Theoretical Insights on the Pre-image Resolution in Machine Learning

Paul Honeine

Theoretical Insights on the Pre-image Resolution in Machine Learning

Paul Honeine[a,*]

[a] *Univ Rouen Normandie, INSA Rouen Normandie, Université Le Havre Normandie, Normandie Univ, LITIS UR 4108, F-76000 Rouen, France*

Abstract

While many nonlinear pattern recognition and data mining tasks rely on embedding the data into a latent space, one often needs to extract the patterns in the input space. Estimating the inverse of the nonlinear embedding is the so-called pre-image problem. Several strategies have been proposed to address the estimation of the pre-image; However, there are no theoretical results so far to understand the pre-image problem and its resolution. In this paper, we provide theoretical underpinnings of the resolution of the pre-image problem in Machine Learning. These theoretical results are on the gradient descent optimization, the fixed-point iteration algorithm and Newton's method. We provide sufficient conditions on the convexity/nonconvexity of the pre-image problem. Moreover, we show that the fixed-point iteration is a Newton update and prove that it is a Majorize-Minimization (MM) algorithm where the surrogate function is a quadratic function. These theoretical results are derived for the wide classes of radial kernels and projective kernels. We also provide other insights by connecting the resolution of this problem to the gradient density estimation problem with the so-called mean shift algorithm.

Keywords: Machine Learning, Pattern Recognition, Pre-image Problem, Fixed-point Iteration, Newton's Method, Majorize-Minimization Algorithm

*Corresponding author:
Email address: paul.honeine@univ-rouen.fr (Paul Honeine)

1. Introduction

In Machine Learning (ML), there has been an increasing interest in the embedding principle for nonlinear pattern recognition and data mining, driven by kernel machines and the revival of deep neural networks. The backbone of these machines is the preprocessing of the data with a nonlinear transformation (in kernel machines) or a cascade of nonlinear transformations (in deep neural networks). Such transformations embed the data into a latent space (often called feature space) where data-processing techniques can be easily carried out. A major bottleneck is that one often needs to extract patterns in the input space, not in the latent space. It is therefore necessary to represent in the input space the results obtained in the latent space. This inverse map of the nonlinear embedding is the so-called pre-image.

Establishing the pre-image is generally an ill-posed problem. Instead of aiming for an exact pre-image, one estimates an approximate pre-image. Finding the pre-image is a hard optimization problem since the objective function is inherently nonlinear. Several strategies have been proposed to address the estimation of the pre-image [22]. While methods inspired from dimensionality reduction and manifold learning literature operate as black-boxes, gradient-descent and fixed-point iteration techniques provide a direct resolution of the pre-image problem. Nevertheless, the underlying mechanism is still unclear. To the best of our knowledge, there are no theoretical results that allow to clearly analyze the pre-image problem and its resolution.

In this paper, we provide theoretical insights on the pre-image problem and its resolution. This problem has been known for about 20 years [33], and is still of great interest nowadays, as shown recently in many frameworks: when dealing with nonlinear dictionary learning [40, 31] and matrix completion [17]; when dealing with structured input spaces, such as in interpretable time series analytics [36], graph edit distances [23, 24], representation learning on graphs [10] and structured prediction [15]; generative machines including generative kernel PCA [29] and multiview generation [30].

The main idea behind this work is to provide a novel viewpoint on the pre-image problem and its resolution, thanks to recent advances in the literature of gradient density estimation. However, such analogy is difficult to make and challenging due to several reasons (see Appendix A for more details): Kernel-based ML and kernel density estimation (also known as Parzen window estimator in statistics) do not share the same foundations, namely not all valid kernels in the former are valid for the latter, and vice versa. Moreover, the parameters in kernel-based ML are arbitrary, making the pre-image problem more difficult to address due to its nonlinearity and nonconvexity in general. Therefore, these issues make the theoretical analysis of the pre-image problem very challenging and needs to be handled with care.

This paper presents first results on the first-order optimization of the pre-image problem (Section 3) and then on the second-order optimization (Section 4), providing sufficient conditions on the convexity and nonconvexity of the problem under study. We then provide connections to other optimization problems, by proving that the fixed-point iteration is a Newton's step and that it is a Majorize-Minimization (MM) algorithm (Section 5). For the sake of clarity, these theoretical results are carried out on the radial kernels, with the Gaussian kernel being the cornerstone, and then extended to the class of projective kernels, such as the polynomial kernels (Appendix B). Experiments illustrate the main theoretical results (Section 6). These results can be extended to other frameworks, such as infinitely wide neural networks with neural network Gaussian process, neural tangent kernels and neural kernels without tangents [34], and representation learning with kernels [16]. See also [1].

The main contributions of this paper in understanding the pre-image problem and its resolution are as follows, with the main results highlighted in Table 1:

- We provide solid foundations to the resolution of the pre-image problem, by borrowing some ideas from the literature of gradient of the density estimation with the mean shift algorithm; This provides a novel point of view on the pre-image problem in ML.

Table 1: A birdview of the main results in this paper, according to the investigated class of kernels and further underlying conditions

Main theoretical results	Radial kernels	Projective kernels	Conditions
Bound on the norm of the gradient	Theorem 8		Gaussian kernel
Sufficient condition on the nonconvexity	Theorem 14	Theorem 23	
Sufficient condition on the convexity	Theorem 15	Theorem 24	
Fixed-point iteration as a Newton update	Theorem 16	Theorem 25	Piecewise-constant derivative kernels
Fixed-point iteration as a quadratic optimization	Theorem 17	Theorem 26	
Fixed-point iteration as an MM algorithm	Theorem 18	Theorem 27	$\alpha_i \geq 0$

- We provide some theoretical insights on the Hessian of the objective function at hand, including sufficient conditions for its positive definiteness and thus the convexity of the optimization problem.
- We establish a relationship between the fixed-point iteration technique and Newton's method and demonstrate that a fixed-point iteration is a Majorize-Minimization.
- We derive general theoretical results for the wide classes of radial kernels and projective kernels.

2. A Primer on the Pre-image Problem in ML

ML methods aim to infer the structure of the data from a set of available samples. While conventional ML methods operate using linear models, extensions to nonlinear models can be investigated by processing the data with a nonlinear map to some space, prior to the application of conventional algorithms.

Let $\mathcal{X}$ be the input space, endowed with the inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$, where $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = \boldsymbol{x}_i^\top \boldsymbol{x}_j$ when dealing with a vector space. Let $\phi(\cdot)$ be a nonlinear transformation, mapping the data from the input space $\mathcal{X}$ to some latent space $\mathcal{H}$. Then, considering a set of n training samples, denoted $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathcal{X}$, the resulting inference model in the latent space $\mathcal{H}$ takes the form

$$\psi = \sum_{i=1}^{n} \alpha_i \, \phi(\boldsymbol{x}_i), \tag{1}$$

for some parameters $\alpha_1, \alpha_2, \dots \alpha_n$ to be estimated. This representer theorem is well-known in spline interpolation theory, kernel-based methods, deep neural networks, and more generally in inverse problems and ML [37].

The kernel trick provides a mathematically elegant means to derive powerful nonlinear variants of classical linear techniques, by replacing the inner product operator in the latent space with a positive definite kernel $\kappa(\cdot,\cdot)$, i.e., satisfying

$$\sum_{i,j} \xi_i \xi_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$$

for all $\xi_i, \xi_j \in \mathbb{R}$ and $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$. A large class of nonlinear kernels implicitly defines the nonlinear embedding and the corresponding latent space, namely

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}) \rangle_{\mathcal{H}}.$$

A preferred choice of positive definite kernels is the Gaussian kernel defined by

$$\kappa_\sigma(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\tfrac{1}{2\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right), \tag{2}$$

for a bandwidth parameter σ, and the polynomial kernel of degree p defined by

$$\kappa_p(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\tfrac{1}{\sigma}\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + c\right)^p,$$

for some parameters $c \geq 0$ and $\sigma > 0$. While kernel-based methods rely on positive definite kernels, extensions to kernels not satisfying this fundamental property exist. Of particular interest is the class of conditionally positive definite kernels [32], satisfying

$$\sum_{i,j} \xi_i \xi_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0 \text{ for } \sum_i \xi_i = 0.$$

Examples of such kernels are the negative distance kernel $\kappa_{\mathrm{nd}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = c - \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$ and its extension to any positive power p [6, 21].

2.1. The pre-image problem

For nonlinear pattern recognition and data mining, a major bottleneck is that one needs to extract patterns in the data space, not in the "implicit" latent

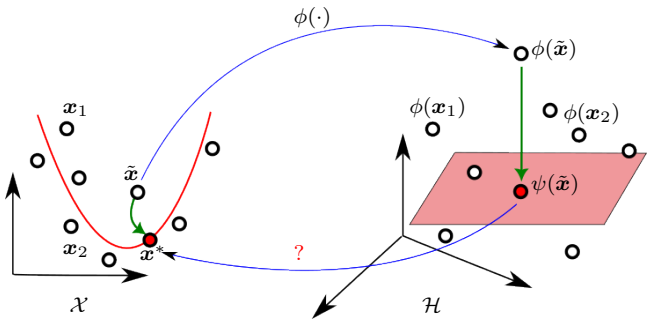

Figure 1: Illustration of the pre-image resolution ("?") as an inverse of the data embedding ($\phi(\cdot)$), allowing to provide nonlinear patterns (given here as a parabola) in the input space $\mathcal{X}$ from linear patterns (given here as a plane) in the latent space $\mathcal{H}$. A linear projection in the latter is associated to a nonlinear one in the former (the two respective green arrows).

one. In other words, we aim to estimate $\boldsymbol{x}^*$ whose image, with the mapping function $\phi(\cdot)$, is ψ, as illustrated in Figure 1. However, most elements of the latent space do not lie in the image of the implicit embedding, namely ψ does not have a valid pre-image in general. This issue can be easily illustrated using the Gaussian kernel, since the representer theorem (1) becomes

$$\psi(\cdot) = \sum_{i=1}^{n} \alpha_i \, \exp\left(-\tfrac{1}{2\sigma^2}\|\boldsymbol{x}_i - \cdot\,\|^2\right),$$

which is a linear combination of Gaussians centered at the input data. However, it is well known that a sum-of-Gaussians centered at different points cannot be written as a single Gaussian $\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x}^* - \cdot\,\|^2\right)$ for some $\boldsymbol{x}^*$, namely, the pre-image of ψ.

