Effective regions and kernels in continuous sparse regularisation, with application to sketched mixtures

Yohann De Castro
Institut Camille Jordan - IUF - Centrale Lyon

Rémi Gribonval
Inria, CNRS, ENS de Lyon, Université Claude Bernard Lyon 1, LIP, UMR 5668, 69342, Lyon cedex 07, France

Nicolas Jouvin
INRAE, AgroParisTech, Université Paris-Saclay, MIA-Paris, 91120, Palaiseau

September 4, 2025

Abstract

This paper advances the general theory of continuous sparse regularisation on measures with the Beurling-LASSO (BLASSO). This TV-regularized convex program on the space of measures allows to recover a sparse measure using a noisy observation from an appropriate measurement operator. While previous works have uncovered the central role played by this operator and its associated kernel in order to get estimation error bounds, the latter requires a technical local positive curvature (LPC) assumption to be verified on a case-by-case basis. In practice, this yields only few LPC-kernels for which this condition is proved. At the heart of our contribution lies the *kernel switch*, which uncouples the *model* kernel from the LPC assumption: it enables to leverage any known LPC-kernel as a *pivot* kernel to prove error bounds, provided embedding conditions are verified between the model and pivot RKHS. We increment the list of LPC-kernels, proving that the "sinc-4" kernel, used for signal recovery and mixture problems, does satisfy the LPC assumption. Furthermore, we also show that the BLASSO localisation error around the true support decreases with the noise level, leading to *effective* near regions. This improves on known results where this error is fixed with some parameters depending on the model kernel. We illustrate the interest of our results in the case of translation-invariant mixture model estimation, using bandlimiting smoothing and sketching techniques to reduce the computational burden of BLASSO.

***K*eywords** Continuous sparse regression, Beurling Lasso, Sketching, Mixture models

Contents

1 Introduction

Sparse regularisation, particularly methods based on ℓ^1 minimisation, has emerged as a cornerstone technique to solve high-dimensional linear inverse problems across numerous domains. These approaches have revolutionized signal processing, statistical learning, and computational imaging by enabling reconstruction from severely limited observations (Foucart and Rauhut 2013; Candès et al. 2006). Despite their remarkable success, traditional sparse methods rely on discretisation, introducing inherent limitations in both computational efficiency and theoretical precision. Consequently, there has been growing interest in continuous counterparts to ℓ^1 minimisation that operate in a grid-free manner. These "off-the-grid" approaches not only mitigate discretisation errors but also potentially offer more efficient solvers, sharper theoretical guarantees, and increased robustness to sampling limitations see *e.g.* Candès and Fernandez-Granda 2014; Tang et al. 2013; Duval and Peyré 2015; Duval and Peyré 2017. Such continuous formulations have proved particularly valuable when the underlying target structure inherently exists in a continuous domain, such as in the case of mixtures of distributions or when the parameters of interest are not naturally discretized.

A promising extension of ℓ^1 minimisation to continuous domains involves encoding both positions and weights as a discrete Radon measure, replacing the ℓ^1 norm with the total variation norm. This approach treats a measure as "sparse" when it consists of a finite sum of few Dirac masses. The resulting infinite-dimensional optimisation problem is cast over the space of measures and known as Beurling-LASSO (BLASSO). The latter has emerged as a powerful framework across multiple disciplines. In signal processing, BLASSO has demonstrated effectiveness in sparse deconvolution and super-resolution (Candès and Fernandez-Granda 2014; Duval and Peyré 2015; Boyer et al. 2017). Its applications extend to statistics for continuous sparse regression and off-the-grid compressed sensing (De Castro and Gamboa 2012; Poon et al. 2023), and to machine learning for mixture model estimation (De Castro et al. 2021). Practical implementations have yielded significant advances in diverse fields: single-molecule fluorescent microscopy (Boyd et al. 2017), neuronal spike sorting (Ekanadham et al. 2014), statistical mixture modeling (Gribonval et al. 2020; De Castro et al. 2021), and training of shallow neural networks (Bach 2017). In this paper, we motivate our analysis and illustrate our results in the case of mixture model estimation (De Castro et al. 2021), although our theoretical contributions apply broadly to all BLASSO applications.

Our main contributions are the four guarantees presented on top of Figure 1 which apply to the general formulation of the BLASSO with applications to mixture models, continuous sparse regression, neural network training and signal processing, among others. Figure 1 depicts various concepts which appeared in recent studies of BLASSO:

- **Sketching** is a powerful technique that allows for the reduction of the dimensionality of data while preserving essential information (Gribonval et al. 2021a). In the context of BLASSO, sketching can be

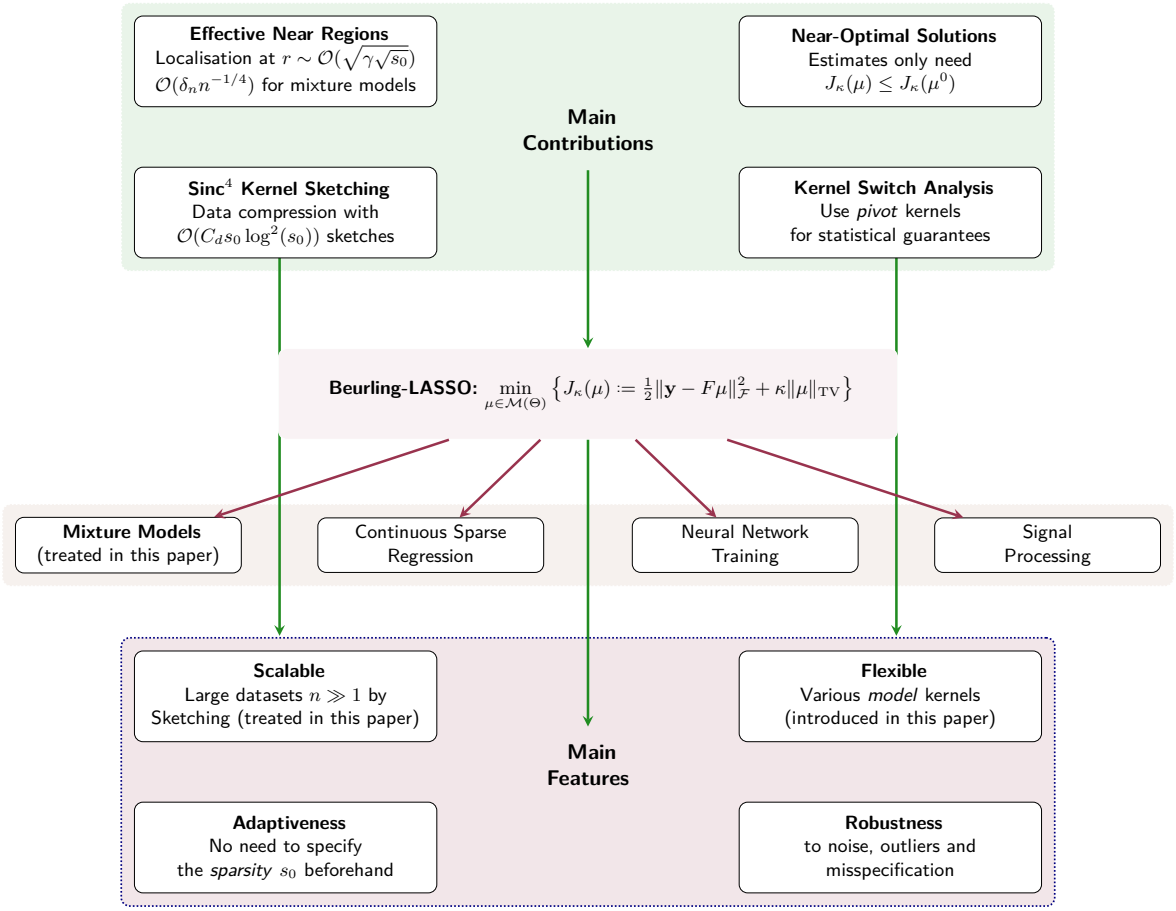

Figure 1: The figure summarizes the main contributions and applications of the paper, highlighting the effective near regions, kernel switch analysis, near-optimal solutions, and sinc kernel sketching. It also emphasizes the scalability, flexibility, robustness, and adaptive nature of the proposed approach in continuous sparse regression problems. Further details can be found in Section 1.3. In this figure, the *model kernel* is the kernel associated to the measurement operator F and the *pivot kernel* is the kernel used to prove the error bounds. The latter can be different from the model kernel, as shown in this paper. The notation C_d is a constant that depends on the dimension d of the parameter space $\mathcal{X}$.

employed to efficiently handle large datasets, making it feasible to apply continuous sparse regularisation even when the sample size is significantly larger than the number of parameters. By leveraging sketching techniques, we can achieve computational efficiency without sacrificing the quality of the estimates (Poon et al. 2023). This is particularly relevant in modern applications where data sizes are rapidly increasing.

- **Adaptiveness** refers to the ability of a method to automatically adjust its complexity based on the underlying data structure. In the context of BLASSO, adaptiveness means that the method can effectively identify and exploit the sparsity of the underlying model without requiring prior knowledge of the number of components or their locations (Duval and Peyré 2015; De Castro et al. 2021). This is crucial in practical scenarios where the true model may not be known a priori, and it allows for more flexible modeling of complex data structures.

- **Robustness** is a key consideration in learning methods, especially when dealing with real-world data that may contain noise, outliers, or other forms of contamination. The BLASSO framework is designed to be robust to various types of perturbations, including noise, outliers and misspecification to non-sparse models, the reader may refer to Poon et al. (2023) for further details.

The combination of sketching, adaptiveness, and robustness makes BLASSO a powerful tool for continuous sparse regularisation in a wide range of applications.

1.1 Continuous sparse regularisation with BLASSO

Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact set, referred to as the *parameter space*, and consider $(\mathcal{M}(\mathcal{X}), \|\cdot\|_{TV})$ the space of Radon measures, defined as the topological dual space of $(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$, the continuous functions endowed with the infinity norm. Let $\mathcal{F}$ be a separable Hilbert space and let $F : \mathcal{M}(\mathcal{X}) \to \mathcal{F}$ be a linear map, referred to as the forward measurement operator. We define the BLASSO problem as

$$\widehat{\mu} \in \operatorname*{arg\,min}_{\mu \in \mathcal{M}(\mathcal{X})} J_\kappa(\mu) \quad \text{where} \quad J_\kappa(\mu) \coloneqq \frac{1}{2}\|\boldsymbol{y} - F\mu\|_{\mathcal{F}}^2 + \kappa\|\mu\|_{\mathrm{TV}}\,, \tag{1.1–J_κ}$$

with $\kappa > 0$ a regularisation parameter and $\boldsymbol{y} \in \mathcal{F}$ some observation. One can prove that a solution $\widehat{\mu} \in \mathcal{M}(\mathcal{X})$ to BLASSO (1.1–J_κ) exists under the forthcoming hypothesis (1.3–$\mathbf{H}_{\mathsf{mod}}$), see for instance De Castro et al. 2023, Theorem 1.1.

The criterion to be minimized decomposes as the sum of two terms: a *data-fitting* term, comparing some observation $\boldsymbol{y}$ with a candidate $F\mu$ in the separable Hilbert space $\mathcal{F}$; and *TV-regularisation* term, which acts as a sparsity inducing penalty analogous to ℓ^1-regularisation in finite dimension (Foucart and Rauhut 2013; Boyer et al. 2019, Theorem 1.1).

Class of sparse models and inverse problem under consideration

The continuous regression problem under consideration offers a powerful generalisation of classical sparse regression, encoding both weights and positions within a Radon measure framework. Central to our analysis is the *model kernel* K_{mod}, which characterizes the interaction between point masses in the parameter space $\mathcal{X}$ through the measurement operator F. Although it is traditionally referred to simply as "the kernel" in the BLASSO literature (Poon et al. 2023), we choose to deliberately distinguish it from what we will call the *pivot kernel* later in our analysis.

The model kernel K_{mod} is formally defined as the $\mathcal{F}$-inner product of the operator F applied to pairs of Dirac masses at different locations in the parameter space $\mathcal{X}$. Its continuity is a mild technical condition which ensures the existence of solutions to the BLASSO optimisation problem (De Castro et al. 2023). This distinction between *model kernel* and *pivot kernel* represents one of our key contributions, allowing us to extend theoretical guarantees to a broader class of inverse problems than previously possible.

