac{\partial q_{t-1}}{\partial x}|^2 dx \\
& \quad 2 \text{ Calculate } M_k = \left[ \frac{\partial q_t}{\partial x_{i_1}}, \frac{\partial q_t}{\partial x_{i_2}} \right]_{L_2(\mathbb{R}^n)} \\
& \quad 3 \text{ Find } \{v_i\}_1^n \text{ s.t. } M_k v_i = \lambda_1 v_i, \lambda_1 \geq \cdots \geq \lambda_n \\
& \quad 4 \ P_t \leftarrow \sum_{i=1}^k v_i v_i^T \\
& \text{Output: } \mathcal{L} = \text{span}(v_1, \ldots, v_k)
\end{align*}
\]

Again, as in a numerical algorithm for MMD, at step 1, we search over \( q \) given in the form \( \Phi(\theta) = \sum_{s=1}^{4} \lambda_s E_{i_1, \ldots, i_s \sim \mu(1, n)} \frac{\partial^s (p_{\text{data}} - q)(0)}{\partial x_{i_1} \ldots \partial x_{i_s}} |^2 + \lambda E_{z' \sim \xi} \| \frac{\partial \phi}{\partial x}(z') - P_{t-1} \frac{\partial q_{t-1}}{\partial x}(z') \|^2_2 \]

where \( \mu(1, n) \) is the discrete uniform distribution over \( \{1, \ldots, n\} \). To apply the stochastic gradient descent methods we need to have an unbiased estimator of \( \nabla_\theta \Phi(\theta) \) which is equal to:

\[
\sum_{s=1}^{4} \lambda_s E_{i_1, \ldots, i_s \sim \mu(1, n)} \nabla \theta |\frac{\partial^s (p_{\text{data}} - q)(0)}{\partial x_{i_1} \ldots \partial x_{i_s}}|^2 + \lambda E_{z' \sim \xi} \| \frac{\partial \phi}{\partial x}(z') - P_{t-1} \frac{\partial q_{t-1}}{\partial x}(z') \|^2_2
\]

Thus, a natural estimator of the gradient is:

\[
\sum_{s=1}^{4} \lambda_s \sum_{i=1}^{m_1} \nabla \theta |\frac{\partial^s (p_{\text{data}} - q)(0)}{\partial x_{a[i,1]} \ldots \partial x_{a[i,s]}}|^2 + \lambda \sum_{i=1}^{m_2} \nabla \theta |\frac{\partial \phi(\xi_i)}{\partial x}|^2 - P_{t-1} \frac{\partial q_{t-1}}{\partial x}(\xi_i) \|^2_2
\]

where \( \{a[i,s,j]\}_{s,i=1,m_1,j=1,s} \sim \mu(1, n) \) and \( \{\xi_i\}_{i=1}^{m_2} \sim \xi \). Overall, we obtain the following Algorithm.
Algorithm 6 The numerical algorithm for HM. Hyperparameters: \( \lambda, \{ \lambda_i \}_{i=1}^{m_1}, m_2, l, \alpha, \beta_1, \beta_2, \text{nn.} \)

\[
P_0 \leftarrow 0, \theta_0 \leftarrow 0 \\
\text{for } t = 1, \ldots, T \text{ do} \\
\quad \text{while } \theta \text{ has not converged do} \\
\quad \quad \text{Sample } \{a[s, i, j]\}_{s=1, i=1, m_1, j=1}^{m_2} \sim_{iid} U(1, n) \\
\quad \quad \text{Sample } \{\xi_{i,j}\}_{i=1}^{m_2} \sim_{iid} \zeta \\
\quad \quad L \leftarrow \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \nabla_{\theta} \left( \frac{\partial^r (p_{\text{data}} - q_{\theta})}{\partial x_{\theta}(x_{i,j})} \right)^2 + \frac{1}{m_2} \sum_{i=1}^{m_2} \nabla_{\theta} \left\| \frac{\partial q_{\theta}}{\partial x_{\theta}} \right\|^2 \\
\quad \quad P_{t-1} \leftarrow \text{Adagrad}^\theta (\nabla L, \theta, \alpha, \beta_1, \beta_2) \\
\quad \quad \theta_t \leftarrow \theta_t \leftarrow \theta \\
\quad \quad \text{Sample } \{\chi_{i,j}\}_{i=1}^{m_2} \sim_{iid} \zeta \\
\quad \quad \text{Calculate } M_t = \frac{1}{t} \sum_{i=1}^{m_2} \frac{\partial q_{\theta}}{\partial x_{\theta}} \frac{\partial p_{\theta}}{\partial x_{\theta}}^T \\
\quad \quad \text{Find } \{v_i\}_{i=1}^{m_2} \text{ s.t. } M_t v_i = \lambda_i v_i, \lambda_1 \geq \cdots \geq \lambda_n \\
\quad \quad P_t \leftarrow \sum_{i=1}^{m_2} v_i v_i^T \\
\text{Output: } v_1, \ldots, v_k
\]