In order to circumvent this issue, one seeks an approximate solution, i.e., $\boldsymbol{x}^* \in \mathcal{X}$ whose embedding $\phi(\boldsymbol{x}^*)$ is as close as possible to ψ, namely

$$\phi(\boldsymbol{x}^*) \approx \psi. \tag{3}$$

In the following, we consider the key fundamental pre-image problem to provide deep analysis of the pre-image problem and its resolution. From (3), measuring the similarity between $\phi(\boldsymbol{x}^*)$ and ψ in the latent space $\mathcal{H}$ is done using the

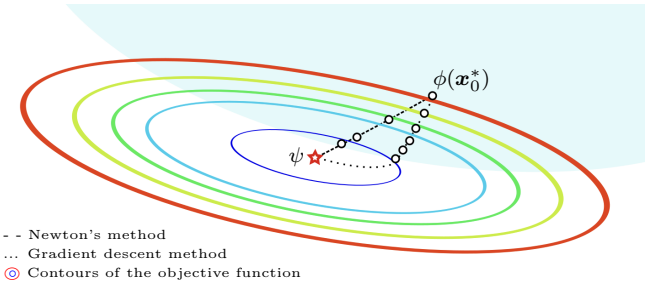

Figure 2: Illustration of the pre-image resolution in the latent space $\mathcal{H}$. The shaded zone corresponds to the image of the input space $\mathcal{X}$ for the function $\phi(\cdot)$, namely any $\phi(\boldsymbol{x})$ belongs to this zone. Any linear combination ψ, as defined by the representer theorem (1), does not belong to this zone in general, and thus it does not have a valid pre-image. To estimate the optimal $\boldsymbol{x}^*$ such that $\phi(\boldsymbol{x}^*) \approx \psi$, we seek to minimize the objective function in (4) (its contours are shown in several colors). Two algorithms are illustrated, starting from a guess $\boldsymbol{x}_0^*$: The first-order optimization with a conventional gradient descent algorithm (dotted path), and the second-order optimization with a Newton's method (dashed path). The major contributions of this paper are theoretical results on the fixed-point iteration technique for solving the pre-image problem, as we demonstrate that it operates as a Newton's method and it is an Majorize-Minimization algorithm (as illustrated in Figure 5).

distance defined in that space, leading to the following optimization problem:

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{2} \Big\| \sum_{i=1}^{n} \alpha_i \, \phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}) \Big\|_{\mathcal{H}}^2, \tag{4}$$

namely,

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathcal{X}} \tfrac{1}{2}\kappa(\boldsymbol{x},\boldsymbol{x}) - \sum_{i=1}^{n} \alpha_i \, \kappa(\boldsymbol{x},\boldsymbol{x}_i), \tag{5}$$

where the term independent of $\boldsymbol{x}$ has been dropped. This optimization problem is the most investigated one in the literature. Moreover, several variants were also proposed, such as including a regularization that penalizes important variations, or considering local smoothing with neighborhood regularization. An illustration of the optimization in the latent space is given in Figure 2.

The structure of the kernel functions provides useful insights to derive appropriate optimization techniques, as one can compute the gradient of the objective function in (5), under the condition of differentiable kernels. For the sake of

clarity, we restrict the presentation to the large class of the radial kernels, since they are the most used in the literature. Appendix B extends these theoretical results to the class of projective kernels, such as the polynomial kernels.

2.2. Radial kernels

In the following, we restrict the presentation to the large class of the radial kernels. To this end, we shall borrow some elements from the literature of the gradient of the density and the mean shift algorithm, such as the notions of profile and shadow [12, 3].

The radial kernels are of the form

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = k(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2), \tag{6}$$

where the function $k(\cdot)$ is called the *profile* of the kernel $\kappa(\cdot, \cdot)$. Let $k'(u)$ be the derivative of the profile function with respect to its argument u, assuming that this derivative exists for all u, except for a finite set of points. Therefore,

$$\nabla_{\boldsymbol{x}} k(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2) = 2(\boldsymbol{x} - \boldsymbol{x}_i)\, k'(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2), \tag{7}$$

where $\nabla_{\boldsymbol{x}}$ denotes the gradient operator with respect to $\boldsymbol{x}$. In the same spirit of [12, Theorem 1], we define the shadow of a kernel, as follows:

Definition 1 (Shadow of a kernel)**.** *A kernel $\kappa(\cdot, \cdot)$ is a shadow of a kernel $\gamma(\cdot, \cdot)$ if and only if their profiles, respectively $k(\cdot)$ and $g(\cdot)$, satisfy*

$$k'(r) = c\, g(r),$$

where c is a positive constant.

Let $k^{(\ell)}(\cdot)$ be the ℓ-th derivative of the function $k(\cdot)$ with respect to its argument. The following proposition is from [14, Proposition 5] (see also [7, Proposition 7.2]).

Proposition 2 (Radial kernels)**.** *A sufficient condition for a function of the form $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = k(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$ to be a positive definite kernel is its complete*

monotonicity, namely it is infinitely differentiable and all its ℓ derivatives satisfy the following condition, for any $u > 0$,

$$(-1)^\ell k^{(\ell)}(u) \geq 0.$$

Since the second derivative is nonnegative, then the following lemma holds.

Lemma 3. *The profile function of a radial kernel is convex.*

Example 4 (Gaussian kernel)**.** *The most-used kernel is the Gaussian kernel,*

$$\kappa_\sigma(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\tfrac{1}{2\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right), \tag{8}$$

for a given bandwidth parameter σ. Its profile function is $k_\sigma(u) = \exp(-\frac{1}{2\sigma^2}u)$, which satisfies Proposition 2, with

$$k_\sigma^{(\ell)}(u) = \left(-\tfrac{1}{2\sigma^2}\right)^\ell k_\sigma(u). \tag{9}$$

Moreover, it turns out that the Gaussian kernel is the only kernel with a profile that is also the profile of its derivatives, up to a multiplicative constant. This is because, for any constant c, $k'(u) = -c\,k(u)$ implies $\int \frac{dk}{k} = \int -c\,du$, namely $\log k(u) - \log k(0) = -c\,r$, which implies $k(u) = k(0)\exp(-cu)$. □

Example 5 (Inverse quadratic kernel)**.** *The inverse quadratic kernel defined by*

$$\kappa_{invquad}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(c + \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right)^{-p}, \tag{10}$$

for some parameters $c > 0$ and $p > 0$, is a valid radial kernel with the profile function $k_{invquad}(u) = (c+u)^{-p}$. □

Example 6 (Epanechnikov kernel)**.** *The Epanechnikov kernel is the truncated negative distance kernel, defined as*

$$\kappa_{Ep}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{cases} c - \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 & \text{if } \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 \leq \rho \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

for some positive parameters ρ and $c \geq 0$. It has the profile function

$$k_{Ep}(u) = \begin{cases} c - u & \text{if } |u| \leq \rho \\ 0 & \text{otherwise} \end{cases}$$

and its derivatives are

$$k'_{Ep}(u) = \begin{cases} -1 & \text{if } |u| \leq \rho \\ 0 & \text{otherwise} \end{cases}$$

which is the profile of the uniform rectangular kernel on the support $[-\rho, \rho]$. *Moreover,* $k''_{Ep}(u) = 0$.

From Proposition 2, it is easy to see that a sufficient condition for this kernel to be positive definite is $c \geq \rho$. *Otherwise, it is conditionally positive definite, since* $\sum_i \xi_i = 0$ *implies* $\sum_{i,j} \xi_i \xi_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = c \sum_{i,j} \xi_i \xi_j - \sum_{i,j} \xi_i \xi_j \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 = 2 \sum_{i,j} \xi_i \xi_j \boldsymbol{x}_i^\top \boldsymbol{x}_j - \sum_{i,j} \xi_i \xi_j \|\boldsymbol{x}_i\|^2 - \sum_{i,j} \xi_i \xi_j \|\boldsymbol{x}_j\|^2 = 2 \| \sum_i \xi_i \boldsymbol{x}_i \|^2 \geq 0$. □

3. First-order Optimization

When considering radial kernels (6), $\kappa(\boldsymbol{x}, \boldsymbol{x}) = k(0)$ is independent of $\boldsymbol{x}$. Thus, the pre-image is obtained by the minimization of the objective function

$$\Xi(\boldsymbol{x}) = -\sum_{i=1}^{n} \alpha_i \, k(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2). \tag{12}$$

To this end, we shall examine its gradient, namely

$$\nabla_{\boldsymbol{x}} \Xi(\boldsymbol{x}) = 2 \sum_{i=1}^{n} \alpha_i \, (\boldsymbol{x}_i - \boldsymbol{x}) \, k'(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2), \tag{13}$$

where we have explored the structure of radial kernels with (7).

3.1. Fixed-point iteration technique

By factorizing the expression of the gradient (13) in the form

$$\nabla_{\boldsymbol{x}} \Xi(\boldsymbol{x}) = 2 \left(-\sum_{i=1}^{n} \alpha_i k'(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2) \right) \left(\boldsymbol{x} - \frac{\sum_{i=1}^{n} \alpha_i k'(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2) \boldsymbol{x}_i}{\sum_{i=1}^{n} \alpha_i k'(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2)} \right), \tag{14}$$

we can notice the following.

The first term is a scalar. It can be viewed as the pattern ψ evaluated at $\boldsymbol{x}$ using the kernel $\gamma(\cdot, \cdot)$, namely the kernel whose shadow is $\kappa(\cdot, \cdot)$ (up to a

multiplicative constant). When dealing with the Gaussian kernel, since this kernel has the same expression as its shadow, then the first term becomes

$$\tfrac{1}{2\sigma^2}\sum_{i=1}^{n}\alpha_i\kappa(\boldsymbol{x},\boldsymbol{x}_i)=\tfrac{1}{2\sigma^2}\psi(\boldsymbol{x}).$$

We have integrated the "-" in the first term of (14) since, when dealing with nonnegative α_i's, this term becomes nonnegative thanks to Proposition 2.