Definition 1 (Model kernel)**.** *Define the model kernel by*

$$K_{\mathsf{mod}}(\boldsymbol{s}, \boldsymbol{t}) \coloneqq \langle F\delta_{\boldsymbol{s}}, F\delta_{\boldsymbol{t}}\rangle_{\mathcal{F}} \quad \text{for all } \boldsymbol{s}, \boldsymbol{t} \in \mathcal{X}\,, \tag{1.2}$$

where $\delta_{\boldsymbol{u}}$ is the Dirac mass at point $\boldsymbol{u} \in \mathcal{X}$. We assume that

$$K_{\mathsf{mod}}(\cdot,\cdot) \text{ is continuous,} \tag{1.3–$\mathbf{H}_{\mathsf{mod}}$}$$

where continuity is understood with respect to the pair of variables. It is a positive definite kernel and we denote as $\mathcal{H}_{\mathsf{mod}}$ its reproducing kernel Hilbert space (RKHS), which is continuously included in $\mathcal{F}$ as discussed in Section 3.

In the following, the class of models $\mathcal{M}_{s,\Delta,\mathfrak{d}}$ is defined as the set of measures that can be written as a sum of s Dirac masses with a *minimal separation* of Δ between two spikes for some distance $\mathfrak{d}$; the noise term Γ is defined as the difference between the observation $\boldsymbol{y}$ and the operator F applied to the *target* measure μ^0; and the *noise level* γ is defined as the norm of the noise term.

Definition 2 (Sparse models)**.** *Let $\mathfrak{d}(\cdot,\cdot)$ be a distance defined on $\mathcal{X}$. Let Δ be positive and let s be greater than 1. We define the class of s-sparse measures with minimal separation Δ as*

$$\mathcal{M}_{s,\Delta,\mathfrak{d}} \coloneqq \Big\{\mu \,:\, \mu = \sum_{k=1}^{s} a_k \delta_{\boldsymbol{x}_k} \text{ and } \min_{k,\ell} \mathfrak{d}(\boldsymbol{x}_k, \boldsymbol{x}_\ell) \geq \Delta\Big\}\,, \tag{1.4}$$

where Δ is referred to as the minimal separation, and s as the sparsity.

Remark 1.1. *The class of sparse models is usually defined with respect to the so-called* Fisher-Rao *distance associated to some positive type kernel $K(\cdot,\cdot) : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$. While previous works choose this kernel to be the model kernel (Poon et al. 2019; Poon et al. 2023), this paper shows error bounds choosing a possibly different kernel.*

The inverse problem under consideration is defined as follows. We consider a target measure $\mu^0 \in \mathcal{M}_{s_0,\Delta_0,\mathfrak{d}}$, for some sparsity s_0, distance $\mathfrak{d}$ and minimal separation Δ_0, and we define the noise term and noise level respectively as

$$\Gamma \coloneqq \boldsymbol{y} - F\mu^0\,, \tag{1.5a}$$

$$\text{resp. } \gamma \coloneqq \|\Gamma\|_{\mathcal{F}}\,. \tag{1.5b}$$

Given the noisy observation $\boldsymbol{y}$, our goal is to estimate the true positions $\boldsymbol{x}_k^0$ and true weights a_k^0 from a solution $\widehat{\mu}$ to BLASSO (1.1–J_κ). The error term Γ can capture the error due to model misspecification, the presence of outliers, and noise (either from sampling or measurement, see De Castro et al. 2021; Poon et al. 2023).

Estimation error bounds via dual certificates

The analysis of BLASSO estimation error bounds (1.7) below hinges on the concept of the so-called *far* and *near* regions, which define spatial partitioning around the support points of the target measure μ^0. These regions serve a fundamental role in characterizing the behaviour of any solution $\widehat{\mu}$ to (1.1–J_κ). Specifically, the BLASSO exhibits strong reconstruction properties in near regions—where $\widehat{\mu}$ closely approximates μ^0—while maintaining controlled behaviour in far regions. This spatial characterisation provides precise quantitative guarantees on the support recovery capabilities of the estimator, referred to as the localisation property of BLASSO. We formally define these regions as follows.

Definition 3 (Far and Near regions)**.** *Let $\mu^0 \in \mathcal{M}_{s_0,\Delta_0,\mathfrak{d}}$, let $r > 0$, we define the near region of $\boldsymbol{x}_k^0$ as the closed ball centered at $\boldsymbol{x}_k^0$ for the distance $\mathfrak{d}$*

$$\mathcal{N}_k^{\text{reg}}(r) \coloneqq \left\{\boldsymbol{x} \in \mathcal{X},\quad \mathfrak{d}(\boldsymbol{x}, \boldsymbol{x}_k^0) \leq r\right\}, \tag{1.6a}$$

and the far region as the complement

$$\mathcal{F}^{\text{reg}}(r) \coloneqq \mathcal{X} \setminus \mathcal{N}^{\text{reg}}(r), \quad \text{with: } \mathcal{N}^{\text{reg}}(r) = \bigcup_{k=1}^{s_0} \mathcal{N}_k^{\text{reg}}(r). \tag{1.6b}$$

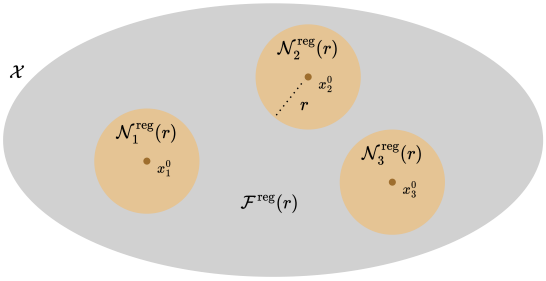

Figure 2: Two-dimensional illustration of near (light marroon) and far (light gray) regions with fixed radius r and a distance $\mathfrak{d}(\boldsymbol{s}, \boldsymbol{t}) \propto \|\boldsymbol{s} - \boldsymbol{t}\|^2$.

Existing bounds on estimation errors proved in the literature (Azais et al. 2015; Candès and Fernandez-Granda 2014; Poon et al. 2023) depend on some fixed $r > 0$, that may depend on Δ_0, and are expressed as follows:

- *Control of the far region:*
$$|\widehat{\mu}|\left(\mathcal{F}^{\text{reg}}(r)\right) \lesssim_d \gamma\sqrt{s_0}\,, \tag{1.7a}$$
- *Control of all the near regions:*
$$\left|\widehat{\mu}\left(\mathcal{N}_k^{\text{reg}}(r)\right) - a_k^0\right| \lesssim_d \gamma\sqrt{s_0}\,, \tag{1.7b}$$
- *Detection level:* For all borelian $A \subset \mathcal{X}$ such that $|\widehat{\mu}|(A) \gtrsim_d \gamma\sqrt{s_0}$, there exists $\boldsymbol{x}_k^0$ such that
$$\mathfrak{d}(A, \boldsymbol{x}_k^0) \coloneqq \min_{\boldsymbol{t} \in A} \mathfrak{d}(t, \boldsymbol{x}_k^0) \lesssim_d r\,, \tag{1.7c}$$

where $\lesssim_d$ denotes the inequality up to a multiplicative constant that may depend on the dimension d, $\widehat{\mu}$ is a solution to (1.1–J_κ) with regularisation parameter $\kappa \simeq \gamma/\sqrt{s_0}$ and $|\widehat{\mu}|$ denotes the absolute part of $\widehat{\mu}$.

Remark 1.2 (On regularisation parameter calibration)**.** *The regularisation parameter κ is crucial for the performance of BLASSO. In practice, it is often calibrated using cross-validation techniques or other model selection criteria. The choice of κ can significantly impact the balance between bias and variance in the estimation process. In our analysis, we assume that κ is chosen to be proportional to the noise level γ and the sparsity s_0, specifically $\kappa \simeq \gamma/\sqrt{s_0}$. This choice ensures that the regularisation term effectively controls the trade-off between fitting the data and promoting sparsity in the estimated measure.*

When s_0 is not known, other choices of the regularisation parameter can be considered, such as $\kappa \simeq \gamma$ which is a more realistic choice in practice, yet less optimal in terms of the theoretical guarantees. In this case, the error bounds (1.7) scale as s_0 instead of $\sqrt{s_0}$, which is a less favorable scaling, yet still provides useful information about the localisation of the solution, see Remark 3.2 below.

This error analysis of BLASSO (1.1–J_κ) relies on controlling both the noise level γ and the Bregman divergence of the total variation norm

$$\mathcal{D}_{\eta^0}(\mu \,||\, \mu^0) := \|\mu\|_{\mathrm{TV}} - \|\mu^0\|_{\mathrm{TV}} - \langle \eta^0, \mu - \mu^0 \rangle_{\mathcal{C}(\mathcal{X})\times\mathcal{M}(\mathcal{X})} \tag{1.8}$$

between the underlying $\mu^0 \in \mathcal{M}_{s_0,\Delta_0,\mathfrak{d}}$ and a solution $\mu = \widehat{\mu}$ (De Castro and Gamboa 2012; Azais et al. 2015) or, as we will see, any *approximate* solution to (1.1–J_κ). This error analysis requires that

- η^0 is an element of the RKHS $\mathcal{H}_{\mathsf{mod}}$,
- η^0 is a sub-gradient of the total variation norm at point μ^0 with specific properties on the near regions $\mathcal{N}_k^{\mathsf{reg}}(r)$ and far region $\mathcal{F}^{\mathsf{reg}}(r)$,

Such η^0 is referred to as a *non-degenerate dual certificate of radius r* (De Castro and Gamboa 2012; Candès and Fernandez-Granda 2014; Duval and Peyré 2015). When such a certificate exists, the control on Bregman divergence directly translates into the estimation error bounds of Equations (1.7), yielding precise guarantees on both support recovery and parameter estimation accuracy (Duval and Peyré 2017).

Construction of dual certificates under local positive curvature assumption (LPC)

Crucially, a sufficient condition for BLASSO error analysis to hold is that the model kernel verifies a so-called *local positive curvature* assumption (**LPC**, Poon et al. 2023, Assumption 1), given in Section 3.1. The **LPC** ensures the existence of a non-degenerate dual certificate η^0 of radius r for any target measure μ^0 in the sparse model class $\mathcal{M}_{s_0,\Delta_0,\mathfrak{d}_{\mathsf{FR}}}$ where Δ_0 is taken sufficiently large with respect to r. There, the distance $\mathfrak{d}_{\mathsf{FR}}$ is the so-called *Fisher-Rao distance* induced by the model kernel, which we will detail later on in Section 3.1.

To the best of our knowledge, the literature (Azais et al. 2015; Candès and Fernandez-Granda 2014; Poon et al. 2019; Poon et al. 2023) has proved that the estimation error bounds of Equations 1.7 hold, but only for *fixed* near regions of size $r = r_0$ *irrespective of the noise level*. However, in view of of Equations (1.7), the smaller the radius r, the closer the BLASSO localizes around the true support, hence the sharper the error bounds. One of our key contributions is showing that *these bounds extend to near regions whose size can adapt to the noise level γ, decreasing as $\mathcal{O}(\sqrt{\gamma\sqrt{s_0}})$*, which we call *effective* near regions. This is a substantial improvement for small noise levels since the **LPC** is a property *intrinsic* to the model kernel, with parameters independent of μ^0 and $\boldsymbol{y}$, and thus of the noise level γ.

Previous studies on kernels for BLASSO

Fortunately, the existence of dual certificates has alreadt been proved for some kernels. Some of these have also been shown to verify the **LPC** assumption. Pioneering work (Candès and Fernandez-Granda 2014) has proved that the non-degenerate dual certificate exist for the Jackson kernel:

- **Jackson kernel** (Candès and Fernandez-Granda 2014) on $\mathcal{X} = [0,1)^d$ identified to the d-dimensional torus $(\mathbb{R}\setminus\mathbb{Z})^d$. The kernel is the fourth power of the Dirichlet kernel, i.e.

$$K(\boldsymbol{s},\boldsymbol{t}) = \prod_{i=1}^{d} \kappa(s_i - t_i), \quad \text{with } \kappa(t) = \left(\frac{\sin\left(\left(\frac{N}{2}+1\right)\pi t\right)}{\left(\frac{N}{2}+1\right)\sin(\pi t)}\right)^4 \text{ and } N \geq 1\,. \tag{1.9a}$$

  This kernel satisfies the **LPC** (Poon et al. 2023, Appendix C).