I A NUMERICAL ALTERNATING SCHEME FOR WD

By Theorem 6 the task 13 is equivalent to \( \min_{\mu \in \mathcal{P}_k} W(\mu, \mu_{\text{data}}) \), or to the following task:

\[
I(\phi) \rightarrow \inf_{\phi \in \mathcal{F}_k} \left\| \phi - \phi_{\text{data}} \right\|
\]

where \( I(T_{\phi}) = W(\mu, \mu_{\text{data}}) \) if \( \mu \in \mathcal{P}_k \) and \( I(\phi) = \infty \), if otherwise. The alternating scheme \( \mathcal{F}_k \) is designed to solve the penalty form of the problem, i.e.

\[
I(\phi) + \lambda R(\phi) \rightarrow \min_{\phi \in \mathcal{S}(\mathbb{R}^n)}
\]

which is equivalent to

\[
W(\phi, \mu_{\text{data}}) + \lambda R(\phi) \rightarrow \min_{\phi \in \mathcal{S}(\mathbb{R}^n)}
\]

where \( \mathcal{S}_p(\mathbb{R}^n) \subseteq \mathcal{S}(\mathbb{R}^n) \) is a set of Schwartz functions that can serve as pdf: \( \phi(x) \geq 0, \int_{\mathbb{R}^n} \phi(x) dx = 1 \). A numerical version of the alternating scheme requires additional specifications on: a) how to minimize over \( \phi \) at step 1, and b) how to estimate \( M_{\phi_t} \).

I.1 HOW TO MINIMIZE OVER \( \phi \)?

In the case of WD, the minimization step of the alternating scheme makes the following:

\[
\phi_t \leftarrow \arg \min_{\phi \in \mathcal{S}_p(\mathbb{R}^n)} W(\phi, \mu_{\text{data}}) + \lambda \left\| \phi - P_{t-1} S_{\phi_{t-1}} \right\|^2
\] (22)

where \( S_f = \sqrt{O(M)} |xf(x)| \).

For a numerical implementation of that step we need to choose some family of functions that is dense in \( \mathcal{S}_p(\mathbb{R}^n) \) (or, rich enough to approach the solution \( \mu^* \)). Following the tradition of GAN research let us assume that the family is given in the following form:

\[
\mathcal{H} = \{ \phi_\theta | \phi_\theta(x) \text{ is pdf of random vector } g_\theta(z), z \sim p(z), \theta \in \Theta \}
\] (23)

where \( \{ g_\theta | \theta \in \Theta \} \) is a parameterized family of smooth functions (usually, a neural network) and \( p(z) \) is some fixed distribution (usually, the gaussian distribution). Following [Arjovsky et al., 2017], we make the assumption \( \mathcal{H}_{\epsilon} \) in a numerical algorithm we need an access to a procedure that samples according to \( \phi_{\theta}(x) \), not the function itself.

Assumption 1. \( \| g_{\theta}(x') - g_{\theta}(z) \| \leq L(\theta, z)(| |\theta' - \theta| | + | |x' - z| |) \) where

\[
\mathbb{P}_{z \sim p(z)} L(\theta, z) < +\infty
\]

\[2\text{If } \mathcal{H} \subseteq \mathcal{S}(\mathbb{R}^n) \text{ is not satisfied, then we can choose } \mathcal{H}_{\epsilon} = \{ \phi_{\theta} \ast G_{\epsilon}^{\theta} | \theta \in \Theta \} \text{ for a very small } \epsilon.\]
Thus, instead of solving \(22\) we solve:
\[
\phi_t \leftarrow \arg\min_{\phi \in \mathcal{H}} W(\phi, \mu_{\text{data}}) + \lambda ||S_\phi - P_{t-1}S_{\phi_{t-1}}||^2
\]

taking into account that \(\phi_{t-1} \in \mathcal{H}\).