The second term is the vector $\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{x})$, where

$$\boldsymbol{\mu}(\boldsymbol{x})=\frac{\sum_{i=1}^{n}\alpha_i k'(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)\,\boldsymbol{x}_i}{\sum_{i=1}^{n}\alpha_i k'(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)}$$

is a weighted combination of the $\boldsymbol{x}_i$'s with the weights depending on the α_i's and the kernel $\gamma(\cdot,\cdot)$. Since the first term of the expression of the gradient (14) is a scalar, nonnegative when dealing with nonnegative coefficients α_i's, and the second one a vector in $\mathcal{X}$, then $\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{x})$ and $\nabla_{\boldsymbol{x}}\Xi(\boldsymbol{x})$ share the same direction, which is the one that minimizes the objective function (12). The expression $\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{x})$ can be viewed as a weighted mean shift. Indeed, in the special case of constant positive coefficients α_i for all $i=1,2,\ldots,n$, it boils down to the so-called *mean shift*. In the following, we take advantage of the literature on the mean shift algorithm and extend it to the case of a weighted mean shift in order to provide a deep analysis on the pre-image problem and its resolution.

The fixed-point iteration technique can be derived by nullifying the gradient of (12), namely the last term in (14). This leads to the fixed-point iteration that updates a guess $\boldsymbol{x}_t^*$ at iteration t to the new estimate $\boldsymbol{x}_{t+1}^*=\boldsymbol{\mu}(\boldsymbol{x}_t^*)$, namely

$$\boldsymbol{x}_{t+1}^*=\frac{\sum_{i=1}^{n}\alpha_i k'(\|\boldsymbol{x}_t^*-\boldsymbol{x}_i\|^2)\,\boldsymbol{x}_i}{\sum_{i=1}^{n}\alpha_i k'(\|\boldsymbol{x}_t^*-\boldsymbol{x}_i\|^2)}. \tag{15}$$

If this sequence converges to some $\boldsymbol{x}_\infty^*$, then the last term in the gradient expression (14) vanishes, so does the gradient at this point. In other words, if convergence, then its limit is a critical point of the objective function (12).

Example 7 (Gaussian kernel)**.** *For the Gaussian kernel, the gradient in* (13) *becomes*

$$\nabla_{\boldsymbol{x}}\Xi(\boldsymbol{x})=-\frac{2}{\sigma^2}\sum_{i=1}^{n}\alpha_i\big(\boldsymbol{x}_i-\boldsymbol{x}\big)\exp(-\tfrac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{x}_i\|^2).$$

Since this kernel has the same expression as its derivative, up to a multiplicative factor, then the fixed-point iteration technique becomes, as in [27],

$$\boldsymbol{x}_{t+1}^* = \frac{\sum_{i=1}^n \alpha_i \, \kappa_\sigma(\boldsymbol{x}_t^*, \boldsymbol{x}_i) \; \boldsymbol{x}_i}{\sum_{i=1}^n \alpha_i \, \kappa_\sigma(\boldsymbol{x}_t^*, \boldsymbol{x}_i)}. \tag{16}$$

□

3.2. Gradient descent

The gradient descent is one of the simplest optimization techniques. In its simplest form, $\boldsymbol{x}_t^*$ is updated to $\boldsymbol{x}_{t+1}^*$ by stepping in the opposite direction of the gradient, namely

$$\boldsymbol{x}_{t+1}^* = \boldsymbol{x}_t^* - \eta_t \, \nabla_{\boldsymbol{x}} \Xi(\boldsymbol{x}_t^*),$$

with the gradient $\nabla_{\boldsymbol{x}} \Xi(\boldsymbol{x})$ given in (14), where η_t is the stepsize parameter. The following theorem shows that the norm of the gradient is bounded.

Theorem 8. *When considering the Gaussian kernel, the norm of the gradient is upper bounded as follows:*

$$\|\nabla_{\boldsymbol{x}} \Xi(\boldsymbol{x})\| \leq \tfrac{2}{\sigma\sqrt{e}} \|\alpha\|_1.$$

Proof. Let $\boldsymbol{g}_i(\boldsymbol{x}) = 2\, k'(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2)\, (\boldsymbol{x}_i - \boldsymbol{x})$ for $i = 1, 2, \ldots, n$, then $\nabla_{\boldsymbol{x}} \Xi(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i \, \boldsymbol{g}_i(\boldsymbol{x})$ for any radial kernel. Thus, we have

$$\|\nabla_{\boldsymbol{x}} \Xi(\boldsymbol{x})\| = \Big\| \sum_{i=1}^n \alpha_i \boldsymbol{g}_i(\boldsymbol{x}) \Big\| \leq \sum_{i=1}^n |\alpha_i| \, \|\boldsymbol{g}_i(\boldsymbol{x})\| \,, \tag{17}$$

from the triangular inequality. In the following, we aim to provide an upper bound on the norm of each $\boldsymbol{g}_i(\boldsymbol{x})$, which is the gradient of $k(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2)$. The maximum norm of the gradient lies at the inflexion point of this function.

When dealing with the Gaussian kernel, the inflexion points are obtained at $\|\boldsymbol{x}_i - \boldsymbol{x}\|^2 = \sigma^2$ (which is a straightforward generalization to higher dimensions of the 1D case, where nullifying the second derivative implies $|x_i - x| = \sigma$). At the inflexion points, the norm of $\boldsymbol{g}_i(\boldsymbol{x})$ becomes

$$\|\boldsymbol{g}_i(\boldsymbol{x})\| = \Big\| -\tfrac{2}{\sigma^2} \big(\boldsymbol{x}_i - \boldsymbol{x}\big) \exp\big(-\tfrac{1}{2\sigma^2} \|\boldsymbol{x} - \boldsymbol{x}_i\|^2\big) \Big\| = \tfrac{2}{\sigma\sqrt{e}}.$$

As this expression is independent of the $\boldsymbol{x}_i$'s, the upper bound on the gradient norm (17) can be simplified to

$$\|\nabla_{\boldsymbol{x}}\Xi(\boldsymbol{x})\| \leq \tfrac{2}{\sigma\sqrt{e}} \sum_{i=1}^{n} |\alpha_i|.$$

□

By using norm inequalities, we can get an upper bound in terms of the ℓ_2-norm or ∞-norm of $\boldsymbol{\alpha}$, since $\|\cdot\|_1 \leq \sqrt{n}\|\cdot\|_2$ and $\|\cdot\|_1 \leq n\|\cdot\|_\infty$. This allows to provide upper bounds for different ML methods, as most of them define some constraints on the α_i's. This is the case of considering centroids, with $\alpha_i = 1/n$ which implies an upper bound of $\frac{2}{\sigma\sqrt{e}}$. The kernel PCA algorithm also imposes some constraints, as given in the following example.

Example 9. *For the kernel PCA, the most used normalization is the unit-norm of the eigenvectors, which yields $\lambda_k\|\boldsymbol{\alpha}_k\|_2^2 = 1$, where λ_k denotes the Gram-eigenvalue associated to the k-th principal axis. This leads to the following bound on the gradient norm:*

$$\|\nabla_{\boldsymbol{x}}\Xi(\boldsymbol{x})\|^2 \leq \frac{4n}{\sigma^2\lambda_k e}.$$

Sometimes, a whitening normalization is recommended to have an equal variance in each direction [35]. This is done using the normalization $\lambda_k\|\boldsymbol{\alpha}_k\|_2 = 1$, which implies $\|\nabla_{\boldsymbol{x}}\Xi(\boldsymbol{x})\|^2 \leq \frac{4n}{\sigma^2\lambda_k^2 e}$. □

The gradient descent method has several drawbacks. The stepsize parameter η_t needs to be determined; However, a line-search procedure is computationally expensive. Moreover, the objective function (12) is inherently nonlinear and clearly nonconvex. Thus, a gradient descent algorithm must be run many times with different starting values, in hope that a feasible solution will be amongst the local minima obtained over the runs. To overcome these difficulties of the gradient descent and provide an adapted step size parameter, we shall investigate in the following Newton's method.

4. Second-order Optimization

In this section, we examine the second-order optimization to solve the pre-image problem, by considering Newton's method with an update of the form

$$\boldsymbol{x}_{t+1}^* = \boldsymbol{x}_t^* - \Big(\nabla^2_{\boldsymbol{x}_t^*}\Xi(\boldsymbol{x}_t^*)\Big)^{-1}\ \nabla_{\boldsymbol{x}_t^*}\Xi(\boldsymbol{x}_t^*), \tag{18}$$

where one needs to inverse the Hessian matrix $\nabla^2_{\boldsymbol{x}}\Xi(\boldsymbol{x})$ of the objective function (12). The developments conducted in this section shall allow us to connect the corresponding Newton's method to the fixed-point iteration technique.

4.1. The Hessian matrix

The Hessian matrix of (12) is the Jacobian matrix of its gradient, namely

$$\begin{aligned}\nabla^2_{\boldsymbol{x}}\Xi(\boldsymbol{x}) &= 2\sum_{i=1}^{n}\alpha_i\,\nabla_{\boldsymbol{x}}\big((\boldsymbol{x}_i-\boldsymbol{x})\,k'(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)\big)\\ &= -2\sum_{i=1}^{n}\alpha_i\Big(k'(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)\boldsymbol{I} + 2k''(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)(\boldsymbol{x}_i-\boldsymbol{x})(\boldsymbol{x}_i-\boldsymbol{x})^\top\Big).\end{aligned} \tag{19}$$

Example 10. *For the Gaussian kernel, we get from* (9) *the expression of the Hessian matrix*

$$\tfrac{1}{\sigma^2}\sum_{i=1}^{n}\alpha_i\,k_\sigma(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)\left(\boldsymbol{I}-\tfrac{1}{\sigma^2}(\boldsymbol{x}_i-\boldsymbol{x})(\boldsymbol{x}_i-\boldsymbol{x})^\top\right).$$

For the Epanechnikov kernel, the Hessian matrix can also be easily computed, as $k''(u)$ *vanishes from* (19) *while* $k'(u)$ *is a simple rectangular function.* □

The Hessian measures the local curvature of the function $\Xi(\cdot)$ and provides a viewpoint on its convexity. To show this, we rewrite this Hessian matrix as

$$-2\sum_{i=1}^{n}\alpha_i\,k'(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)\Big(\boldsymbol{I}+2\frac{k''(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)}{k'(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)}(\boldsymbol{x}_i-\boldsymbol{x})(\boldsymbol{x}_i-\boldsymbol{x})^\top\Big). \tag{20}$$

Since we have $k'(\cdot)\leq 0$ and $k''(\cdot)\geq 0$ from Proposition 2, then the above expression can be viewed as a linear combination of matrices of the form $\boldsymbol{I}-\boldsymbol{Q}_i$, where $\boldsymbol{Q}_i=\eta_i(\boldsymbol{x}_i-\boldsymbol{x})(\boldsymbol{x}_i-\boldsymbol{x})^\top$ with nonnegative η_i defined by

$$\eta_i = 2\frac{k''(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)}{-k'(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)}.$$

Moreover, the matrices $\boldsymbol{Q}_i$ are positive definite, since we have for all $\boldsymbol{u}$

$$\boldsymbol{u}^\top \boldsymbol{Q}_i \boldsymbol{u} = 2\frac{k''(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)}{-k'(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)} \left\|\boldsymbol{u}^\top(\boldsymbol{x}_i-\boldsymbol{x})\right\|^2 \geq 0.$$

By writing the Hessian as (20), namely with the difference of two positive definite matrices $\boldsymbol{I}$ and $\boldsymbol{Q}_i$, we get some insights on the underlying mechanism. Indeed, the Hessian is not positive definite in general, as the difference of two positive definite matrices is not positive definite in general.