Using a similar technique as in Candès and Fernandez-Granda 2014, a low-pass spherical harmonic kernel has been investigated in Bendory et al. 2015 for *spike detection* on the sphere.

- **Low-pass spherical harmonic** (Bendory et al. 2015) on the 2-sphere $\mathcal{X} = \mathbb{S}^2 \subset \mathbb{R}^3$, given by

$$K(\boldsymbol{s}, \boldsymbol{t}) = C(N) \sum_{k=0}^{N} \zeta(k/N) P_{k,3}(\langle \boldsymbol{s}, \boldsymbol{t} \rangle)\,,$$

  where $N \geq 1$ is some degree, $C(N)$ is a normalizing constant, $P_{k,3}$ is the univariate ultraspherical Gegenbauer polynomial of order 3 and degree k, and $\zeta(t)$ is a smooth non-negative univariate function equal to 1 for t between 0 and $1/2$, 0 for t greater than 1 and less than 1 otherwise.
  The **LPC** has not been proved for this kernel although the existence of a dual certificate has been proven via alternative techniques.

Recently, two others kernels have been proved to satisfy the **LPC**, namely

- **Gaussian kernel** (Poon et al. 2023, Appendix D) given by

$$K(\boldsymbol{s}, \boldsymbol{t}) = \Theta_{\boldsymbol{\Omega}}(\boldsymbol{s} - \boldsymbol{t}) \quad \text{where} \quad \Theta_{\boldsymbol{\Omega}} := \exp\big(-\frac{1}{2}\|\cdot\|_{\Omega}^2\big)\,, \tag{1.9b}$$

  and $\|\boldsymbol{t}\|_{\boldsymbol{\Omega}}^2 = \langle \boldsymbol{t}, \boldsymbol{\Omega}^{-1}\boldsymbol{t} \rangle$ for some positive definite matrix $\boldsymbol{\Omega}$;

- **Laplace transform kernel** (Poon et al. 2023, Appendix E). This kernel is defined on $\mathcal{X} = \mathbb{R}_+^d$ and arises when studying continuous sampling of the Laplace transform of a positive signal, defined as $\mathcal{L}[\mu](\boldsymbol{\omega}) = \int_{\mathcal{X}} e^{-\boldsymbol{\omega}^\top \boldsymbol{t}}\, \mathrm{d}\mu(\boldsymbol{t})$. It was studied in Denoyelle et al. (2019) and Poon et al. (2019) and is not to be confused with the *Laplace kernel*. In particular, it is not translation-invariant and is given by

$$K_\alpha(\boldsymbol{s}, \boldsymbol{t}) = \prod_{i=1}^{d} \kappa(s_i + \alpha_i, t_i + \alpha_i), \quad \text{with } \kappa(a, b) = \frac{2\sqrt{ab}}{a+b} \text{ and } \alpha_i > 0. \tag{1.9c}$$

In this paper, we will prove in Theorem 3 that the fourth power of the sinus cardinal kernel, coined sinc-4, also satisfies the **LPC**.

- **Sinc-4 kernel**

$$K(\boldsymbol{s}, \boldsymbol{t}) = \Psi_\tau(\boldsymbol{t} - \boldsymbol{s}) \quad \text{where} \quad \Psi_\tau(\boldsymbol{u}) := \mathrm{sinc}^4\left(\frac{\boldsymbol{u}}{4\tau}\right),\ \tau > 0\,, \tag{1.9d}$$

  where the sinc^4 function is defined on $\mathbb{R}^d$ as the product over coordinates of the fourth power of the univariate sinus cardinal $\sin(u)/u$. This kernel was introduced by De Castro et al. (2021, Section 4.1) to derive recovery guarantees for mixture models. Their analysis does not allow to directly prove the **LPC**.

Importantly, these kernels (except the low-pass spherical harmonic) will serve as "pivot kernels" in our theoretical analysis—a central concept to our contribution. A pivot kernel is one that satisfies the local positive curvature assumption (**LPC**) and enables the construction of non-degenerate dual certificates for a larger class of BLASSO problem with *other* model kernels. As a key innovation, we establish that the error bounds (1.7) for the estimator $\widehat{\mu}$ *remain valid even when the model kernel and pivot kernel differ*, provided certain embedding conditions are satisfied, the latter being formalized in Assumption 2 and the so-called *kernel switch*.

Kernel name	Domain $\mathcal{X}$	$K(\boldsymbol{s}, \boldsymbol{t})$	Assumption 1 (**LPC**)	**LPC**-parameters
Jackson	$(\mathbb{R}\backslash\mathbb{Z})^d$	(1.9a)	Poon et al. (2023, Appendix C)	$r_0 = \mathcal{O}(1)$, $\Delta_0 = \mathcal{O}\big(d^{\frac{1}{2}} s_0^{\frac{1}{4}}\big)$
Gaussian	$\mathbb{R}^d$	(1.9b)	Poon et al. (2023, Appendix D)	$r_0 = \mathcal{O}(1)$, $\Delta_0 = \mathcal{O}\big(\log(s_0)\big)$
Laplace	$\mathbb{R}^d$	(1.9c)	Poon et al. (2023, Appendix E)	$r_0 = \mathcal{O}(1)$, $\Delta_0 = \mathcal{O}\big(d + \log(d^{\frac{3}{2}} s_0)\big)$
Sinc-4	$\mathbb{R}^d$	(1.9d)	This work - Theorem 3	$r_0 = \mathcal{O}\big(\frac{1}{d}\big)$, $\Delta_0 = \mathcal{O}\big(d^{\frac{7}{4}} s_0^{\frac{1}{4}}\big)$

Table 1: Table of known *pivot* kernels verifying the **LPC** assumption, with associated references for proofs. The parameters s_0 and r_0 are the sparsity and radius of the near regions, respectively. The parameter Δ_0 is the minimal separation between two spikes in the target measure.

1.2 Continuous sparse regression and sketching

A key motivation for the introduction of pivot kernels is to address model kernels arising from sketched problems, which we now detail. The principle of sketching techniques is to compress a dataset into a smaller vector, the *sketch*, by computing averages (over the samples of the dataset) of m non-linear random features. This compression is designed to retain statistical guarantees for the underlying learning task (Gribonval et al. 2021a). Recent studies (Gribonval et al. 2021a; Belhadji and Gribonval 2024) have shown that the corresponding sketched forward operator (formally defined below in (1.10a)) acts as a quasi-isometry when restricted to the set of sparse measures whose support points are sufficiently separated. For continuous sparse regression, BLASSO analyses for forward operators defined via sketching have been investigated in (Poon et al. 2019; Poon et al. 2023). This has led to the development of a corresponding *sketched* BLASSO formulation.

The general goal is to estimate a sparse target measure $\mu^0 \in \mathcal{M}_{s,\Delta,\mathfrak{d}}$ from a sketch $\boldsymbol{y}_{\mathsf{sketch}} \in \mathbb{C}^m$, by solving the corresponding BLASSO program (1.1–J_κ), with the sketched forward operator F_{sketch}, which definition involves the sketching law Λ. Thus the corresponding setting is as follows:

- *Hilbert space* $\mathcal{F}_{\mathsf{sketch}}$: consider the standard dot product on $\mathbb{C}^m$, $\mathcal{F}_{\mathsf{sketch}} = \mathbb{C}^m$,
- *Forward measurement* $F_{\mathsf{sketch}} : \mathcal{M}(\mathcal{X}) \to \mathbb{C}^m$ given by

$$(F_{\mathsf{sketch}}\mu)_i = \frac{1}{\sqrt{m}} \int_{\mathcal{X}} \varphi_{\boldsymbol{\omega}_i}(\boldsymbol{t}) \mathrm{d}\mu(\boldsymbol{t}) \,, \quad i = 1, \dots, m \,, \tag{1.10a}$$

  where $\varphi_{\boldsymbol{\omega}} : \mathcal{X} \to \mathbb{C}$ is a complex-valued continuous function, referred to as the *sketching function*.
- *Sketching law* Λ: the sequence of *i.i.d.* random vectors $(\boldsymbol{\omega}_i)_i$ is drawn with respect to Λ,

$$\boldsymbol{\omega}_i \underset{\text{i.i.d.}}{\sim} \Lambda \,, \quad i = 1, \dots, m \,. \tag{1.10b}$$

Thus, in line with Definition 1, we define the *sketched* model kernel by

$$K_{\mathsf{sketch,mod}}(\boldsymbol{s}, \boldsymbol{t}) := \left\langle F_{\mathsf{sketch}}\delta_{\boldsymbol{s}}, F_{\mathsf{sketch}}\delta_{\boldsymbol{t}} \right\rangle_{\mathbb{C}^m} = \frac{1}{m} \sum_{i=1}^{m} \varphi_{\boldsymbol{\omega}_i}(\boldsymbol{s}) \overline{\varphi}_{\boldsymbol{\omega}_i}(\boldsymbol{t}) \,, \tag{1.11a}$$

where $\overline{\varphi}_{\boldsymbol{\omega}}$ denotes complex conjugation. This kernel depends on the draw of the sketch, and taking its expectation naturally leads to the *population* model kernel

$$K_{\mathsf{mod}}(\boldsymbol{s}, \boldsymbol{t}) := \mathbb{E}_{\boldsymbol{\omega} \sim \Lambda} \left[\varphi_{\boldsymbol{\omega}}(\boldsymbol{s}) \overline{\varphi}_{\boldsymbol{\omega}}(\boldsymbol{t}) \right] . \tag{1.11b}$$

Note that every draw of the sketch model kernel meets Definition 1 and satisfies the continuity assumption (1.3–$\mathbf{H}_{\mathsf{mod}}$) with the construction (1.10). Furthermore, its limit as m tends to infinity is almost surely the population model kernel, for fixed $\boldsymbol{s}$ and $\boldsymbol{t}$.

An interesting feature of the sketched problem is to reduce the computational cost induced by the model kernel by considering an empirical version of rank m. Moreover, Poon et al. (2023) showed that the error bound analysis of the BLASSO also holds for its sketched counterpart, provided that K_{mod} satisfies the **LPC** assumption, under additional probabilistic assumptions on the tails of the feature function $\varphi_{\boldsymbol{\omega}}$ and its derivatives, which are detailed later in Assumption 3.

However, proving the **LPC** assumption has to be done on a case by case basis, possibly leading to tedious calculations. Instead, we propose in this paper to *switch* to another kernel (known to satisfy the **LPC**) using an alternative (continuous) sketching function $\psi_{\boldsymbol{\omega}} : \mathcal{X} \to \mathbb{C}$. We define the associated kernel by

$$K(\boldsymbol{s}, \boldsymbol{t}) := \mathbb{E}_{\boldsymbol{\omega} \sim \Lambda} \left[\psi_{\boldsymbol{\omega}}(\boldsymbol{s}) \overline{\psi}_{\boldsymbol{\omega}}(\boldsymbol{t}) \right] , \tag{1.12–$\mathbf{H}_{\psi_{\boldsymbol{\omega}},\Lambda}$}$$

where we assume that $K(\cdot, \cdot)$ is real valued and the expectation is taken with respect to the same distribution Λ. Given a model kernel K_{mod}, we characterize a broader class of kernels K that can be used as *pivot* in the BLASSO analysis, provided that $\mathcal{H}_{\mathsf{mod}}$ is continuously included in their RKHS (see Figure 3). Our result offers an alternative route to obtain the controls of Equations 1.7 when considering a specific BLASSO problem.

1.3 Main contributions

To date, only a limited set of kernels, such as those listed in Table 1, have been rigorously proved to enjoy the local positive curvature property (Assumption 1). This scarcity can restrict the theoretical analysis of BLASSO

estimators and the application of sketching techniques. This paper introduces a significant advancement by demonstrating that an existing kernel satisfying the **LPC**—termed a *pivot kernel*—can be used to derive statistical guarantees for BLASSO problems involving a *different* model kernel K_{mod}, including the case of a sketching kernel $K_{\text{sketch,mod}}$. This is possible provided that the RKHS of the pivot kernel is continuously embedded into the RKHS of the model kernel, a condition formalized in the forthcoming Assumption 2.