The Kantorovich-Rubinstein duality theorem gives us that:
\[
W(\phi_\theta, \mu_{\text{data}}) = \max_{f: ||f||_1 \leq 1} \mathbb{E}_{x \sim \mu_{\text{data}}} [f(x)] - \mathbb{E}_{z \sim p}(f(\phi(z))]
\]
which turns \(22\) into the following minimax task:
\[
\phi_t \leftarrow \arg\min_{\phi \in \mathcal{H}} \max_{f: ||f||_1 \leq 1} \mathbb{E}_{x \sim \mu_{\text{data}}} [f(x)] - \mathbb{E}_{z \sim p}(f(g_\phi(z)))] + \lambda ||S_\phi - P_{t-1}S_{\phi_{t-1}}||^2 \tag{24}
\]

In practice, we choose a family of functions \(\mathcal{L} = \{f_w | w \in \mathcal{W}\}\) and internal maximization is made over \(w \in \mathcal{W}\) with an additional penalty term that penalizes a violation of the Lipschitz condition: \(\forall x : ||f_x|| \leq 1\).

A family of minimax algorithms for the minimization of \(W(\phi_\theta, \mu_{\text{emp}})\) was developed in a series of papers [Arjovsky et al. (2017); Gulrajani et al. (2017); Wei et al. (2018)]. The standard minimax scheme that gained popularity in GAN literature iterates two steps: a) \(n\) times make a gradient ascent over \(w \in \mathcal{W}\), b) make a gradient descent over \(\theta\). The task \(24\) can be viewed as a Wasserstein GAN with an additional regularization term \(\lambda T(\theta)\) where \(T(\theta) = ||S_\phi - P_{t-1}S_{\phi_{t-1}}||^2\). To adapt these algorithms to the minimization of our function, we only need to have an unbiased estimator of the gradient \(\frac{\partial T}{\partial \theta}\). This estimator is needed for the generator to make its gradient descent step. The discriminator’s part of the algorithm (in which we maximize over Lipschitz functions \(f_w\)) can be set in a standard fashion — we choose Petzka et al. (2018)’s version, in which the term \(\max\{0, \frac{\partial f_w}{\partial \phi}(\xi x + (1 - \xi)g_\phi(z))\} \geq 1\) enforces Lipschitz condition (see step (**)) of the Algorithm [7].

### Algorithm 7 Numerical algorithm for WD. We use \(M(x, y) = e^{-\frac{|x-y|^2}{2\alpha}}\) and default values of \(\lambda = 10, \Lambda = 100, n_{\text{critic}} = 5, m = 40, l = 10000n, \alpha = 0.00001, \beta_1 = 0.5, \beta_2 = 0.9\)

\[
P_0 \leftarrow 0, \theta_0 \leftarrow 0
\]

for \(t = 1, \ldots, T\) do
  Minimax realization of \(\min_\theta W(\phi_\theta, \mu_{\text{emp}}) + \lambda T(\theta)\) (**):
    while \(\theta\) has not converged do
      for \(s = 1, \ldots, n_{\text{critic}}\) do
        Discriminator updates \(w\)
        Sample \(\{z_i\}_{i=1}^m, \{z_i'\}_{i=1}^m \sim p(z)\)
        \(\phi \leftarrow -\frac{1}{m} \sum_{i=1}^m f_w(g_\phi(z_i)) + \frac{1}{m} \sum_{i=1}^m \Xi(\theta, z_i, z_i')\) (\(\Xi\) is defined in equation \(25\))
        \(\theta \leftarrow \text{Adan}(\nabla \theta L, \theta, \alpha, \beta_1, \beta_2)\)
    \(\theta_t \leftarrow \theta\)
    Realization of step (**): \(g_\theta(z)\)
    Sample \(\{z_i\}_{i=1}^m, \{z_i'\}_{i=1}^m \sim p(z)\)
    \(M_t \leftarrow \sum_{i,j} g_\theta(z_i)g_\theta(z_i')^T M(g_\theta(z_i), g_\theta(z_i'))\)
    Find \(\lambda_i \geq \cdots \lambda_n\) s.t. \(\sum_i \lambda_i v_i = \lambda_1 v_1, \lambda_1 \geq \cdots \geq \lambda_n\)
    \(P_t \leftarrow \sum_{i=1}^k v_i v_i^T\)
  Output: \(v_1, \ldots, v_k\)