Before providing some insights using linear algebra, we give the following lemma that is due to the nonnegativity of $k'(\cdot)$.

Lemma 11. *If the coefficients α_i's are nonnegative, then the Hessian matrix is a difference of two positive definite matrices, the first one being the identity matrix and the second one is the sum of rank-one matrices (up to some nonlinear scaling).*

While this result is restricted to nonnegative coefficients, we aim in the following to consider the general case.

4.2. Conditions on the positive definiteness of the Hessian

In this section, we aim to provide a sufficient condition for the positive definiteness of the Hessian, by providing a lower bound on its eigenvalues. To this end, we first bring to mind Weyl's interlacing theorem for the eigenvalues of the sum of two matrices. We write it here for a rank-one update.

Lemma 12 (Weyl's theorem)**.** *Consider a single rank-one update of the form $\boldsymbol{A} \pm \boldsymbol{u}\boldsymbol{u}^\top$. Weyl's theorem for its smallest eigenvalue verifies the inequalities:*

$$\begin{aligned} \lambda_{\min}(A) &\leq \lambda_{\min}(A+\boldsymbol{u}\boldsymbol{u}^\top) \leq \lambda_{\min}(A)+\|\boldsymbol{u}\|^2 \\ \lambda_{\min}(A)-\|\boldsymbol{u}\|^2 &\leq \lambda_{\min}(A-\boldsymbol{u}\boldsymbol{u}^\top) \leq \lambda_{\min}(A) \end{aligned}$$

Proof. Weyl's theorem is well-known for the sum of two arbitrary matrices, of the form $|\lambda_j(A+B)-\lambda_j(A)| \leq \|B\|$ for any symmetric matrices A and B [20,

Theorem 8.1.5 & Corollary 8.1.6]. Considering a rank-one matrix $B = \boldsymbol{u}\boldsymbol{u}^\top$, which has $\lambda_{\min}(B) = 0$ and $\lambda_{\max}(B) = \|\boldsymbol{u}\|^2$, we get

$$\lambda_{\min}(A) \leq \lambda_{\min}(A + \boldsymbol{u}\boldsymbol{u}^\top) \leq \lambda_{\min}(A) + \|\boldsymbol{u}\|^2.$$

Replacing A with $A - \boldsymbol{u}^\top\boldsymbol{u}$ leads to the second pair of inequalities in Lemma 12. □

The following Lemma extends Weyl's theorem to multiple rank-one updates.

Lemma 13. *For any matrix given as a linear combination of n rank-one updates, of the form $H = H_0 - \sum_{i=1}^n \nu_i \boldsymbol{r}_i\boldsymbol{r}_i^\top$ for an arbitrary matrix H_0, vectors $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$, and scalars $\nu_1, \ldots, \nu_n$, we have*

$$\lambda_{\min}(H_0) - \sum_{\substack{i=1 \\ \nu_i>0}}^n \nu_i \|\boldsymbol{r}_i\|^2 \leq \lambda_{\min}(H) \leq \lambda_{\min}(H_0) - \sum_{\substack{i=1 \\ \nu_i<0}}^n \nu_i \|\boldsymbol{r}_i\|^2. \tag{21}$$

Proof. Let $\mathbb{1}_{\nu>0}$ be the indicator function, namely $\mathbb{1}_{\nu>0} = 1$ if $\nu > 0$ and 0 otherwise. Then, the left-hand-side dual inequalities in Lemma 12 can be summarized in the single expression

$$\lambda_{\min}(A) \leq \lambda_{\min}(A - \nu\,\boldsymbol{u}\boldsymbol{u}^\top) + \mathbb{1}_{\nu>0}\,\nu\,\|\boldsymbol{u}\|^2,$$

for any scalar ν, where $\nu = -1$ leads to $A + \boldsymbol{u}\boldsymbol{u}^\top$ and $\nu = +1$ leads to $A = -\boldsymbol{u}\boldsymbol{u}^\top$.

Now, we apply this inequality in a chain rule to $H = H_0 - \nu_1\,\boldsymbol{r}_1\boldsymbol{r}_1^\top - \nu_2\,\boldsymbol{r}_2\boldsymbol{r}_2^\top \ldots - \nu_n\,\boldsymbol{r}_n\boldsymbol{r}_n^\top$. Thus, we get

$$\begin{aligned}
\lambda_{\min}(H_0) &\leq \lambda_{\min}(H_0 - \nu_1\boldsymbol{r}_1\boldsymbol{r}_1^\top) + \mathbb{1}_{\nu_1>0}\,\nu_1\,\|\boldsymbol{r}_1\|^2 \\
&\leq \lambda_{\min}(H_0 - \nu_1\boldsymbol{r}_1\boldsymbol{r}_1^\top - \nu_2\boldsymbol{r}_2\boldsymbol{r}_2^\top) + \mathbb{1}_{\nu_1>0}\,\nu_1\,\|\boldsymbol{r}_1\|^2 + \mathbb{1}_{\nu_2>0}\,\nu_2\,\|\boldsymbol{r}_2\|^2 \\
\cdots \qquad &\leq \lambda_{\min}(H) + \sum_{i=1}^n \mathbb{1}_{\nu_i>0}\,\nu_i\,\|\boldsymbol{r}_i\|^2,
\end{aligned}$$

which concludes the proof for the left-hand-side of (21). The right-hand-side can be obtained by considering the right-hand-side dual inequalities in Lemma 12, namely $\lambda_{\min}(A - \nu\,\boldsymbol{u}\boldsymbol{u}^\top) \leq \lambda_{\min}(A) - \mathbb{1}_{\nu<0}\,\nu\,\|\boldsymbol{u}\|^2$ for any scalar ν. By applying this inequality n times for each rank-one update, we get

$$\lambda_{\min}(H) \leq \lambda_{\min}(H_0) - \sum_{i=1}^n \mathbb{1}_{\nu_i<0}\,\nu_i\,\|\boldsymbol{r}_i\|^2,$$

which concludes the proof. □

Back to our problem, we use these results to provide a condition for the Hessian to be not positive definite, and a condition for being positive definite.

Theorem 14. *A sufficient condition to not have a positive definite Hessian is*

$$-\sum_{i=1}^{n} \alpha_i \, k'(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2) < 2 \sum_{\substack{i=1 \\ \alpha_i<0}}^{n} \alpha_i \, k''(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2) \, \|\boldsymbol{x}_i-\boldsymbol{x}\|^2,$$

thus the pre-image is a saddle point of (12).

Proof. We rewrite the Hessian matrix (19) as follows $\nabla^2_{\boldsymbol{x}} \Xi(\boldsymbol{x}) = \nu_0 \, \boldsymbol{I} - \sum_{i=1}^{n} \alpha_i \, \boldsymbol{r}_i \boldsymbol{r}_i^\top$, with $\boldsymbol{r}_i = 2\sqrt{k''(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)}\,(\boldsymbol{x}_i - \boldsymbol{x})$ and $\nu_0 = -2\sum_{i=1}^{n} \alpha_i \, k'(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)$, where we have used the nonnegativity of $k''(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)$ (Proposition 2). From Lemma 13, the right-hand-side of (21) becomes

$$\lambda_{\min}(\nabla^2_{\boldsymbol{x}} \Xi(\boldsymbol{x})) \leq \lambda_{\min}(\nu_0 \boldsymbol{I}) - \sum_{\substack{i=1 \\ \alpha_i<0}}^{n} \alpha_i \, \|\boldsymbol{r}_i\|^2.$$

Therefore, a sufficient condition for having at least one negative eigenvalue is obtained by setting the upper bound below zero, which concludes the proof. □

One can see that this sufficient condition is not verified when dealing with constant nonnegative coefficients α_i's, since $k'(\cdot) \leq 0$ for any radial kernel. The following theorem provides a sufficient condition on the positive definiteness of the Hessian, by considering an upper bound on its smallest eigenvalues.

Theorem 15. *A sufficient condition for the Hessian to be positive definite is*

$$2 \sum_{\substack{i=1 \\ \alpha_i>0}}^{n} \alpha_i \, k''(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2) \, \|\boldsymbol{x}_i-\boldsymbol{x}\|^2 < -\sum_{i=1}^{n} \alpha_i \, k'(\|\boldsymbol{x}-\boldsymbol{x}_i\|^2).$$

Proof. The proof follows the same outline as in the proof of the previous theorem. Considering the left-hand-side of (21), Lemma 13 implies

$$\lambda_{\min}(\nu_0 \boldsymbol{I}) - \sum_{\substack{i=1 \\ \alpha_i>0}}^{n} \alpha_i \, \|\boldsymbol{r}_i\|^2 \leq \lambda_{\min}(\nabla^2_{\boldsymbol{x}} \Xi(\boldsymbol{x})).$$

Therefore, we get a sufficient condition for positive eigenvalues of the Hessian by imposing the positivity of the left-hand-side. □

It is easy to understand this theorem through some special cases. For instance, if all coefficients are negative, this condition is not satisfied. Moreover, when dealing with the Gaussian kernel and constant nonnegative coefficients α_i's, the sufficient condition in Theorem 15 boils down to

$$\sum_{i=1}^{n} k(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2) \left\| \tfrac{1}{\sigma}(\boldsymbol{x}_i - \boldsymbol{x}) \right\|^2 < \sum_{i=1}^{n} k(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2). \tag{22}$$

This result can be related to Theorem D.4 in [8], where the author studied the estimation of the modes of a finite mixture of M normal distributions. In this case, the derived sufficient condition for the positive definiteness of the associated Hessian matrix is $\sum_{m=1}^{M} p(m|\boldsymbol{x}) \left\| (\boldsymbol{x}_i - \boldsymbol{x})/\sigma \right\|^2 < 1$, where the value 1 of the right-hand-side is due to the unitarity axiom of the probability distributions. Therefore, one can provide an analogy between this result and (22). Nevertheless, Theorem 15 provides a more general result for any radial kernel and for any set of weighting coefficients α_i's.