This *kernel switch* principle, illustrated in Figure 3, substantially broadens the scope of continuous sparse regression. It allows practitioners to leverage the established theoretical guarantees of well-characterized pivot kernels while employing model kernels that might be more computationally advantageous or better suited to specific problem structures. Our contributions to the theory of continuous sparse regularisation, stemming from this principle and other novel analyses, are fourfold:

- **Effective near region** – In the literature, the size r of the near regions (1.6a) in the error bounds (1.7) is fixed and prescribed by the parameter r_0 of the **LPC** condition satisfied by the model kernel K_{mod}. Instead, we show that the BLASSO can localize up to $r \sim \sqrt{\gamma\sqrt{s_0}}$.
  Consequences for Mixture Modeling: When the model class is fixed and the noise level decreases as $\mathcal{O}(1/\sqrt{n})$, as in the mixtures problem where n is the sample size, one gets effective near regions of size essentially $\mathcal{O}(n^{-\frac{1}{4}})$. We extend this analysis to sketching techniques for kernels of the form (1.12–$\mathbf{H}_{\psi_{\boldsymbol{\omega}},\Lambda}$) with a sketch size m proportional to $\mathcal{O}(s_0 \log^2(s_0))$ up to polynomial factors in the dimension d and logarithmic in the size of the compact $\mathcal{X}$, see Assumption 3.
- **Kernel switch** – We introduce the so-called pivot kernel analysis. It shows that the statistical error bounds (1.7) still holds with *effective* near regions when the pivot kernel satisfies the **LPC** (Assumption 1) and its RKHS continuously embeds into the RKHS of the model (Assumption 2).
  Consequences for Mixture Modeling: For translation invariant kernels, this latter assumption can be expressed as a simple control between the ratio of characteristic functions of the pivot and the model kernels, see (2.6–$\mathbf{H}_{\phi}$). We extend this analysis to sketching techniques for pivot kernels of the form (1.12–$\mathbf{H}_{\psi_{\boldsymbol{\omega}},\Lambda}$), with a sketch size m proportional to $\mathcal{O}(s_0 \log^2(s_0))$ up to log factors in the size of the compact $\mathcal{X}$ and polynomial factors in the dimension d, see Assumption 3.
- **Sketching with the** sinc-4 **pivot kernel** Ψ_τ – For the first time, we show that the sinc-4 satisfies the **LPC** and Assumption 3.
  Consequences for Mixture Modeling: In particular, it answers the question whether one can use sketching techniques for mixture modeling using bandlimiting smoothing, as detailed in Section 2. Our result

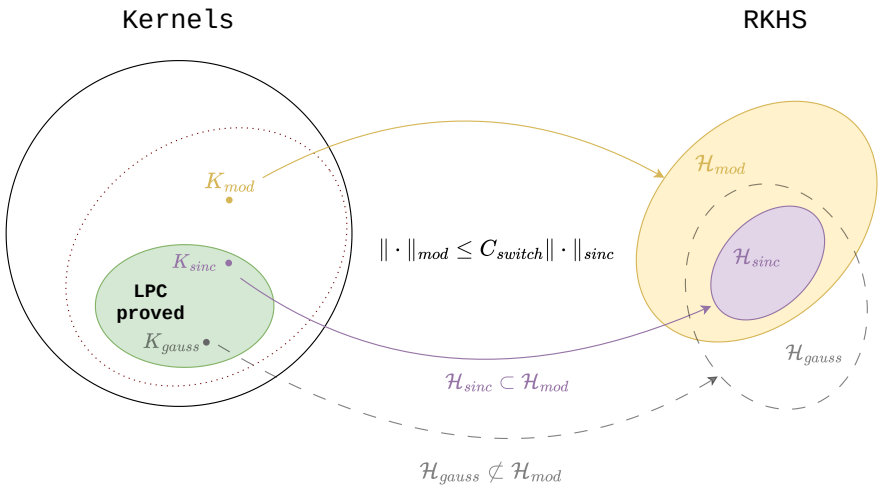

Figure 3: Illustration of the kernel switch principle. Left: A small set of pivot kernels known to satisfy the local positive curvature assumption (**LPC**) enables theoretical guarantees for a much larger class of model kernels. Right: A model kernel K_{mod} is admissible for our statistical guarantees if there exists a pivot LPC kernel (here K_{sinc}) whose RKHS is continuously embedded into the model RKHS. The error bounds remain valid with an additional scaling factor C_{switch}. In this example, using the sinc kernel as pivot is valid because its RKHS is included in the model RKHS, while the Gaussian kernel cannot serve as a pivot since its RKHS contains functions with faster-decaying Fourier transforms than the model allows.

holds for any template distribution ϕ (the elementary distribution involved in the mixture model (2.1)) satisfying (2.6–$\mathbf{H}_\phi$), which asserts that the characteristic function of the template ϕ is positive on low frequencies.

- **Near-optimal solutions as estimators with guarantees** – We uncover an interesting feature of our convex programming problem, the statistical error bounds (1.7) hold for *any* μ such that $J_\kappa(\mu) \leq J_\kappa(\mu^0)$, not only the minimum. In practice, one does not need to solve exactly BLASSO (1.1–J_κ), any sufficiently near solution will do. This is reminiscent of related properties for the LASSO (Foucart and Rauhut 2013).

Contributions for Mixture Modeling Our contribution on sketching sparse regularisation of mixtures is detailed in the rest of the paper with a special focus on the sinc-4 kernel Ψ_τ as pivot. This lead to S2Mix, a more practical algorithm for the sketched Supermix problem (De Castro et al. 2021), see Section 2. In the context of mixture modeling, the consequences of our results are as follows: S2Mix is a computationally efficient algorithm to estimate the parameters of a mixture model, even for large datasets. It achieves a sketch size complexity of $\mathcal{O}(s_0 \log^2(s_0))$, up to polynomial factors in the dimension d and logarithmic in the size of the compact $\mathcal{X}$, which can be significantly lower than the sample complexity n; S2Mix is designed to work with a wide range of mixture models, including those with non-Gaussian template distribution ϕ, since it only requires that the latter satisfies the embedding condition (2.6–$\mathbf{H}_\phi$). This makes it a versatile tool for various applications in machine learning and statistics. The algorithm is robust to noise and model misspecification, as it can handle a wide range of noise levels and model assumptions. This robustness is particularly important in real-world applications where data may be noisy or incomplete. Finally, S2Mix provides theoretical guarantees for parameter recovery (1.7), ensuring that the estimated parameters are close to the true parameters of the mixture model with high probability.

1.4 Related works

Mixture models using BLASSO have been thoroughly investigated by De Castro et al. (2021), who employed a smoothing kernel on the data points (as described in detail in Section 2 of this paper) combined with an ad-hoc construction of non-degenerate dual certificates η^0. Their approach was inspired by the pioneering work of Candès and Fernandez-Granda (2014), whose proof can be reread under the viewpoint developed in this paper as using the Dirichlet kernel as model kernel and the Jackson kernel (Dirichlet kernel raised to the power of 4) as pivot kernel for the super-resolution deconvolution problem on the torus. Building directly on this methodology, De Castro et al. (2021) utilized a proof technique which can also be revisited as relying on a sinc smoothing kernel with a sinc-4 pivot kernel. Notably, both papers implicitly leverage a kernel switch analysis without formally characterizing the general principles of this phenomenon. Furthermore, De Castro et al. (2021) did not provide a proof that the sinc-4 kernel satisfies the local positive curvature assumption (**LPC**), making it impossible to derive sketching statistical error bounds using the framework later developed by Poon et al. (2023).

In their comprehensive treatment, Poon et al. (2023) formalized the **LPC** condition to analyze sketched BLASSO problems across various models. For Gaussian mixture models with fixed covariance, they specifically established that the Gaussian kernel with matching covariance satisfies the **LPC** (Poon et al. 2023, Section 2.3 and Appendix D.5.1). However, their analysis remains fundamentally tied to Gaussian mixtures with Gaussian smoothing kernels. In contrast, our kernel switch analysis significantly broadens the applicability of continuous sparse regularization by accommodating, under mild assumptions detailed in Section 2, a much wider range of mixture densities and smoothing kernels.

This article establishes that near minimizers of the BLASSO problem—defined as solutions with an objective value no greater than that of the target $J_\kappa(\mu^0)$—share the same statistical guarantees as the exact minimizer itself. Although this property is implicit in the proofs of previous works on Bregman divergence (Candès and Fernandez-Granda 2014; Duval and Peyré 2015; Duval and Peyré 2017; Azais et al. 2015; Poon et al. 2023), it has not been explicitly stated. A similar concept exists in the Compressed Sensing (CS) literature, where guarantees for exact minimizers of LASSO-related problems are often proven via more general results that also apply to near minimizers (see e.g., Foucart and Rauhut 2013, Theorems 4.18 & 4.19, 4.21 & 4.23).

The near-minimizer property is of great practical importance, as it justifies the use of approximate solvers. This is especially relevant for large-scale or high-dimensional problems where finding the exact minimizer is computationally infeasible. Recent algorithms to solve the BLASSO problem, such as Sliding Frank-Wolfe (Denoyelle et al. 2019) and stochastic versions of Conic Particle Gradient Descent (CPGD) (Chizat 2019; De Castro et al. 2023), are designed precisely to find such near minimizers with convergence guarantees. However, while these CPGD-based methods can find a near minimizer in a finite number of steps, the required number of steps scales exponentially with the dimension d, currently limiting their application to moderate-dimensional regimes where d is less than approximately 100.

Mixture models represent a fundamental framework in modern statistics, with diverse applications spanning machine learning, biology, genetics, and astronomy (McLachlan and Peel 2004; Fruhwirth-Schnatter et al. 2019). Traditional inference for these models typically relies on the non-convex maximum-likelihood paradigm solved via the Expectation-Maximisation (EM) algorithm (McLachlan and Krishnan 2007). A critical challenge in mixture modeling—determining the optimal number of components s_0—is conventionally addressed post-inference as a model selection problem (Celeux et al. 2019). To handle large-scale datasets, the computational statistics community has developed sophisticated methodological tools including variational inference (Blei et al. 2017) and amortized inference techniques (Kingma 2014; Tomczak 2024), though theoretical guarantees for parameter recovery remain relatively limited. The sparse continuous regression perspective on mixture modeling, recently introduced by De Castro et al. (2021) and Poon et al. (2023), offers compelling theoretical guarantees for parameter recovery while demonstrating robustness to noise and model misspecification. This approach complements practical algorithms developed to solve the BLASSO optimisation problem, such as sliding Frank-Wolfe (Denoyelle et al. 2019) and particle-based methods (Chizat 2019; De Castro et al. 2023), providing both theoretical foundations and computational efficiency.

Outline

The remainder of the paper is organized as follows:

- Section 2 introduces the sketching BLASSO problem for mixture models. We first present a resolution-based approach to mixture modeling and show how it can be formulated as a continuous sparse regularisation problem. We then develop a sketched version of the framework, called S2Mix, which achieves significant computational efficiency through random Fourier features. We derive statistical guarantees for this approach, demonstrating how the sinc-4 kernel serves as an effective pivot kernel that enables parameter estimation with near-optimal sample complexity.
- Section 3 presents our three main theoretical contributions. First, our effective near regions analysis in Theorem 1 demonstrates that the size of near regions can adapt to the noise level, yielding localisation guarantees of order $\mathcal{O}(\sqrt{\gamma\sqrt{s_0}})$ that improve with increasing sample size. Second, our kernel switch technique in Theorems 1 and 2 establishes that statistical error bounds hold when using a different kernel for analysis than for measurement, provided the RKHS embedding condition (Assumption 2) is satisfied. Third, we prove in both theorems that near-optimal solutions—*i.e.,* measures with objective value not exceeding that of the target measure—enjoy the same statistical guarantees as the exact minimizer, which is particularly valuable for practical implementations. This section also includes a rigorous analysis proving that the sinc-4 kernel satisfies the local positive curvature assumption required for sketching.