I.2 How to estimate \(\frac{\partial T}{\partial \theta}\) and \(M_{\phi_\theta}\) ?

Another important aspect of the numerical algorithm is the complexity of estimating the matrix \(M_{\phi_\theta}\) at step (**). The following theorem shows that we only need to sample \(z \sim p(z)\) a sufficient number of times to estimate \(\frac{\partial T}{\partial \theta}\) and \(M_{\phi_\theta}\).

**Theorem 14.** If \(\phi_\theta\) is pdf of the random vector \(g_\theta(z), z \sim p(z)\), then
Using lemma 5 we have:

\[
\frac{\partial T}{\partial \theta} = E_{z,z' \sim \rho} \frac{\partial \Xi(\theta, z, z')}{\partial \theta}
\]

where

\[
\Xi(\theta, z, z') = (g_\theta(z) \cdot g_\theta(z'))M(g_\theta(z), g_\theta(z')) - 2(g_\theta(z) \cdot P_{t-1}g_{\theta_{t-1}}(z'))M(g_\theta(z), g_{\theta_{t-1}}(z'))
\]

and RHS is well-defined.

I.2.1 Definition of \( \mathcal{H} \)

Specifically, for robust PCA/outlier pursuit applications, we define \( \phi_\theta(x) \) as a probability density function of the random vector \( \mathbf{a} + \mathbf{b} \), where \( \mathbf{a}, \mathbf{b} \) are independent and \( \mathbf{a} \) is the \( i \)-th column of matrix \( \theta_1 \in \mathbb{R}^{n \times N} \) (where \( i \sim U(1, N) \) is sampled uniformly from \( \{1, \ldots, N\} \)), \( \mathbf{b} = g_{\theta_2}(\mathbf{c}) \), \( \mathbf{c} \sim \mathcal{N}(0, I_n) \) and \( g_{\theta_2} : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is a neural network with weights \( \theta_2 \). Thus, \( \theta = (\theta_1, \theta_2) \). It can be checked that \( \mathcal{H}_t \), defined in this way, satisfies the Assumption 1. We specifically introduce the random vector \( \mathbf{a} \) here because, according to Theorem 6, the ultimate solution of the problem corresponds to \( \theta_1 = Y \) and \( \mathbf{b} = 0 \). This guarantees that the solution is approachable from set \( \mathcal{H} \).

I.3 Proof of theorem

We need to following lemma.

**Lemma 5.** \( ||S_\phi - PS_\psi||^2 = E_{x,y \sim \phi}(x \cdot y)M(x, y) + E_{x,y \sim \psi}(x \cdot Py)M(x, y) - 2E_{x \sim \phi, y \sim \psi}(x \cdot Py)M(x, y) \)

**Proof of lemma.**

\[
||S_\phi - PS_\psi||^2 = ||\sqrt{O(M)}[x\phi(x)] - P\sqrt{O(M)}[x\psi(x)]||^2 = \\
= ||\sqrt{O(M)}[x\phi(x) - Px\psi(x)]||^2 = \sum_{i=1}^{n} ||\sqrt{O(M)}[x_i\phi(x) - (Px)_i\psi(x)]||^2 = \\
\sum_{i=1}^{n} (x_i\phi(x)[O(M)x_i\phi(x)]) + (Px)_i\psi(x)[O(M)(Px)_i\psi(x)]) - 2((Px)_i\psi(x)[O(M)x_i\phi(x)]) = \\
E_{x,y \sim \phi}(x \cdot y)M(x, y) + E_{x,y \sim \psi}(x \cdot Py)M(x, y) - 2E_{x \sim \phi, y \sim \psi}(x \cdot Py)M(x, y)
\]