5. Connections to Optimization Methods

In this section, we provide further insights on the pre-image resolution. First, we provide the equivalence between the fixed-point iteration and Newton's method. Moreover, we show that the former is an MM algorithm.

5.1. The fixed-point iteration as a Newton step

Applying Newton's method to solve the pre-image problem leads to the update rule (18), where one needs to inverse the Hessian matrix given in (19). The following theorem on the pre-image problem allows to show the equivalence between Newton's method and the fixed-point iteration technique.

Theorem 16. *If the profile $k(\cdot)$ has a piecewise-constant derivative $k'(\cdot)$, then the fixed-point iteration* (15) *is equivalent to a Newton update* (18)*.*

Proof. If $k'(\cdot)$ is piecewise constant, then $k''(u) = 0$ for all $u \in \mathbb{R}$ and the second term in (19) vanishes. By introducing the remaining term in (18) and

the multiplicative expression of the gradient in (14), we get

$$\begin{aligned}\boldsymbol{x}_{t+1}^* &= \boldsymbol{x}_t^* - \left(\nabla^2_{\boldsymbol{x}_t^*}\Xi(\boldsymbol{x}_t^*)\right)^{-1} \nabla_{\boldsymbol{x}_t^*}\Xi(\boldsymbol{x}_t^*) \\ &= \boldsymbol{x}_t^* + \left(\frac{\sum_{i=1}^n \alpha_i k'(\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2)\,\boldsymbol{x}_i}{\sum_{i=1}^n \alpha_i k'(\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2)} - \boldsymbol{x}_t^*\right),\end{aligned}$$

which corresponds to a step in the fixed-point method. □

This theorem shows that the direction and stepsize in the fixed-point iteration technique are exactly the direction and stepsize of the Newton's method, when dealing with a kernel that has a piecewise constant profile, such as the Epanechnikov kernel or other kernels proposed in [5].

Theorem 16 is restricted to kernels with a profile $k(\cdot)$ that is a shadow of a piecewise-constant profile $k'(\cdot)$. In the following, we generalize this result to any kernel. The following theorem examines the optimization mechanism operated by the fixed-point iteration. See Figure 5 in Section 6 for an illustration.

Theorem 17. *The fixed-point iteration* (15) *at a guess* $\boldsymbol{x}_t^*$ *seeks the optimum of the quadratic function*

$$q(\boldsymbol{x}) = -\sum_{i=1}^n \alpha_i\, k'(\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2)\, \|\boldsymbol{x} - \boldsymbol{x}_i\|^2 - C(\boldsymbol{x}_t^*),$$

with $C(\boldsymbol{x}_t^*) = \sum_{i=1}^n \alpha_i \left(k(\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2) - k'(\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2)\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2\right)$ *independent of* $\boldsymbol{x}$*. Furthermore, the quadratic function* $q(\cdot)$ *is tangent to* $\Xi(\cdot)$ *at* $\boldsymbol{x}_t^*$*.*

Proof. First of all, it is easy to see that $\Xi(\boldsymbol{x}_t^*) = q(\boldsymbol{x}_t^*)$. Now, taking the first derivative of $q(\boldsymbol{x})$ with respect to its argument, we have

$$\nabla_{\boldsymbol{x}} q(\boldsymbol{x}) = 2\sum_{i=1}^n \alpha_i\, k'(\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2)(\boldsymbol{x}_i - \boldsymbol{x}), \tag{23}$$

which is exactly the gradient $\nabla_{\boldsymbol{x}}\Xi(\boldsymbol{x}_t^*)$ given in (13). Therefore, $q(\cdot)$ is tangent to $\Xi(\cdot)$ at $\boldsymbol{x}_t^*$. Finally, the optimum of the quadratic form is obtained when its gradient (23) is nullified, which yields the fixed-point iteration at $\boldsymbol{x}_t^*$. Furthermore, since $q(\boldsymbol{x})$ is quadratic, then Newton's method finds its exact optimum. □

Besides the relations given in this theorem between $q(\cdot)$ and $\Xi(\cdot)$, we provide next other connections. For instance, besides the fact that both functions share

the same tangent at a given point $\boldsymbol{x}_t^*$, they also share the same local curvature when $k''(\cdot) = 0$. Indeed, the Hessian of $q(\boldsymbol{x})$ with respect to its argument is

$$\nabla^2_{\boldsymbol{x}} q(\boldsymbol{x}_t^*) = -2 \sum_{i=1}^{n} \alpha_i \, k'(\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2) \boldsymbol{I}. \tag{24}$$

This is also the first term of $\nabla^2_{\boldsymbol{x}} \Xi(\boldsymbol{x})$ at $\boldsymbol{x}_t^*$, as given in (19), when $k''(\cdot) = 0$, namely the profile $k(\cdot)$ is linear in its argument. By examining this Hessian, the positivity (resp. negativity) of $-2 \sum_{i=1}^{n} \alpha_i \, k'(\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2)$ determines the minimization (resp. maximization) of the quadratic function in Theorem 17. When dealing with nonnegative coefficients α_i's, this yields a positive definite Hessian and thus a minimization problem, since $k'(\cdot) \leq 0$ from (7).

5.2. The fixed-point iteration as an MM algorithm

Next, we take advantage of the convexity of the profile function for radial kernels (Lemma 3). The following theorem proves that the fixed-point iteration technique is a Majorize-Minimization (MM) algorithm, by showing that it is a bounded optimization with $q(\boldsymbol{x})$ being an upper bound on $\Xi(\boldsymbol{x})$. See Figure 5.

Theorem 18. *For nonnegative coefficients α_i's, the fixed-point iterative technique is an MM algorithm, where the quadratic function $q(\cdot)$ defined in Theorem 17 is the surrogate function that majorizes the objective function* (12).

Proof. First of all, it is easy to see that at any guess $\boldsymbol{x}_t^*$, we have $\Xi(\boldsymbol{x}_t^*) = q(\boldsymbol{x}_t^*)$. Next, we show that $\Xi(\boldsymbol{x}) \leq q(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{X}$. Since the profile function $k(\cdot)$ is convex from Lemma 3, then $k(u) \geq k(v) + k'(v)(u - v)$. By substituting u with $\|\boldsymbol{x} - \boldsymbol{x}_i\|^2$ and v with $\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2$ for any $\boldsymbol{x}$, $\boldsymbol{x}_t^*$ and $\boldsymbol{x}_i$, we get

$$k(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2) \geq k(\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2) + k'(\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2)(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2 - \|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2).$$

Since this expression is valid for all $\boldsymbol{x}_i$, then the inequality holds also for any combination with some nonnegative coefficients α_i's, namely

$$\begin{aligned} -\sum_{i=1}^{n} \alpha_i k(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2) \leq -\sum_{i=1}^{n} \alpha_i \Big(& k'(\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2)(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2) \\ & + k(\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2) - k'(\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2)(\|\boldsymbol{x}_t^* - \boldsymbol{x}_i\|^2) \Big), \end{aligned}$$

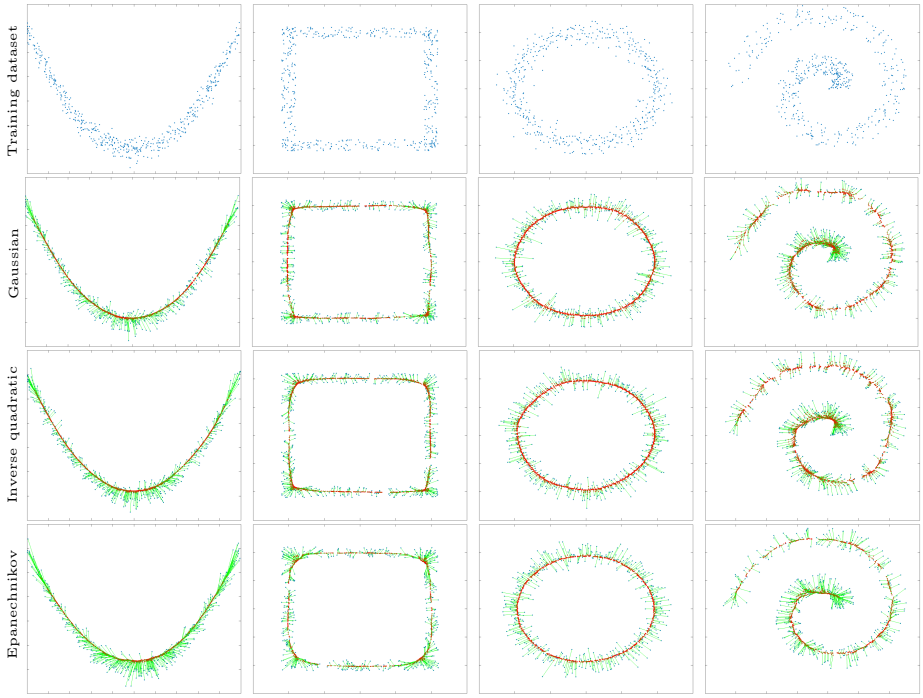

Figure 3: Illustrations of the pre-image resolution (red ·) of the centroid of the data for 4 datasets (blue ·), using the Gaussian, Inverse quadratic and Epanechnikov kernels.

which means that $\Xi(\boldsymbol{x}) \leq q(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{X}$. Finally, the optimal solution of the surrogate function $q(\boldsymbol{x})$ is obtained by nullifying its gradient (23), yielding the fixed-point iteration. This concludes the proof. □

6. Experiments

In this section, we illustrate the main theoretical results on some datasets.