Technical proofs, additional mathematical background, and supporting material are provided in the appendix. A comprehensive list of notation precedes the references.

2 Sketching mixture models via continuous sparse regression

This section focuses on the application of our general theoretical framework to the problem of mixture model estimation. We present specific consequences of the broader results that will be detailed in Section 3. The key novelties for mixture modeling include the ability of our proposed algorithm, S2Mix, to handle a wide variety of template densities beyond the commonly assumed Gaussian. Furthermore, we establish new localisation error bounds for parameter recovery, demonstrating improved scaling as the sample size increases. These advancements significantly enhance the practical utility and theoretical understanding of continuous sparse regularisation for mixture models.

2.1 A resolution-based approach of mixture modeling

We demonstrate our theoretical contributions through their application to the important problem of mixture model estimation. Mixture models represent fundamental statistical tools used across machine learning, signal processing, and computational biology to model heterogeneous populations. Before introducing the technical details, let us sketch the key ideas: we reformulate mixture estimation as a continuous sparse regression problem where the unknown mixture components—both their weights and locations—are encoded as a sparse discrete measure. This measure can then be recovered using BLASSO, even when working with compressed data through sketching techniques.

In this section, we first present a resolution-based approach to mixture modeling via continuous sparse regression. We then develop an efficient sketching algorithm, coined S2Mix, that enables parameter estimation from large datasets with near-optimal sample complexity. Crucially, we establish novel statistical guarantees for this approach by leveraging the sinc-4 kernel Ψ_τ as a pivot kernel within our theoretical framework. This demonstrates how our effective near regions analysis and kernel switch technique deliver practical improvements for an important class of statistical models.

We are interested in estimating the mixture positive weights $(a_k^0)_k$ and parameters $(\boldsymbol{x}_k^0)_k$ from a n-sample $\boldsymbol{Z} := \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n\}$ of a translation invariant mixture model with s_0-components:

$$f^0 := \sum_{k=1}^{s_0} a_k^0 \phi(\cdot - \boldsymbol{x}_k^0) \quad \text{with} \quad \sum_k a_k^0 = 1 \text{ and } \boldsymbol{x}_k^0 \in \mathcal{X}\,, \tag{2.1}$$

where $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$. Here, ϕ is a probability density function on $\mathbb{R}^d$ called the *template distribution*. Popular instances include mixtures of Gaussians with fixed and known covariance, or mixtures of Laplace distributions. More generally, our proposed framework will also encompass mixtures of any distribution with known characteristic function, such as mixtures of stable distributions for instance.

From a non-parametric statistics perspective, estimating a mixture model amounts to estimate the density f^0, which can be done via kernel density estimation (*i.e.,* smoothing the empirical sample distribution), namely

$$L_\tau \hat{f}_n := \lambda_\tau \star \hat{f}_n \quad \text{where} \quad \hat{f}_n := \frac{1}{n} \sum_{j=1}^{n} \delta_{\boldsymbol{z}_j}\,.$$

The *smoothing* function $\lambda_\tau(\cdot) = \tau^{-d}\lambda(\cdot/\tau)$ is defined by the smoothing kernel λ and its *bandwidth* parameter τ. In this work, we focus on the sinus cardinal smoothing kernel which is a standard choice in the literature (De Castro et al. 2021), given by

$$\lambda(\boldsymbol{u}) := \operatorname{sinc}(\boldsymbol{u}) = \prod_{i=1}^{d} \frac{\sin(u_i)}{u_i} \quad \text{so that} \quad \lambda_\tau(\boldsymbol{x}) = \tau^{-d} \operatorname{sinc}\left(\frac{\boldsymbol{x}}{\tau}\right). \tag{2.2}$$

This function acts as a low-pass filter in the Fourier domain, other filters are possible but we will consider the sinus cardinal function in this section. The crux of non-parametric methods is then to optimize the mean quadratic loss of the estimator with respect to the bandwidth parameter τ_n, usually depending on n, the latter translating into a bias-variance trade-off.

In this work, we follow a different path considering that the bandwidth parameter τ will be fixed by the practitioner and will not depend on the sample size n but rather on the *minimal separation* between parameters $(\boldsymbol{x}_k^0)_k$, also referred to as *Rayleigh limit*. Precisely, we assume that the bandwidth $\tau > 0$ is such that

$$\tau \le \tau_{\max}(\mu^0) := \frac{\min_{1 \le k \ne \ell \le s_0} \left\|\boldsymbol{x}_k^0 - \boldsymbol{x}_\ell^0\right\|_2}{147.77\, {s_0}^{1/4} d^{7/4}}\,. \tag{2.3–$\mathbf{H}_\tau$}$$

As we will see, this hypothesis ensures that the bandwidth is small enough to resolve the mixture components.

Now, consider a putative parameter θ given by

$$\theta := \left\{s;\, a_1, \dots, a_s;\, x_1, \dots x_s\right\} \quad \text{and} \quad \mu_\theta := \sum_{k=1}^{s} a_k \delta_{\boldsymbol{x}_k}\,,$$

with $s \ge 1$ the number of mixing components and μ_θ is the lifting of the putative parameter on the space of measures, where $\delta_{\boldsymbol{t}}$ denotes the Dirac mass at location $t \in \mathcal{X}$.

The mixture model density defined in Equation (2.1) can be seen as an image of a discrete measure through the convolution operator:

$$\Phi: \quad \mu \in \mathcal{M}(\mathcal{X}) \quad \mapsto \quad \Phi\mu := \phi \star \mu = \int \phi(\cdot - \boldsymbol{x})\, \mathrm{d}\mu(\boldsymbol{x}) = \int \phi_{\boldsymbol{x}}\, \mathrm{d}\mu\,, \tag{2.4}$$

where $\phi_{\boldsymbol{x}} := \phi(\cdot - \boldsymbol{x})$. This mapping is reminiscent of the so-called Kernel Mean Embedding of μ in Machine Learning, and they exactly corresponds when the template ϕ is of positive type (Muandet et al. 2017).

The mixture problem seeks to fit a parameter θ from the observation of $\hat{f}_n$, namely

$$\theta \xrightarrow{\text{Non-linear}} \left[f_\theta := \Phi\mu_\theta = \sum_{k=1}^{s} a_k \phi(\cdot - t_k) \right] \xrightarrow{\text{Sampling}} \hat{f}_n \,.$$

where the step $\theta \to f_\theta$ is a non-linear mapping of the parameters to the mixture density. Directly solving the corresponding inverse problem is hard, both because of the non-linearity as well as the fact that those two objects do not belong to the same functional space. On the other hand, lifting the problem on the space of measures yields the following linear problem

$$\mu_\theta \xrightarrow{\text{Linear}} \left[f_\theta = \Phi\mu_\theta \right] \xrightarrow{\text{Embedding}} \left[L_\tau f_\theta = (L_\tau \circ \Phi)\mu_\theta \right] \approx L_\tau \hat{f}_n \xleftarrow{\text{Embedding}} \hat{f}_n, \tag{2.5}$$

where we have embedded the candidate $L_\tau f_\theta$ and the observation $L_\tau \hat{f}_n$ into the same Hilbert space $\mathcal{F}_\tau$, the RKHS of λ_τ.

2.2 Continuous sparse regularisation for mixture estimation – Supermix

Recent advances have established a powerful connection between continuous sparse regularization techniques developed for super-resolution problems and the statistical challenge of mixture estimation (De Castro et al. 2021; Poon et al. 2023). Indeed, as we have just seen, mixture modeling can be reformulated as the problem of recovering an unknown sparse discrete measure μ^0 from observations obtained through a linear measurement operator $F = L_\tau \circ \Phi$ mapping from the space of finite Radon measures $\mathcal{M}(\mathcal{X})$ to a Hilbert space $\mathcal{F}$, a possible approach to address it is to seek a (near) optimum of the BLASSO problem (1.1–J_κ).

This formulation offers the significant advantage that the number of components s_0 need not be specified in advance. Speciminimalfically, this approach involves

- *Hilbert space*: $\mathcal{F} := \mathcal{F}_\tau$, the RKHS associated to the smoothing kernel $\lambda_\tau(t-s)$ defined in (2.2). As a translation invariant reproducing kernel, it admits an explicit characterisation in the Fourier domain (see Appendix C.3).
- *Forward measurement operator* $F := L_\tau \circ \Phi : \mathcal{M}(\mathcal{X}) \to \mathcal{F}_\tau$.
- *Observation* $\boldsymbol{y} := L_\tau \hat{f}_n$ which is of the form

  $$\boldsymbol{y} = F\mu^0 + \Gamma_n \quad \text{with} \quad \Gamma_n := L_\tau \hat{f}_n - \mathbb{E}[L_\tau \hat{f}_n],$$

  the latter being a centered noise term due to n-sampling with $\mathbb{E}[L_\tau \hat{f}_n] = L_\tau f^0 = F\mu^0$.

Hence, one can use a sparse regression estimation procedure to fit the mixture model, which was coined *Supermix* by De Castro et al. (2021). For the specific problem at hand, the model kernel defined in (1.2) is translation invariant and can be described in the Fourier domain (see Appendix C.3). Indeed, one has (see Remark 3.1)

$$\begin{aligned} K_{\text{mod}}(\boldsymbol{s},\boldsymbol{t}) &:= \langle F\delta_{\boldsymbol{s}}, F\delta_{\boldsymbol{t}} \rangle_{\mathcal{F}} = \langle \lambda_\tau \star \phi_{\boldsymbol{s}}, \lambda_\tau \star \phi_{\boldsymbol{t}} \rangle_{\mathcal{F}_\tau}, \\ &= \frac{1}{(2\pi)^d} \int \mathfrak{F}[\lambda_\tau](\boldsymbol{\omega}) \cdot |\mathfrak{F}[\phi]|^2(\boldsymbol{\omega}) \cdot e^{+\imath \boldsymbol{\omega}^\top (\boldsymbol{t}-\boldsymbol{s})} \,\mathrm{d}\boldsymbol{\omega}, \\ &= \mathfrak{F}^{-1}[\mathfrak{F}[\lambda_\tau] \cdot |\,\mathfrak{F}[\phi]|^2](\boldsymbol{t}-\boldsymbol{s}), \\ &= (\lambda_\tau \star \phi \star \check{\phi})(\boldsymbol{t}-\boldsymbol{s})\,, \end{aligned}$$

where $\check{\phi}(\boldsymbol{x}) = \phi(-\boldsymbol{x})$. At first glance, the natural road to derive estimation error bounds is trying to prove **LPC** (Assumption 1) for this translation invariant kernel. However, this would require controlling non-trivial quantities involving the model kernel and its associated Fisher-Rao metric. Alternatively, De Castro et al. (2021) proposed to use the sinc-4 kernel $K_{\text{pivot}}(\boldsymbol{s},\boldsymbol{t}) = \Psi_\tau(\boldsymbol{t}-\boldsymbol{s})$ defined in Equation (1.9d) as a proxy to build non-degenerate certificates, implicitly relying on a *kernel switch* analysis. The latter holds under the assumption that the RKHS $\mathcal{H}_{\text{pivot}}$, defined by K_{pivot}, is continuously included in $\mathcal{H}_{\text{mod}}$, defined by K_{mod}. In the special case where both K_{pivot} and K_{mod} are translation invariant, this translates into an assumption on the boundedness of the ratio of their Fourier transforms (see *e.g.* Zhang and Zhao (2013, Corollary 3.2) and Appendix C.6):

$$C_{\text{switch}} = C_{\text{switch}}(\tau,\phi) := \sup_{\boldsymbol{\omega} \in [-1/\tau, 1/\tau]^d} \sqrt{\frac{\mathfrak{F}[\Psi_\tau]}{\mathfrak{F}[\lambda_\tau]\,|\,\mathfrak{F}[\phi]|^2}(\boldsymbol{\omega})} < +\infty, \tag{2.6–$\mathbf{H}_\phi$}$$

with the convention $\frac{0}{0} = 0$. This is a special instance of the general results discussed in Section 3. The scaling of this constant C_{switch} with respect to the bandwidth parameter τ depends on the smoothness of $\mathfrak{F}[\phi]$ as discussed in De Castro et al. (2021, Section 5.2). In particular, the case of *supersmooth* densities defined hereafter leads to a scaling in

$$\exists p \in [1, +\infty],\ \alpha, \beta > 0, \quad \mathfrak{F}[\phi](\boldsymbol{\omega}) \propto e^{-\alpha \|\boldsymbol{\omega}\|_p^\beta} \implies C_{\text{switch}} = \mathcal{O}_d\Big(\tau^{d/2} e^{\alpha \left(\frac{d^{1/p}}{\tau}\right)^\beta}\Big),$$

where the constant in $\mathcal{O}_d$ may depend exponentially in the dimension d, but it does not depend on τ. This case encompasses *e.g.* Gaussian, multivariate Cauchy, product of univariate Cauchy, or more generally of centered stable distributions with known scale parameter and zero skewness. As an example, when ϕ is a Gaussian density, the scaling is given by $C_{\text{switch}} = \mathcal{O}_d(\tau^{d/2} e^{d/2\tau^2})$.