**Proof of theorem** Using lemma 5 we have:

\[
T(\theta) = E_{x,y \sim \phi_\theta}(x \cdot y)M(x, y) + E_{x,y \sim \phi_{t-1}}(x \cdot P_{t-1}y)M(x, y) - \\
2E_{x \sim \phi_\theta, y \sim \phi_{t-1}}(x \cdot P_{t-1}y)M(x, y) = \\
E_{z,z' \sim \rho}(g_\theta(z) \cdot g_\theta(z'))M(g_\theta(z), g_\theta(z')) + \\
E_{z,z' \sim \rho}(g_{\theta_{t-1}}(z) \cdot P_{t-1}g_{\theta_{t-1}}(z'))M(g_{\theta_{t-1}}(z), g_{\theta_{t-1}}(z')) - \\
2E_{z,z' \sim \rho}(g_\theta(z) \cdot P_{t-1}g_{\theta_{t-1}}(z'))M(g_\theta(z), g_{\theta_{t-1}}(z'))
\]

The second term does not depend on \( \theta \). Therefore,

\[
\frac{\partial T}{\partial \theta} = \frac{\partial}{\partial \theta} E_{z,z' \sim \rho} \Xi(\theta, z, z')
\]

where

\[
\Xi(\theta, z, z') = (g_\theta(z) \cdot g_\theta(z'))M(g_\theta(z), g_\theta(z')) - 2(g_\theta(z) \cdot P_{t-1}g_{\theta_{t-1}}(z'))M(g_\theta(z), g_{\theta_{t-1}}(z'))
\]
According to Algorithm 2, the next goal is to estimate \( M \). The gradient of the function \( \Phi(z) \) becomes:

\[
\frac{\partial}{\partial \theta} \mathbb{E}_{x,z} \Phi(\theta, z) = \mathbb{E}_{x,z} \frac{\partial \Phi(\theta, z)}{\partial \theta}
\]

The fact that \( M_{\phi_0} = \mathbb{E}_{x,z} \Phi(\theta, z) \) is obvious from the definition \( M_{\phi_0} = \mathbb{E}_{x,y} \Phi(\theta, x) \).

### J A Numerical Alternating Scheme for SDR

For a binary classification case, given a labeled dataset \( \{(x_i, y_i)\}_{i=1}^N \), \( x_i \in \mathbb{R}^n \), \( y_i \in \{0, 1\} \) we formulate the sufficient dimension reduction problem as the minimization task:

\[
J(f) = \mathbb{E}_{(z, c) \sim \mu_{data}, \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)} L(c, f(z + \epsilon)) \rightarrow \min_{f \in \mathcal{F}}
\]

where \( L(c, y) = -c \log(y) - (1 - c) \log(1 - y) \).

We apply the alternating scheme in the dual space (Algorithm 2) to this task. We set \( M(x, y) = \xi(x - y) \), where \( \xi \) is a strictly positive probability density function. A numerical version of the scheme is given below (Algorithm 8).

At every iteration \( t = 1, \ldots, T \) of the Algorithm 2 we solve the task (in our case \( \hat{I} = J \)):

\[
\hat{\phi}_t \leftarrow \arg \min_{\hat{\phi} \in \mathcal{P}} \hat{I}(\hat{\phi}) + \lambda \left\| \frac{\partial \hat{\phi}}{\partial x} - \hat{P}_{t-1} \frac{\partial \hat{\phi}_{t-1}}{\partial x} \right\|_{L_2, \hat{\zeta}(\mathbb{R}^n)}^2
\]

In a numerical version of the algorithm we assume that \( \hat{\phi} \) is given as a neural network \( f_\theta \), i.e. our task becomes:

\[
\theta_t \leftarrow \arg \min_{\theta} J(f_\theta) + \lambda \mathbb{E}_{\xi \sim \hat{\zeta}} \left\| \frac{\partial f_\theta}{\partial x}(\xi) - \hat{P}_{t-1} \frac{\partial f_{\theta_{t-1}}}{\partial x}(\xi) \right\|_{L_2, \hat{\zeta}(\mathbb{R}^n)}^2
\]