6.1. Illustration on a simple ML problem

We consider the simple ML problem of estimating the centroid of a dataset, where the representer theorem (1) boils down to $\psi = \frac{1}{n}\sum_{i=1}^{n} \phi(\boldsymbol{x}_i)$. In order to illustrate the resolution of the pre-image problem, we consider four datasets in 2D. These datasets consist of 500 samples generated within the shape of a

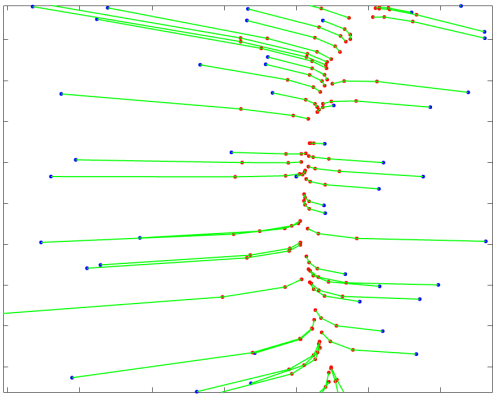

Figure 4: A zoom from the spiral results of Figure 3 that shows the 3 iterations are enough, from a starting point (blue ·) to the 3 estimates (red ·) including the trajectory (solid line).

noisy parabola, frame, ring, and spiral. Figure 3 shows these datasets in its top row, while the other rows present the pre-image results (red dots) obtained by the fixed-point iteration (15) with only 3 iterations. The second row is obtained using the Gaussian kernel, with its bandwidth set to $\sigma = 0.2$, the third row using the inverse quadratic kernel (10) with $c = 1$ and $p = 10$, and the fourth row using the Epanechnikov kernel (11) with $c = 1$ and $\rho = 0.5$. To the best of our knowledge, this is the first time that the relevance of the pre-image is demonstrated using a simple centroid method and the first time that the inverse quadratic and Epanechnikov kernels are used. It is worth noting that the latter is seldom used in kernel-based ML; However, we have proven its usefulness as well as its positive definiteness as it verifies the sufficient condition of Example 6.

To provide more details on these results, Figure 4 shows a zoom from the spiral results given in Figure 3, showing how the 3 iterations are enough to converge to the underlying manifold. This figure also illustrates how the fixed-point iteration technique is an adapted gradient descent / Newton update, as discussed in detail in this paper (see Section 3 and also Theorem 16).

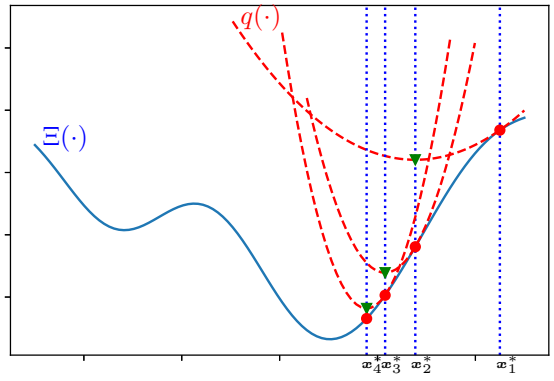

Figure 5: Illustration in 1D of solving the pre-image as a quadratic optimization step. For any guess solution $\boldsymbol{x}_t^*$ (●), the quadratic function $q(\cdot)$ (dashed parabola) is tangent to $\Xi(\cdot)$ (solid line) at $\boldsymbol{x}_t^*$, and its minimum (▼) corresponds to the fixed point iteration yielding $\boldsymbol{x}_{t+1}^*$.

6.2. Connections to optimization methods

We illustrate the equivalence between the fixed-point iteration technique and Newton's method. More precisely, the following experimental results illustrate and support Theorem 17, which proves that the fixed-point iteration technique seeks the optimization of a quadratic function, and Theorem 27, which proves that it is an MM algorithm. To this end, we use a dataset of 20 samples uniformly generated in 1D, with the corresponding coefficients $\alpha_i's$ uniformly generated between 0 and 1 with the sum-to-one constraint and using the Gaussian kernel. Figure 5 illustrates the results. We can see that, starting from any guess solution $\boldsymbol{x}_t^*$, the quadratic function $q(\cdot)$ defined in Theorem 17 is tangent to the pre-image objective function $\Xi(\cdot)$ (12) at that starting solution, and its minimum corresponds to the fixed point iteration yielding the resulting solution. Moreover, $q(\cdot)$ is the surrogate function that majorizes the objective function, as demonstrated in Theorem 27.

7. Conclusion and Future Work

In this paper, we provided theoretical insights on the resolution of the pre-image problem in ML. These results were on the gradient descent optimization,

the fixed-point iteration technique and Newton's method. Connections to other optimization procedures were also proposed, such as demonstrating that the fixed-point iteration is an MM algorithm. These theoretical results were inspired by recent advances on the mean shift algorithm for mode seeking within the framework of gradient density estimation.

Nevertheless, the pre-image problem (arbitrary weights, including nonnegative ones, as well as different kernel definitions) is quite more complex and more challenging than the mode seeking problem (dealing exclusively with probability density functions). It turns out that the resolution of the former boils down to the mean shift algorithm in very specific cases (Gaussian kernel with coefficients α_i's corresponding to a probability distribution), which allows to validate the derived results. For instance, Theorems 15 and 24 (on a sufficient condition for the convexity of the pre-image problem) boil down to Theorem D.4 in [8] (on the convexity of the mode seeking problem) when considering as a special case the Gaussian kernel with normal distributions and constant nonnegative coefficients. In the same sense, Theorems 18 and 27 are more general than Theorem 4 in [18].

As a future work, the convergence of the resolution of the pre-image problem is of great interest. While the derivations carried out in this paper provide insights to the convergence of the fixed-point iteration algorithm, we did not provide a rigorous proof of the convergence. Indeed, the convergence of the mean shift algorithm for mode estimation of probability density functions is still an open question, even though it was introduced almost 50 years ago [19]. This problem has attracted the interest of many researchers recently, with several attempts have been made to prove its convergence. It turns out that the most well-known proofs are incorrect. This is the case of the proof in [13], which relies on the incorrect assumption that the mode estimate sequence generated by the mean shift algorithm is a Cauchy sequence and hence converges; However, this claim is not correct, as pointed out in [26]. Another convergence proof was claimed in [9], where the mean shift with the Gaussian kernel is shown to be an expectation-maximization (EM) algorithm, and hence converges; However,

as pointed out in [2], the EM algorithm may not converge without additional conditions. The more recent proof in [4] also suffers from the flaws of the proof in [13], as pointed out in [38]. A proof for one-dimensional space was provided in [2] under mild conditions. See [38, 39] for a review of proof attempts and their flaws, as well as recent proofs under some mild conditions.

Acknowledgments

This work was supported by the French National Research Agency, grant APi (ANR-18-CE23-0014).

Appendix A. Gradient Density vs. Pre-image Problem in ML

This appendix presents the gradient density estimation problem and provides connections between solving it and the resolution of the pre-image while pointing out the difficulties to carry out such analogy.

Kernel density estimation, also known as the Parzen window estimator in statistics, has been largely investigated in the literature. For a set of available samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathcal{X}$, it is defined by

$$\widehat{p}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha \, k(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2), \tag{A.1}$$

where $k(\cdot)$ is a smoothing kernel, written here in the profile notation defined in (6), and α is a positive normalizing factor that makes the kernel integrate to one. A fundamental property of a density function is its modes, which are the values at which it has local maxima, namely the values that are most likely to be sampled from the underlying probability density. Since these maxima are located at the nullification of the gradient of the density, researchers studied in [19] the estimation of the density gradient, by taking the derivative of the density function estimate (A.1) with respect to $\boldsymbol{x}$, namely

$$\nabla_{\boldsymbol{x}} \widehat{p}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha \nabla_{\boldsymbol{x}} k(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2).$$

By setting it to zero, the *mean shift algorithm* was obtained to seek these modes. The guarantees of asymptotic unbiasedness, consistency, and uniform consistency of this estimate were also studied in [19].

We can view the kernel density estimation as a special case of the pre-image objective function, as the latter is more general with arbitrary coefficients α_i's; Thus, estimating the modes in the former can be related to estimating the pre-image in the latter. However, such analogy needs to be handled with care, as described next due to two major issues.

On one hand, a major issue is that not all kernels used in kernel density estimate are valid positive definite kernels and, vice versa, not all positive definite kernels are valid for kernel density estimate. This is due to the definition of the kernels in each domain. In kernel density estimation, a smoothing kernel is a nonnegative real-valued integrable function that integrates to one, is symmetric about the origin, and may also have other properties such as finite second moment. In kernel-based ML, kernels should verify to be positive definite (Mercer theorem). While the Gaussian kernel (Example 4) is a valid kernel for both kernel-based machines and kernel density estimation (up to a normalizing factor), this is not the case of most other kernels. Of particular interest in kernel density estimation is the Epanechnikov kernel, which is the positive part of a parabola. While this kernel is optimal in the asymptotic mean integrated squared error sense (under some conditions), it is indefinite (*i.e.,* non-positive definite kernel) [28]. Moreover, not all positive definite kernels are valid for kernel density estimation, such as projective kernels in general.

On the other hand, the coefficients α_i's in ML are arbitrary, while they are constant in kernel density estimation with a positive value[1] . Consequently, the derivations conducted in this paper and the theoretical analysis are more difficult to carry out than in the domain of gradient density estimation. For instance,

[1] A generalization of the density function was considered in [11] for arbitrary coefficients α_i's, leading to a generalized mode estimation, with the analysis conducted under several assumptions, such as the coefficients are uniformly bounded random variables.

Theorem 15 gives a sufficient condition for the Hessian to be positive definite, for any radial kernel and without any condition on the values of the coefficients α_i's. It turns out that this result boils down to Theorem D.4 in [8] when considering as a special case the Gaussian kernel associated to a normal distribution, under the unitarity axiom of the probability distributions, and constant nonnegative coefficients. In the same sense, Theorem 18 is more difficult to establish than Theorem 4 in [18]; however, our proof is straightforward using only conventional function calculus, as opposed to their proof that investigated dimensionality decomposition, change of variable and reparameterization, as well as geometry.