In addition, the scaling of the noise level $\gamma_n = \|\Gamma_n\|_{\mathcal{F}_\tau}$ defined in Equations 1.5 can be controlled with high probability over the draw of the sample $\boldsymbol{Z}$ given by the following Lemma

Lemma 2.1 (Control of the noise level γ_n)**.** *Let $\alpha > 0$. Then with probability at least $1 - \alpha$, it holds that*

$$\gamma_n \leq C_\alpha \frac{\tau^{-\frac{d}{2}}}{\sqrt{n}}, \tag{2.7}$$

where $C_\alpha := 2\sqrt{1 + C_1 \log(C_2/\alpha)}$ is a constant only depending on α, while C_1 and C_2 are universal constants.

Proof. The result is proved in De Castro et al. (2021, Lemma 16) using concentration of U-process and Arcones and Giné (1993, Proposition 2.3). A closely related proof is given for the sketched case in Lemma 2.2, see also Remark 2.5. □

Thus, we can tune the regularisation κ of the BLASSO problem (1.1–J_κ), which is crucial for the statistical recovery result. This leads to the following proposition, which is an instance of the upcoming Theorem 1 for the Supermix problem, using the sinc-4 kernel Ψ_τ as pivot. It can also be seen as a refined version of De Castro et al. (2021, Theorem 10) as commented below.

Proposition 2.1 (Statistical guarantees for the Supermix problem)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact set, and the target $\mu^0 \in \mathcal{M}(\mathcal{X})$ be a sparse measure of s_0 spikes $\{\boldsymbol{x}_1^0, \dots, \boldsymbol{x}_{s_0}^0\}$, with minimal separation $\Delta = \min_{k \neq l} \|\boldsymbol{x}_k^0 - \boldsymbol{x}_l^0\|^2$. Fix the bandwidth parameter $\tau \leq \tau_{\max}(\mu^0)$ as in Equation (2.3–$\mathbf{H}_\tau$). Take any template distribution ϕ such that it achieves a finite* kernel switch *constant $C_{\text{switch}}(\tau, \phi) < +\infty$ in Equation (2.6–$\mathbf{H}_\phi$). Consider the forward measurement operator $F : \mu \in \mathcal{M}(\mathcal{X}) \to \lambda_\tau \star \phi \star \mu \in \mathcal{F}_\tau$, where λ_τ is the smoothing kernel defined in Equation (2.2). Let $\boldsymbol{Z} = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n\}$ be an i.i.d. sample from $\phi \star \mu^0$, and $\hat{f}_n$ its associated empirical distribution. Define the following BLASSO problem, named Supermix, with observation $\boldsymbol{y} = \lambda_\tau \star \hat{f}_n$*

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} \left\{ J_\kappa(\mu) = \frac{1}{2} \|\boldsymbol{y} - F\mu\|_{\mathcal{F}_\tau}^2 + \kappa \|\mu\|_{\text{TV}} \right\}. \tag{2.8–Supermix}$$

Take any diverging sequence δ_n and set the near and far regions' effective radius *$r_n = \delta_n n^{-1/4}$. Fix the regularization $\kappa = \frac{1}{\sqrt{2} C_{\text{switch}}} \frac{C_\alpha \tau^{-\frac{d}{2}}}{\sqrt{n}} \frac{1}{\sqrt{s_0}}$. Then, for any near optimal $\mu \in \mathcal{M}(\mathcal{X})$ such that $J_\kappa(\mu) \leq J_\kappa(\mu^0)$, $n \geq c_d \delta_n^4$ where c_d depends polynomially on d, and $\alpha > 0$, we have with probability $1 - \alpha$ over the drawing of $\boldsymbol{Z}$:*

- *Control of the far region:*

$$|\mu|\big(\mathcal{F}^{\text{reg}}(r)\big) \leq \frac{256\sqrt{2}\, C_\alpha C_{\text{switch}} \tau^{-\frac{d}{2}}}{23\, \delta_n^2} \sqrt{s_0}, \tag{2.9a}$$

- *Control of all the near regions: for all $k \in [s_0]$,*

$$\Big|\mu\big(\mathcal{N}_k^{\text{reg}}(r)\big) - a_k^0\Big| \leq \frac{1536\sqrt{2}\, C_\alpha C_{\text{switch}} \tau^{-\frac{d}{2}}}{23\, \delta_n^2} \sqrt{s_0} \tag{2.9b}$$

- *Detection level: for all Borelian $A \subset \mathcal{X}$ such that $|\mu|(A) > \frac{256\sqrt{2} C_\alpha C_{\text{switch}} \tau^{-\frac{d}{2}}}{23 \delta_n^2} \sqrt{s_0}$, there exists $\boldsymbol{x}_k^0$ such that*

$$\min_{\boldsymbol{t} \in A} \frac{1}{2\sqrt{3}\tau} \|\boldsymbol{t} - \boldsymbol{x}_k^0\|_2 \leq r_n, \tag{2.9c}$$

The proof is given in Appendix A.7, and is an application of the main theorem of Section 3, Theorem 1. Note that the result is stated using the optimal proportionality constant for the regularisation $\kappa \propto \gamma/\sqrt{s_0}$, as commented in Remark 3.2. Moreover, this result is a digest and improvement of De Castro et al. (2021, Theorem 10) as made clear by the following remarks.

Remark 2.1 (Notational comparison with SuperMix)**.** *When comparing with De Castro et al. (2021), the following notational equivalences are relevant:*

- *The expression $1/(4\tau)$, related to inverse bandwidth (or spatial resolution) in this paper, corresponds to their notation m.*
- *Our noise level, $\gamma_n = \|\Gamma_n\|$, corresponds to their notation ρ_n.*
- *Our constant $C_{\mathsf{switch}}(\tau, \phi)$ from Equation (2.6–$\mathbf{H}_\phi$) corresponds to their notation $C_m(\varphi, \lambda)$.*
- *Their radius is denoted as ϵ and does not depend on n (see Remark 2.3 below).*

Remark 2.2 (Model set with sinc-4 pivot)**.** *The sinc-4 pivot kernel Ψ_τ induces the so-called Fisher-Rao distance, properly defined in Equation (3.1), which is here just a rescaled version of the Euclidean distance by a factor $2\sqrt{3}\tau$. Consequently, for a measure to belong to the model class $\mathcal{M}_{s_0,\Delta_0,\mathfrak{d}_\mathfrak{g}}$, its support points $\{\boldsymbol{x}_k^0\}$ must satisfy $\|\boldsymbol{x}_k^0 - \boldsymbol{x}_l^0\|_2 \geq 2\sqrt{3}\tau\Delta_0$. The parameter Δ_0 is the minimal separation inherent to the Local Positive Curvature (**LPC**) property of Ψ_τ, which is established for this kernel in Theorem 3. Assumption (2.3–$\mathbf{H}_\tau$) ensures that the chosen bandwidth τ aligns the actual separations in the target measure μ^0 with this **LPC**-derived requirement for Ψ_τ. Since the bandwidth τ is chosen by the practitioner, this assumption means that they must chose τ small enough to resolve the mixture components, which is a standard assumption in the literature (De Castro et al. 2021).*

Remark 2.3 (Effective near regions)**.** *One of the novelties of Proposition 2.1 is that the radius of the near regions $r_n = \delta_n n^{-1/4}$ now depends on the number of samples. Moreover, any sequence δ_n diverging to infinity is acceptable provided that $r_n \to 0$. The latter controls a trade-off between how fast effective near regions narrow around the true support of μ^0 at a rate $r_n = \delta_n n^{-1/4}$, and how slow the estimator $\widehat{\mu}$ detects the effective near and far regions at a rate $v_n = \delta_n^{-2}$.*

Thus, one have different scalings of the near regions given by the choice of the diverging sequence δ_n

- *Polynomial: $\delta_n = n^\alpha$ where $0 < \alpha < 1/4$. The effective near regions scales in $n^{-1/4+\alpha}$ and $v_n = n^{-2\alpha}$.*
- *Logarithmic: $\delta_n = \sqrt{\log n}$. In this case the effective near regions narrows at the rate $(\log(n)/n)^{1/4}$ while $v_n = \log(n)$.*
- *Log-Polynomial: $\delta_n = n^\alpha \log^\beta n$ where $0 < \alpha < 1/4$ and $\beta \geq 0$. The effective near regions scales in $n^{-1/4+\alpha} \log^\beta n$ and $v_n = n^{-2\alpha} \log^{-2\beta} n$.*

2.3 Statistical error bounds for *sketched* mixture models

While the preceding results are interesting from a theoretical perspective, a major practical difficulty arise when trying to solve the Supermix as the objective function J_κ, or its gradient, involves d-dimensional integrals. In this work, we focus on a *sketched* version of the Supermix BLASSO problem leading to computational gains.

Random Fourier features The main idea revolves on an approximation of the translation invariant model kernel K_{mod} using weighted random Fourier features (Rahimi and Recht 2007; Gribonval et al. 2021b). The latter stems from the representation of a translation invariant K_{mod} as an expectation over a free *sketch* distribution, which is then approximated by m Monte-Carlo samples.

Akin to importance sampling in the Fourier domain, this *sketching distribution* Λ may be chosen different from (a properly normalized version of) the spectral measure $U_\tau := \mathfrak{F}[\lambda_\tau]/(2\pi)^d$ of the smoothing kernel. Formally, let Λ be *any p.d.f. with support containing the hypercube* $\|\boldsymbol{\omega}\|_\infty \leq 1/\tau$, which is the support of U_τ. Then, the

model kernel is amenable to sketching as it can be written in the form of Equation (1.11b) as follows:

$$\begin{aligned} K_{\text{mod}}(\boldsymbol{t},\boldsymbol{s}) &= \int \overbrace{\frac{\mathfrak{F}[\lambda_\tau]}{(2\pi)^d}}^{U_\tau}(\boldsymbol{\omega})|\mathfrak{F}[\phi]|^2(\boldsymbol{\omega})e^{+\imath\boldsymbol{\omega}^\top(\boldsymbol{t}-\boldsymbol{s})}\,\mathrm{d}\boldsymbol{\omega}, \\ &= \int \Lambda(\boldsymbol{\omega})\frac{U_\tau}{\Lambda}(\boldsymbol{\omega})|\mathfrak{F}[\phi]|^2(\boldsymbol{\omega})e^{+\imath\boldsymbol{\omega}^\top(\boldsymbol{t}-\boldsymbol{s})}\,\mathrm{d}\boldsymbol{\omega}, \\ &= \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[\varphi_{\boldsymbol{\omega}}(\boldsymbol{s})\overline{\varphi_{\boldsymbol{\omega}}(\boldsymbol{t})}\right], \end{aligned} \tag{2.10a}$$

$$\text{where } \varphi_{\boldsymbol{\omega}}(\boldsymbol{t}) := W(\boldsymbol{\omega})\ \mathfrak{F}[\phi](\boldsymbol{\omega})\, e^{-\imath\boldsymbol{\omega}^\top\boldsymbol{t}}, \quad \text{with: } W(\boldsymbol{\omega}) = \sqrt{\frac{U_\tau}{\Lambda}(\boldsymbol{\omega})}. \tag{2.10b}$$

Then, this leads to a Monte-Carlo approximation of the kernel, defining the *sketched* model kernel as in (1.11a)

$$K_{\text{sketch,mod}}(\boldsymbol{s},\boldsymbol{t}) := \frac{1}{m}\sum_{i=1}^{m}\varphi_{\boldsymbol{\omega}_i}(\boldsymbol{s})\overline{\varphi}_{\boldsymbol{\omega}_i}(\boldsymbol{t})\,, \tag{2.10c}$$

where the sequence of *i.i.d.* random vectors $(\boldsymbol{\omega}_i)_i$ is drawn with respect to Λ.