The gradient of the function \( \Phi(\theta) = J(f_\theta) + \lambda \mathbb{E}_{\xi \sim \hat{\zeta}} \left\| \frac{\partial f_\theta}{\partial x}(\xi) - \hat{P}_{t-1} \frac{\partial f_{\theta_{t-1}}}{\partial x}(\xi) \right\|_{L_2, \hat{\zeta}(\mathbb{R}^n)}^2 \) equals:

\[
\frac{\partial \Phi(\theta)}{\partial \theta} = \mathbb{E}_{(z, c) \sim P_{data}, \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)} \frac{\partial}{\partial \theta} L(c, f_\theta(z + \epsilon)) + \lambda \mathbb{E}_{\xi \sim \hat{\zeta}} \frac{\partial}{\partial \theta} \left[ \frac{\partial f_\theta}{\partial x}(\xi) - \hat{P}_{t-1} \frac{\partial f_{\theta_{t-1}}}{\partial x}(\xi) \right] \]

That is why \( \nabla_\theta L \) (given to Adam optimizer in the gradient descent loop) in the Algorithm 8 is an unbiased estimator of \( \frac{\partial \Phi(\theta)}{\partial \theta} \). Thus, in the “while loop” we find optimal \( \hat{\phi}_t = f_{\theta_t} \).

According to Algorithm 2, the next goal is to estimate \( M_t = \left[ \text{Re} \left( \frac{\partial \hat{\phi}_t}{\partial x}, \frac{\partial \hat{\phi}_t}{\partial \hat{z}} \right) \right]_{L_2, \hat{\zeta}(\mathbb{R}^n)} \). It is easy to see that

\[
M_t = \mathbb{E}_{x \sim \hat{\zeta}} \frac{\partial \hat{\phi}_t}{\partial x}(x) \frac{\partial \hat{\phi}_t}{\partial \hat{z}}(x) = \mathbb{E}_{x \sim \hat{\zeta}} \frac{\partial f_{\theta_t}}{\partial x}(x) \frac{\partial f_{\theta_t}}{\partial \hat{z}}(x)\]

From the last we see that the matrix \( M_t \) can be estimated by sampling \( x \sim \hat{\zeta} \) a sufficient number of times (the parameter \( l \) in our algorithm). All the rest is identical to Algorithm 2.

The regression version of the algorithm can be obtained by setting \( L(c, c') = (c - c')^2 \). Implementations for different databases can be found at [github].
Algorithm 8 The numerical alternating scheme for SDR. We use $\nu = 1.0$, $\hat{\zeta}(\mathbf{x}) = G_{0,8}(\mathbf{x})$ and default values of $\tilde{\lambda} = 10$, $m \approx 50$, $m' = 100$, $l = 30000$, $\alpha = 0.0001$, $\beta_1 = 0.5$, $\beta_2 = 0.9$

$P_0 \leftarrow 0, \theta_0 \leftarrow 0$

for $t = 1,\cdots,T$ do

while $\theta$ has not converged do

Sample $\{(\mathbf{z}_i, c_i)\}_{i=1}^m \sim P_{\text{data}}$

Sample $\{\mathbf{e}_i\}_{i=1}^m \sim \mathcal{N}(\mathbf{0}, \nu^2 I_n)$

Sample $\{\mathbf{\xi}_i\}_{i=1}^{m'} \sim \hat{\zeta}$

$L \leftarrow \frac{1}{m} \sum_{i=1}^m L(c_i, f_\theta(\mathbf{z}_i + \mathbf{e}_i))$

$\theta \leftarrow \text{Adam}(\nabla_\theta L, \theta, \alpha, \beta_1, \beta_2)$

$\theta_t \leftarrow \theta$

Sample $\{\chi_i\}_{i=1}^l \sim \hat{\zeta}$

Calculate $M_t = \frac{1}{l} \sum_{i=1}^l \left( \frac{\partial f_\theta(\chi_i)}{\partial \mathbf{x}} \right)^T$

Find $\{\mathbf{v}_i\}_{i=1}^l$ s.t. $M_t \mathbf{v}_i = \lambda_i \mathbf{v}_i$, $\lambda_1 \geq \cdots \geq \lambda_n$

$P_t \leftarrow \sum_{i=1}^k \mathbf{v}_i \mathbf{v}_i^T$

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