Appendix B. Theoretical Results for Projective Kernels

Projective kernels, also called inner-product kernels, are based on the inner product between samples in the input space. All these kernels are of the form

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = f(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle), \tag{B.1}$$

for some real function $f(\cdot)$ defined on real values. By analogy with the definitions carried out for the radial kernels, we shall call $f(\cdot)$ the profile of the kernel. Well-known nonlinear projective kernels with their expressions are given next.

Example 19 (Projective kernels)**.** *The most used projective kernels are the polynomial kernels defined as*

$$\kappa_p(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\tfrac{1}{\sigma}\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + c)^p, \tag{B.2}$$

for $\sigma > 0$, $p \in \mathbb{N}_+$ and $c \geq 0$ (also called homogeneous polynomial kernel when $c = 0$), and the exponential kernel defined as

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(\tfrac{1}{\sigma}\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle), \tag{B.3}$$

for some positive bandwidth parameter σ. Other valid kernels are the Vovk's real polynomial, of the form $\frac{1-\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle^p}{1-\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}$, as well as the Vovk's infinite polynomial $(1 - \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)^{-1}$. The Sigmoid kernel $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\frac{1}{\sigma}\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + c)$ is not a valid kernel in general (depending on the parameters (σ, c)). □

The following result is given in [7, Proposition 7.1].

Proposition 20 (Projective kernels [7])**.** *Three necessary conditions for a function $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = f(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)$ to be a positive definite kernel are, for any nonnegative u (namely $u = \|\boldsymbol{x}\|^2$ for all $\boldsymbol{x} \in \mathcal{X}$):*

$$f(u) \geq 0, \qquad f'(u) \geq 0, \qquad f'(u) + u f''(u) \geq 0.$$

In contrast with radial kernels, which have convex profiles as stated in Lemma 3, this is not the general case of projective kernels. However, there are many situations where the projective kernels have a convex profile as stated next, where the proof is straightforward by considering the nonnegativity of the second derivative of their profiles.

Proposition 21 (Projective kernels with convex profiles)**.** *The following projective kernels have a convex profile:*

- *The exponential kernel* (B.3)*.*
- *The polynomial kernel* (B.2) *for any even power $p \geq 2$.*
- *The polynomial kernel* (B.2) *for any odd power $p \geq 3$ with $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \geq -c\sigma$, such as $\mathcal{X}$ is in the positive orthant.*

The objective function of the pre-image problem (4) becomes

$$\Xi_{\mathrm{P}}(\boldsymbol{x}) = \frac{1}{2} f(\langle \boldsymbol{x}, \boldsymbol{x} \rangle) - \sum_{i=1}^{n} \alpha_i \, f(\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle). \tag{B.4}$$

Its gradient in terms of $f(\cdot)$ and its derivative $f'(\cdot)$ is

$$\nabla_{\boldsymbol{x}} \Xi_{\mathrm{P}}(\boldsymbol{x}) = f'(\langle \boldsymbol{x}, \boldsymbol{x} \rangle) \, \boldsymbol{x} - \sum_{i=1}^{n} \alpha_i \, f'(\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle) \, \boldsymbol{x}_i.$$

From this, we can provide the following fixed-point iterative technique:

$$\boldsymbol{x}^*_{t+1} = \frac{\sum_{i=1}^{n} \alpha_i \, f'(\langle \boldsymbol{x}^*_t, \boldsymbol{x}_i \rangle) \, \boldsymbol{x}_i}{f'(\langle \boldsymbol{x}^*_t, \boldsymbol{x}^*_t \rangle)}. \tag{B.5}$$

To the best of our knowledge, this general fixed-point iterative method is novel. The following special case was given in [25].

Example 22. *For the polynomial kernel, the fixed-point iterative method is*

$$\boldsymbol{x}_{t+1}^* = \frac{\sum_{i=1}^n \alpha_i \, \kappa_{p-1}(\boldsymbol{x}_i, \boldsymbol{x}_t^*) \, \boldsymbol{x}_i}{\kappa_{p-1}(\boldsymbol{x}_t^*, \boldsymbol{x}_t^*)}.$$

□

Besides first-order optimization using the gradient descent scheme or a fixed-point iterative technique as given in (B.5), we can also investigate second order optimization, by analogy with Section 4. To this end, we compute the Hessian matrix of the objective function (B.4), namely

$$\nabla_{\boldsymbol{x}}^2 \Xi_{\mathrm{P}}(\boldsymbol{x}) = f'(\langle \boldsymbol{x}, \boldsymbol{x} \rangle) \boldsymbol{I} + 2 f''(\langle \boldsymbol{x}, \boldsymbol{x} \rangle) \boldsymbol{x} \boldsymbol{x}^\top - \sum_{i=1}^n \alpha_i f''(\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle) \boldsymbol{x}_i \boldsymbol{x}_i^\top. \quad \text{(B.6)}$$

The following two theorems for projective kernels are equivalent to Theorems 14 and 15, with the proofs following a similar outline.

Theorem 23. *For projective kernels with a convex profile, a sufficient condition for the Hessian to be not positive definite is*

$$f'(\langle \boldsymbol{x}, \boldsymbol{x} \rangle) + 2 f''(\langle \boldsymbol{x}, \boldsymbol{x} \rangle) \, \|\boldsymbol{x}\|^2 - \sum_{\substack{i=1 \\ \alpha_i < 0}}^n \alpha_i \, f''(\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle) \, \|\boldsymbol{x}_i\|^2 < 0.$$

Proof. Since the profile of the kernel is convex, then its second derivative $f''(u)$ is nonnegative. Thus, we can rewrite the Hessian matrix (B.6) as

$$\nabla_{\boldsymbol{x}}^2 \Xi_{\mathrm{P}}(\boldsymbol{x}) = \nu_0 \, \boldsymbol{I} + 2 f''(\langle \boldsymbol{x}, \boldsymbol{x} \rangle) \, \boldsymbol{x} \boldsymbol{x}^\top - \sum_{i=1}^n \alpha_i \, \boldsymbol{r}_i \boldsymbol{r}_i^\top, \quad \text{(B.7)}$$

with $\nu_0 = f'(\langle \boldsymbol{x}, \boldsymbol{x} \rangle)$ and $\boldsymbol{r}_i = \sqrt{f''(\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle)} \, \boldsymbol{x}_i$, as well as the second term identified also as a rank-one update with the vector $\sqrt{2 f''(\langle \boldsymbol{x}, \boldsymbol{x} \rangle)} \, \boldsymbol{x}$. Lemma 13 leads to $\lambda_{\min}(\nabla_{\boldsymbol{x}}^2 \Xi_{\mathrm{P}}(\boldsymbol{x})) \leq \lambda_{\min}(\nu_0 \boldsymbol{I}) + 2 f''(\langle \boldsymbol{x}, \boldsymbol{x} \rangle) \|\boldsymbol{x}\|^2 - \sum_{\substack{i=1 \\ \alpha_i < 0}}^n \alpha_i \, \|\boldsymbol{r}_i\|^2$. Thus, by setting the right-hand-side to be strictly negative, we get the sufficient condition for having at least one negative eigenvalue. □

Theorem 24. *For projective kernels with a convex profile, a sufficient condition for the Hessian to be positive definite is*

$$\sum_{\substack{i=1 \\ \alpha_i > 0}}^n \alpha_i \, f''(\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle) \, \|\boldsymbol{x}_i\|^2 < f'(\langle \boldsymbol{x}, \boldsymbol{x} \rangle).$$

Proof. By following the proof of the previous theorem, applying Lemma 13 to the Hessian matrix (B.7) yields $\lambda_{\min}(\nu_0 \boldsymbol{I}) - \sum_{\substack{i=1 \\ \alpha_i>0}}^{n} \alpha_i \|\boldsymbol{r}_i\|^2 \leq \lambda_{\min}(\nabla^2_{\boldsymbol{x}} \Xi_{\mathrm{P}}(\boldsymbol{x}))$. Thus, a sufficient condition for positive eigenvalues of the Hessian is obtained by imposing the positivity of the left-hand-side. □

While Theorem 24 is restricted to projective kernels with a convex profile $f(\cdot)$ on $\mathcal{X}$, it turns out that this condition is often satisfied as given in Proposition 21. For example, when dealing with the quadratic homogeneous kernel, the sufficient condition boils down to $\sum_{\substack{i=1 \\ \alpha_i>0}}^{n} \alpha_i \|\boldsymbol{x}_i\|^2 < \|\boldsymbol{x}\|^2$.

By analogy with Theorem 16, we can provide an equivalence between Newton's method and the fixed-point iteration technique under some conditions.

Theorem 25. *For projective kernels, if $f(\cdot)$ has a piecewise-constant derivative, then the fixed-point iteration* (B.5) *is equivalent to a Newton update* (18)*.*

Proof. For piecewise-constant $f'(\cdot)$, $f''(\cdot)$ vanishes from the expression of the Hessian, yielding $\nabla^2_{\boldsymbol{x}} \Xi_{\mathrm{P}}(\boldsymbol{x}) = f'(\langle \boldsymbol{x}, \boldsymbol{x} \rangle) \boldsymbol{I}$. Thus, the Newton update becomes

$$\boldsymbol{x}^*_{t+1} = \boldsymbol{x}^*_t - \left(f'(\langle \boldsymbol{x}^*_t, \boldsymbol{x}^*_t \rangle) \boldsymbol{I}\right)^{-1} \nabla_{\boldsymbol{x}^*_t} \Xi_{\mathrm{P}}(\boldsymbol{x}^*_t) = \frac{\sum_{i=1}^{n} \alpha_i \, f'(\langle \boldsymbol{x}^*_t, \boldsymbol{x}_i \rangle) \, \boldsymbol{x}_i}{f'(\langle \boldsymbol{x}^*_t, \boldsymbol{x}^*_t \rangle)},$$

which corresponds to a step in the fixed-point method (B.5). □

Theorem 17 can be recast for projective kernels as follows, following the same outlines of the proof of the former, as well as the resulting discussions.