Sketching the supermix problem Recalling the definition of Equations 1.10, we are in presence of a sketched version of the Supermix BLASSO problem, coined S2Mix hereafter, with the following characteristics

- *Finite dimensional* Hilbert space: $\mathcal{F}_{\text{sketch}} = \mathbb{C}^m$.
- *Sketching law*: the sequence of *i.i.d.* random vectors $(\boldsymbol{\omega}_i)_i$ is drawn with respect to Λ,

$$\boldsymbol{\omega}_i \underset{\text{i.i.d.}}{\sim} \Lambda\,, \quad i = 1,\dots,m\,. \tag{2.11a}$$

- *Sketched*[1] *forward operator* $F_{\text{sketch}} : \mathcal{M}(\mathcal{X}) \to \mathbb{C}^m$ defined as

$$F_{\text{sketch}}\mu = \frac{1}{\sqrt{m}}\left(\int \varphi_{\boldsymbol{\omega}_i}(\boldsymbol{t})\,\mathrm{d}\mu(\boldsymbol{t})\right)_{i=1}^{m}. \tag{2.11b}$$

  This forward sketched operator yields the sketched model kernel given by Equation (2.10c), which can in turn be viewed as a Monte-Carlo approximation of the model kernel given by Equation (2.10a).

- *Sketch vector*: $\boldsymbol{y}_{\text{sketch}}$, a summary of the dataset easily computed as a weighted version the empirical distribution's Fourier transform taken at the random frequencies $\boldsymbol{\omega}_i$

$$\boldsymbol{y}_{\text{sketch}} := \frac{1}{\sqrt{m}}\left(W(\boldsymbol{\omega}_i)\frac{1}{n}\sum_{j=1}^{n}e^{-\imath\boldsymbol{\omega}_i^\top\boldsymbol{z}_j}\right)_{i=1}^{m}. \tag{2.11c}$$

  The latter is of the form:

$$\boldsymbol{y}_{\text{sketch}} = F_{\text{sketch}}\mu^0 + \Gamma_{\text{sketch}} \quad \text{with: } \Gamma_{\text{sketch}} = \boldsymbol{y}_{\text{sketch}} - F_{\text{sketch}}\mu^0.$$

  The noise term is centered as $\mathbb{E}_{\boldsymbol{Z}}\left[\boldsymbol{y}_{\text{sketch}}\right] = \frac{1}{\sqrt{m}}\left(W(\boldsymbol{\omega}_i)\mathbb{E}_{\boldsymbol{z}\sim\phi\star\mu^0}[e^{-\imath\boldsymbol{\omega}_i^\top\boldsymbol{z}}]\right)_{i=1}^{m} = F_{\text{sketch}}\mu^0$. Hence the data sketch is an empirical version of F_{sketch} at the target point μ^0.

Remark 2.4 (Sketching and compression)**.** *Crucially, computing $\boldsymbol{y}_{\text{sketch}}$ acts as a compression of the original $n \times d$ dataset $\boldsymbol{Z}$ into an m-dimensional sketch in $\mathbb{C}^m$, and is done* only once *prior to learning. This the essence of sketching and compressive learning methods (Gribonval et al. 2021a), seeking statistical guarantees with respect to the sketch size m.*

As in the population case, we can use concentration arguments to control the sketched noise level γ_{sketch}.

[1] Related to the *sketching operator* $\mathcal{A}$ (Gribonval et al. 2021a) which verifies $\boldsymbol{y}_{\text{sketch}} := \mathcal{A}\hat{f}_n \approx \mathcal{A}(\phi \star \mu^0) = F_{\text{sketch}}\mu^0$.

Lemma 2.2. *Let $\alpha \in (0,1)$, $\boldsymbol{Z} = \{\boldsymbol{z}_j\}_{j=1}^n$ be an i.i.d. n-sample from f^0 and $\{\boldsymbol{\omega}_i\}_{i=1}^m$ be an m-sample from Λ. Then, with probability at least $1-\alpha$ over the joint draw of the sample and the sketch, it holds that*

$$\|\Gamma_{\mathsf{sketch}}\|_{\mathbb{C}^m} \leq C_{\alpha,m}\frac{1}{\sqrt{n}},$$

with

$$C_{\alpha,m} = 2\sqrt{\left[\left(\frac{1}{\tau}\right)^d + \frac{1}{2\sqrt{m}}\left\|\frac{U_\tau}{\Lambda}\right\|_\infty \log\left(\frac{2}{\alpha}\right)\right]\left[1 + C_1\log\left(\frac{2C_2}{\alpha}\right)\right]},$$

and $C_1, C_2 > 0$ the same universal constants as in Lemma 2.1.

Remark 2.5. *Note that, as m goes to infinity, we uncover that $C_{\alpha,m}$ tends to $C_{\alpha/2}\left(\frac{4}{\tau}\right)^{\frac{d}{2}}$ and we recover Lemma 2.1, which bounds the noise level in the population case.*

The proof is given in Appendix A.8.1 and, again, we may tune the optimal regularisation κ. Applying the upcoming Theorem 2 to this specific sketched BLASSO, using the sinc-4 pivot kernel Ψ_τ, one can prove the following results for the S2Mix problem.

Proposition 2.2 (Statistical guarantees for S2Mix)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact set, and the target $\mu^0 \in \mathcal{M}(\mathcal{X})$ be a sparse measure of s_0 spikes $\{\boldsymbol{x}_1^0, \dots, \boldsymbol{x}_{s_0}^0\}$, with minimal separation $\Delta = \min_{k\neq l}\|\boldsymbol{x}_k^0 - \boldsymbol{x}_l^0\|^2$. Fix the bandwidth parameter $\tau \leq \tau_{\max}(\mu^0)$ as in Equation (2.3–$\mathbf{H}_\tau$). Take any template distribution ϕ such that it achieves a finite* kernel switch *constant $C_{\mathsf{switch}}(\tau,\phi) < +\infty$ in Equation (2.6–$\mathbf{H}_\phi$). Take any sketching distribution Λ with support containing $[-1/\tau, 1/\tau]^d$ and such that $\|U_\tau/\Lambda\|_\infty < +\infty$. Let $m \in \mathbb{N}^\star$ be the sketch size and consider the following* sketched *BLASSO problem, denoted as sketched Supermix (S2Mix), with sketched forward measurement operator F_{sketch} and observation (data sketch) $\boldsymbol{y}_{\mathsf{sketch}}$ defined in Equations (2.11b) and (2.11c) respectively:*

$$\min_\mu \left\{ J_{\mathsf{sketch},\kappa}(\mu) = \frac{1}{2}\|\boldsymbol{y}_{\mathsf{sketch}} - F_{\mathsf{sketch}}\mu\|_{\mathbb{C}^m}^2 + \kappa\|\mu\|_{\mathrm{TV}} \right\}. \tag{2.12–S2Mix}$$

Suppose the constant $C_\Lambda := \left\|f_\tau^{(4)}/\Lambda\right\|_\infty < +\infty$, where $f_\tau^{(4)} = \mathfrak{F}[\Psi_\tau]/(2\pi)^d$ is the normalized spectral measure of the sinc-4 kernel Ψ_τ (Appendix C.5). Let $C_d, C'_{\mathsf{pivot}} > 0$ be constants that may depend polynomially on the dimension d. Fix the regularisation

$$\kappa = \frac{C_{\alpha,m}}{C'_{\mathsf{pivot}}C_{\mathsf{switch}}\sqrt{n}}\frac{1}{\sqrt{s_0}}.$$

Take any diverging sequence $\delta_n = o(n^{1/4})$ and set the near and far regions' effective radius *$r_n = \delta_n n^{-1/4}$. Let $\alpha > 0$ and the sketch size m be such that*

$$m \geq C_d\, C_\Lambda\, \max(d, \log(s_0))\, s_0\, \log\left(\frac{\max(1, |\mathcal{X}|)\, s_0}{\alpha}\right), \tag{2.13}$$

Then, for any near optimal *$\mu \in \mathcal{M}(\mathcal{X})$ such that $J_{\mathsf{sketch},\kappa}(\mu) \leq J_{\mathsf{sketch},\kappa}(\mu^0)$, and $n \geq c_d\delta_n^4$, where c_d depends polynomially on d, the following controls hold with probability $1-\alpha$ over the joint draws of the sample $\boldsymbol{Z}$ and the sketches $(\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_m)$.*

- *Control of the far region:*

$$|\mu|(\mathcal{F}^{\mathsf{reg}}(r)) \leq C'_{\mathsf{pivot}}C_{\mathsf{switch}}\frac{512}{69}C_{\alpha,m}\left(\frac{1}{\delta_n^2}\right)\sqrt{s_0}\,, \tag{2.14}$$

- *Control of all the near regions: for all $k \in [s_0]$,*

$$|\mu(\mathcal{N}_k^{\mathsf{reg}}(r)) - a_k^0| \leq C'_{\mathsf{pivot}}C_{\mathsf{switch}}\frac{1536}{69}C_{\alpha,m}\left(\frac{1}{\delta_n^2}\right)\sqrt{s_0}\,, \tag{2.15}$$

- *Detection level: for all Borelian $A \subset \mathcal{X}$ such that $|\mu|(A) > C'_{\mathsf{pivot}}C_{\mathsf{switch}}\frac{512}{69}C_{\alpha,m}\left(\frac{1}{\delta_n^2}\right)\sqrt{s_0}$, there exists $\boldsymbol{x}_k^0$ such that*

$$\min_{\boldsymbol{t}\in A}\mathfrak{d}_{\mathfrak{g}}(t, \boldsymbol{x}_k^0) \leq r\,, \tag{2.16}$$

The detailed proof is given in Appendix A.8.

2.4 Extensions

An important consequence of the kernel switch is that one can leverage the previous results for any translation invariant mixture model beyond Gaussian, and any smoothing kernel λ_τ beyond the sinus cardinal, as long as the resulting K_{mod} verifies Equation (2.6–$\mathbf{H}_\phi$). Since the latter is translation invariant, it suffices to show that its Fourier transform is not "too localized". For the sinc-4 pivot, this means $\mathfrak{F}[K_{\mathsf{mod}}]$ should contain low-frequencies and be bounded below so that C_{switch} is finite. One could also consider using another translation invariant pivot kernel, for the Gaussian pivot of Equation (1.9b) (not treated in this paper) this means $\mathfrak{F}[K_{\mathsf{mod}}]$ has a slower decay than the Gaussian.

3 Effective near regions, kernel switch and near-optimal solutions

This section presents the core theoretical contributions of our work, establishing rigorous guarantees for the BLASSO estimator. We begin by detailing the key assumptions underpinning our analysis, notably the local positive curvature (**LPC**) property for pivot kernels and conditions for sketching. Our main results then demonstrate three significant advancements: first, we introduce the concept of *effective near regions*, showing that the localisation accuracy of BLASSO adapts to the noise level, improving as more data becomes available. Second, we formalize the *kernel switch* principle, which allows leveraging well-characterized pivot kernels to analyze BLASSO problems with different, potentially more complex, model kernels, provided a specific embedding condition holds. This greatly expands the range of problems amenable to theoretical guarantees. Third, we prove that these statistical error bounds apply not only to exact minimizers of the BLASSO objective but also to *near-optimal solutions*, a crucial finding for practical algorithm design where exact optimisation is often infeasible. Finally, we provide a dedicated analysis for the sinc-4 kernel, proving it satisfies the necessary **LPC** and sketching assumptions, thereby enabling its use as a robust pivot kernel, particularly for mixture model estimation as discussed in Section 2.