Theorem 26. *For projective kernels, the fixed-point iteration* (B.5) *at a guess $\boldsymbol{x}^*_t$ seeks the optimum of the quadratic function*

$$q(\boldsymbol{x}) = \tfrac{1}{2} f'(\langle \boldsymbol{x}^*_t, \boldsymbol{x}^*_t \rangle) \|\boldsymbol{x}\|^2 - \sum_{i=1}^{n} \alpha_i \, f'(\langle \boldsymbol{x}^*_t, \boldsymbol{x}_i \rangle) \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle - C(\boldsymbol{x}^*_t),$$

where $C(\boldsymbol{x}^*_t) = -\frac{1}{2} f(\langle \boldsymbol{x}^*_t, \boldsymbol{x}^*_t \rangle) + \sum_{i=1}^{n} \alpha_i \, f(\langle \boldsymbol{x}^*_t, \boldsymbol{x}_j \rangle) + f'(\langle \boldsymbol{x}^*_t, \boldsymbol{x}^*_t \rangle) \|\boldsymbol{x}^*_t\|^2 - \sum_{i=1}^{n} \alpha_i \, f'(\langle \boldsymbol{x}^*_t, \boldsymbol{x}_i \rangle) \langle \boldsymbol{x}^*_t, \boldsymbol{x}_i \rangle$ is independent of $\boldsymbol{x}$. Furthermore, the quadratic function $q(\cdot)$ is tangent to $\Xi_P(\cdot)$ at $\boldsymbol{x}^*_t$.

Finally, we recast Theorem 18 for the projective kernels, by showing that the fixed-point iteration technique is an MM algorithm.

Theorem 27. *For nonnegative coefficients α_i's and projective kernels with convex profile, the fixed-point iterative technique is an MM algorithm, where the quadratic function $q(\cdot)$ defined in Theorem 26 is the surrogate function that majorizes the objective function* (B.4).

Proof. The proof follows the same guidelines as the proof of Theorem 18. For any guess $\boldsymbol{x}_t^*$, $\Xi_{\mathrm{P}}(\boldsymbol{x}_t^*) = q(\boldsymbol{x}_t^*)$, and $\Xi_{\mathrm{P}}(\boldsymbol{x}) \leq q(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{X}$. Since $f(\cdot)$ is convex, we use $f(u) \geq f(v) + f'(v)(u - v)$. On one hand, we get by substituting u with $\langle \boldsymbol{x}_t^*, \boldsymbol{x}_t^* \rangle$ and v with $\langle \boldsymbol{x}, \boldsymbol{x} \rangle$ the following inequality:

$$f(\langle \boldsymbol{x}, \boldsymbol{x} \rangle) \leq f(\langle \boldsymbol{x}_t^*, \boldsymbol{x}_t^* \rangle) - f'(\langle \boldsymbol{x}_t^*, \boldsymbol{x}_t^* \rangle)(\langle \boldsymbol{x}, \boldsymbol{x} \rangle - \langle \boldsymbol{x}_t^*, \boldsymbol{x}_t^* \rangle).$$

On the other hand, substituting u with $\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle$ and v with $\langle \boldsymbol{x}_t^*, \boldsymbol{x}_i \rangle$ for any $\boldsymbol{x}$, $\boldsymbol{x}_t^*$ and $\boldsymbol{x}_i$, we get

$$f(\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle) \geq f(\langle \boldsymbol{x}_t^*, \boldsymbol{x}_i \rangle) + f'(\langle \boldsymbol{x}_t^*, \boldsymbol{x}_i \rangle)(\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle - \langle \boldsymbol{x}_t^*, \boldsymbol{x}_i \rangle).$$

By combining this inequality for all $i = 1, 2 \ldots, n$, with some nonnegative coefficients α_i's, with the previous inequality, then the inequality holds also for any combination, we get $\Xi_{\mathrm{P}}(\boldsymbol{x}) \leq q(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$. Finally, the optimal solution of the surrogate function $q(\boldsymbol{x})$ is obtained by nullifying its gradient, which corresponds to the fixed-point iteration. This concludes the proof. □

References

[1] Abedsoltan, A., Belkin, M., Pandit, P., 2023. Toward large kernel models, in: International Conference on Machine Learning, PMLR. pp. 61–78.

[2] Aliyari Ghassabeh, Y., 2013. On the convergence of the mean shift algorithm in the one-dimensional space. Pattern Recognition Letters 34, 1423–1427.

[3] Aliyari Ghassabeh, Y., 2015. A sufficient condition for the convergence of the mean shift algorithm with Gaussian kernel. Journal of Multivariate Analysis 135, 1–10.

[4] Arias-Castro, E., Mason, D., Pelletier, B., 2016. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. Journal of Machine Learning Research 17, 1–28.

[5] Botsch, M., Nossek, J.A., 2008. Construction of interpretable radial basis function classifiers based on the random forest kernel, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 220–227.

[6] vor der Brück, T., Eger, S., Mehler, A., 2015. Complex decomposition of the negative distance kernel, in: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), IEEE. pp. 103–108.

[7] Burges, C., 1999. Geometry and invariance in kernel based methods. Advances in kernel methods: support vector learning , 89–116.

[8] Carreira-Perpinan, M., 2000. Mode-finding for mixtures of gaussian distributions. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 1318–1323.

[9] Carreira-Perpinan, M.A., 2007. Gaussian mean-shift is an EM algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 767–776.

[10] Celikkanat, A., Shen, Y., Malliaros, F., 2022. Multiple kernel representation learning on networks. IEEE Trans. on Knowledge and Data Engineering .

[11] Chen, Y.C., Genovese, C.R., Wasserman, L., 2014. Generalized mode and ridge estimation. arXiv preprint arXiv:1406.1803 .

[12] Cheng, Y., 1995. Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 17, 790–799.

[13] Comaniciu, D., Meer, P., 2002. Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 603–619.

[14] Cucker, F., Smale, S., 2002. On the mathematical foundations of learning. Bulletin of the American Mathematical Society 39, 1–49.

[15] El Ahmad, T., Brogat-Motte, L., Laforgue, P., d'Alché Buc, F., 2024. Sketch in, sketch out: Accelerating both learning and inference for structured prediction with kernels, in: Dasgupta, S., Mandt, S., Li, Y. (Eds.), Proc. of The 27th International Conference on Artificial Intelligence and Statistics, pp. 109–117.

[16] Esser, P., Fleissner, M., Ghoshdastidar, D., 2024. Non-parametric representation learning with kernels, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 11910–11918.

[17] Fan, J., Chow, T.W., 2018. Non-linear matrix completion. Pattern Recognition 77, 378–394.

[18] Fashing, M., Tomasi, C., 2005. Mean shift is a bound optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 471–474.

[19] Fukunaga, K., Hostetler, L., 1975. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory 21, 32–40.

[20] Golub, G.H., Van Loan, C.F., 1996. Matrix computations. Third ed., JHU press.

[21] He, M., He, F., Shi, L., Huang, X., Suykens, J.A., 2023. Learning with asymmetric kernels: Least squares and feature interpretation. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–12.

[22] Honeine, P., Richard, C., 2011. Preimage problem in kernel-based machine learning. IEEE Signal Processing Magazine 28, 77 – 88.

[23] Jia, L., Gaüzère, B., Honeine, P., 2021. A graph pre-image method based on graph edit distances, in: Torsello, A., et al. (Eds.), Proc. IAPR Joint

International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (S+SSPR), Springer International Publishing, Venice, Italy. pp. 216–226.

[24] Jia, L., Ning, X., Gaüzère, B., Honeine, P., Riesen, K., 2023. Bridging distinct spaces in graph-based machine learning, in: Blumenstein, M., Lu, H., Yang, W., Cho, S.B. (Eds.), Proceedings of the 7th Asian Conference on Pattern Recognition (ACPR), Kitakyushu, Japan.

[25] Kwok, J.T., Tsang, I.W., 2003. The pre-image problem in kernel methods, in: Proc. 20th International Conference on Machine Learning, AAAI Press, Washington, DC, USA. pp. 408–415.

[26] Li, X., Hu, Z., Wu, F., 2007. A note on the convergence of the mean shift. Pattern Recogn. 40, 1756–1762.

[27] Mika, S., Schölkopf, B., Smola, A., Müller, K.R., Scholz, M., Rätsch, G., 1999. Kernel PCA and de-noising in feature spaces, in: Proc. Conf. on Advances in Neural Information Processing Systems II, pp. 536–542.

[28] Ong, C.S., Mary, X., Canu, S., Smola, A.J., 2004. Learning with non-positive kernels, in: Proc. 21st International Conference on Machine Learning, p. 81.

[29] Pandey, A., De Meulemeester, H., De Moor, B., Suykens, J.A., 2023. Multi-view kernel PCA for time series forecasting. Neurocomputing 554, 126639.

[30] Pandey, A., Schreurs, J., Suykens, J.A., 2021. Generative restricted kernel machines: a framework for multi-view generation and disentangled feature learning. Neural Networks 135, 177–191.

[31] Salazar, D., Rios, J., Aceros, S., Flórez-Vargas, O., Valencia, C., 2021. Kernel joint non-negative matrix factorization for genomic data. IEEE Access 9, 101863–101875.

[32] Schölkopf, B., 2000. The kernel trick for distances. Advances in neural information processing systems 13.

[33] Schölkopf, B., Mika, S., Burges, C., Knirsch, P., Muller, K.R., Ratsch, G., Smola, A., 1999. Input space versus feature space in kernel-based methods. IEEE Transactions on Neural Networks 10, 1000–1017.

[34] Shankar, V., Fang, A., Guo, W., Fridovich-Keil, S., Schmidt, L., Ragan-Kelley, J., Recht, B., 2020. Neural kernels without tangents, in: Proceedings of the 37th International Conference on Machine Learning, JMLR.org. pp. 1–10.

[35] Tax, D.M.J., Juszczak, P., 2003. Kernel whitening for one-class classification. International Journal of Pattern Recognition and Artificial Intelligence 17, 333–347.

[36] Tran Thi Phuong, T., Douzal, A., Yazdi, S.V., Honeine, P., Gallinari, P., 2020. Interpretable time series kernel analytics by pre-image estimation. Artificial Intelligence 286, 103342.

[37] Unser, M., 2021. A unifying representer theorem for inverse problems and machine learning. Foundations of Computational Mathematics 21, 941–960.

[38] Yamasaki, R., Tanaka, T., 2020. Properties of mean shift. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 2273–2286.

[39] Yamasaki, R., Tanaka, T., 2024. Convergence analysis of mean shift. IEEE Transactions on Pattern Analysis and Machine Intelligence (in press), 1–11.

[40] Zhu, F., Honeine, P., 2017. Online kernel nonnegative matrix factorization. Signal Processing 131, 143 – 153.