3.1 Assumptions

First, we introduce the local positive curvature assumption, which is a key property for BLASSO analysis. This assumption is crucial to establish the statistical guarantees of our results. We also define the Fisher-Rao distance, which is a Riemannian distance associated with a metric tensor denoted $\mathfrak{g}_{\boldsymbol{x}}$ and a kernel $K(\cdot,\cdot)$.

Fisher-Rao distance for general kernels Let $K(\cdot,\cdot)$ be a real valued kernel of positive type. Define the metric tensor (identified as a psd matrix) $\mathfrak{g}_{\boldsymbol{x}} \coloneqq \nabla_1\nabla_2 K(\boldsymbol{x},\boldsymbol{x}) \in \mathbb{R}^{d\times d}$, where ∇_i is the gradient with respect to the ith variable. Define the Fisher-Rao distance by

$$\mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t}) \coloneqq \inf_p \int_0^1 \sqrt{p'(u)^\top \mathfrak{g}_{p(u)} p'(u)}\,\mathrm{d}u\,, \tag{3.1}$$

where the infimum is taken over smooth paths $p : [0,1] \to \mathcal{X}$ such that $p(0) = \boldsymbol{s}$ and $p(1) = \boldsymbol{t}$. The Fisher-Rao distance is thus the infimum of the length of paths connecting two points in the parameter space $\mathcal{X}$.

Remark 3.1 (Metric of translation invariant kernels)**.** *The Fisher-Rao distance is a Riemannian distance associated with the metric tensor $\mathfrak{g}_{\boldsymbol{x}}$ and the kernel $K(\cdot,\cdot)$. For translation invariant kernels defined by*

$$K(\boldsymbol{s},\boldsymbol{t}) = \rho(\boldsymbol{s}-\boldsymbol{t}), \tag{3.2a}$$

the metric tensor is constant and the Fisher-Rao distance is the Mahalanobis distance given by a positive definite matrix $\mathfrak{g}$ which does not depend on the points $\boldsymbol{x}$, it holds

$$\mathfrak{g} = -\nabla^2\rho(0)\,, \tag{3.2b}$$

$$\mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t}) = \sqrt{(\boldsymbol{s}-\boldsymbol{t})^\top \mathfrak{g}(\boldsymbol{s}-\boldsymbol{t})}\,. \tag{3.2c}$$

Local positive curvature assumption We now recall the local positive curvature assumption **LPC** as presented in Poon et al. (2023). The **LPC** property ensures that the kernel $K(\cdot,\cdot)$ behaves well in terms of its derivatives and curvature properties with respect to some Fisher-Rao distance, which is essential for the convergence of the BLASSO estimator.

Assumption 1 (**LPC** with parameters s_0, Δ_0, r_0, $\bar{\varepsilon}_0$ and $\bar{\varepsilon}_2$)**.** *A real valued kernel $K(\cdot,\cdot)$ of positive type satisfies the local positive curvature assumption (**LPC**) if the following holds:*

- *Assume that, for all $0 \leq i,j \leq 2$ such that $i+j \leq 3$,*

$$B_{ij} := \sup_{\boldsymbol{s},\boldsymbol{t}\in\mathcal{X}} \left\| K^{(i,j)}(\boldsymbol{s},\boldsymbol{t}) \right\|_{\boldsymbol{s},\boldsymbol{t}} < +\infty\,. \tag{3.3a}$$

  and denote $B_i := 1 + B_{0i} + B_{1i}$ where $K^{(i,j)}(\boldsymbol{s},\boldsymbol{t})$ is the covariant derivative[2] of order i with respect to the first variable $\boldsymbol{s}$ and of order j with respect to the second variable $\boldsymbol{t}$, and $\|\cdot\|_{\boldsymbol{s},\boldsymbol{t}}$ denotes the operator norm with respect to the corresponding metric tensor given by $K(\cdot,\cdot)$ on the tangent spaces at $\boldsymbol{s}$ and $\boldsymbol{t}$.

- *Assume that there exists $r_0 \in (0, 1/\sqrt{B_{02}})$ such that $K(\cdot,\cdot)$ has positive* curvature constants defined as

$$\begin{aligned}
\bar{\varepsilon}_0 &:= \frac{1}{2}\sup_{\varepsilon\geq 0}\left\{\varepsilon \,:\, K(\boldsymbol{s},\boldsymbol{t}) \leq 1-\varepsilon\,,\ \forall \boldsymbol{s},\boldsymbol{t}\in\mathcal{X} \text{ s.t. } \mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t}) \geq r_0\right\},\\
\bar{\varepsilon}_2 &:= \frac{1}{4}\sup_{\varepsilon\geq 0}\left\{\varepsilon \,:\, -K^{(0,2)}(\boldsymbol{s},\boldsymbol{t})[\boldsymbol{v},\boldsymbol{v}] \geq \varepsilon\|\boldsymbol{v}\|_{\boldsymbol{t}}^2\,,\ \forall \boldsymbol{v}\in\mathbb{T}_{\boldsymbol{t}}\,,\ \forall \boldsymbol{s},\boldsymbol{t}\in\mathcal{X} \text{ s.t. } \mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t}) < r_0\right\},
\end{aligned} \tag{3.3b}$$

  with $\mathfrak{d}_{\mathfrak{g}}$ the Fisher-Rao distance associated with the metric tensor of $K(\cdot,\cdot)$ and $\mathbb{T}_{\boldsymbol{t}}$ the tangent space at point $\boldsymbol{t}$.

- *Assume that there exists $s_0 \geq 2$ such that the set*

$$\left\{\Delta \,:\, 32\sum_{l=2}^{s_0}\|K^{(i,j)}(\boldsymbol{x}_1,\boldsymbol{x}_l)\|_{\boldsymbol{x}_1,\boldsymbol{x}_l} \leq \min\left(\frac{\bar{\varepsilon}_0}{B_0},\frac{2\bar{\varepsilon}_2}{B_2}\right),\ \forall (i,j)\in\{0,1\}\times\{0,2\},\forall\{\boldsymbol{x}_l\}_{l=1}^{s_0}\in\mathcal{S}_\Delta\right\} \tag{3.3c}$$

  is not empty, with $\mathcal{S}_\Delta := \left\{\{\boldsymbol{x}_l\}_{l=1}^{s_0}\in\mathcal{X}^{s_0} \,:\, \min_{k\neq l}\mathfrak{d}_{\mathfrak{g}}(\boldsymbol{x}_k,\boldsymbol{x}_l)\geq\Delta\right\}$. Denote Δ_0 its infimum.

*In this case, the kernel $K(\cdot,\cdot)$ is said to satisfy the **LPC** with parameters s_0, Δ_0, r_0, $\bar{\varepsilon}_0$ and $\bar{\varepsilon}_2$.*

Pivot kernels Next, we introduce the concept of *pivot kernels*, which are a special class of kernels that can be used to analyze the BLASSO problem. Admissible model kernels are those such that the kernel switch condition (Assumption 2) defined below holds. This kernel switch allows us to leverage the properties of the pivot kernel to analyze the model kernel, which may be more complex.

The model kernel introduced in Definition 1.2 is a kernel of positive type associated to a unique Reproducing Kernel Hilbert Space (RKHS) which is denoted by $\mathcal{H}_{\text{mod}}$. One can prove that $\mathcal{H}_{\text{mod}}$ is separable and isometric to the closure in $\mathcal{F}$ of $\mathrm{Im}(F)$, the range of F (De Castro et al. 2023, Appendix A.1),

$$\left(\mathcal{H}_{\text{mod}}, \|\cdot\|_{\mathcal{H}_{\text{mod}}}\right) \simeq \left(\overline{\mathrm{Im}(F)}, \|\cdot\|_{\mathcal{F}}\right) \subseteq \left(\mathcal{F}, \|\cdot\|_{\mathcal{F}}\right),$$

where $\simeq$ denotes an isometry mapping.

Let $F^\star : \mathcal{F} \to (\mathcal{M}(\mathcal{X}), \|\cdot\|_{\mathrm{TV}})^\star$ be the dual linear map of $F : \mathcal{M}(\mathcal{X}) \to \mathcal{F}$. One can prove (De Castro et al. 2023, Lemma A.2) that

$$F^\star \,:\, c\in\mathcal{F} \mapsto \eta_c \in \left(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty\right) \tag{3.4a}$$

is a continuous linear operator, and

$$\langle c, F\nu\rangle_{\mathcal{F}} = \int_{\mathcal{X}} \langle c, F\delta_{\boldsymbol{t}}\rangle_{\mathcal{F}}\,\mathrm{d}\nu(t) = \langle F^\star c,\nu\rangle_{\mathcal{C}(\mathcal{X}),\mathcal{M}(\mathcal{X})} = \int_{\mathcal{X}}\eta_c\,\mathrm{d}\nu(t) \quad \text{for all } c\in\mathcal{F},\ \nu\in\mathcal{M}(\mathcal{X})\,. \tag{3.4b}$$

Now, define

$$\forall c\in\mathcal{F}\,,\quad \eta_c \,:\, t\in\mathcal{X}\mapsto\langle c, F\delta_{\boldsymbol{t}}\rangle_{\mathcal{F}}\in\mathbb{R}\,, \tag{3.4c}$$

we have the following useful proposition.

Proposition 3.1. *If (1.3–$\mathbf{H}_{\text{mod}}$) holds then for all $\eta\in\mathcal{H}_{\text{mod}}$, there exists a unique $c\in\overline{\mathrm{Im}(F)}$ such that $\|\eta\|_{\mathcal{H}_{\text{mod}}} = \|c\|_{\mathcal{F}}$ and $\eta = \eta_c$. Furthermore, it holds that $\eta = F^\star c$.*

[2] We refer to Poon et al. 2023, Section 4.1 for further details on covariant derivatives and their operator norm, and to Appendix B.1 which gives their expression for translation invariant kernels.

Proof. See Appendix B.3 □

We can now turn to the the definition of a pivot kernel $K_{\text{pivot}}(\cdot,\cdot)$ with respect to the model kernel $K_{\text{mod}}(\cdot,\cdot)$.

Assumption 2 (Pivot kernels)**.** *A real valued kernel $K_{\text{pivot}}(\cdot,\cdot)$ of positive type is a pivot kernel with respect to $K_{\text{mod}}(\cdot,\cdot)$ if a) its values at points $(\boldsymbol{t},\boldsymbol{t})$ are normalized, namely*

$$\forall \boldsymbol{t} \in \mathcal{X}\,, \quad K_{\text{pivot}}(\boldsymbol{t},\boldsymbol{t}) = 1\,; \tag{3.5a}$$

and b) its RKHS, denoted by $\mathcal{H}_{\text{pivot}}$, is such that $\mathcal{H}_{\text{pivot}} \subseteq \mathcal{H}_{\text{mod}}$, and it holds that the following identity map $\mathrm{Id}_{\text{pivot,mod}} : (\mathcal{H}_{\text{pivot}}, \|\cdot\|_{\mathcal{H}_{\text{pivot}}}) \to (\mathcal{H}_{\text{mod}}, \|\cdot\|_{\mathcal{H}_{\text{mod}}})$ is continuous, namely

$$C_{\text{switch}} := \left\|\mathrm{Id}_{\text{pivot,mod}}\right\|_{\text{op}} = \sup_{\eta \in \mathcal{H}_{\text{pivot}} \setminus \{0\}} \frac{\|\eta\|_{\mathcal{H}_{\text{mod}}}}{\|\eta\|_{\mathcal{H}_{\text{pivot}}}} < \infty\,. \tag{3.5b}$$

Tail bound assumption on the sketching function derivatives. The third and final assumption is a tail bound on the derivatives of the sketching function $\psi_{\boldsymbol{\omega}}$ defined in Equation (1.12–$\mathbf{H}_{\psi_{\boldsymbol{\omega}},\Lambda}$), which also depends on the sketching distribution Λ. It ensures that the derivatives of the sketching function do not grow too large with respect to the Fisher-Rao distance, which is important to control the sketching error. The tail bounds are defined in terms of a so-called survival function $\bar{\mathbb{P}}_j(t)$, which quantifies the probability that the derivatives of the sketching function exceed a certain threshold t.

Assumption 3 (Tail bound on the sketching function derivatives)**.** *The function $\psi_{\boldsymbol{\omega}}$ and the law $\boldsymbol{\omega} \sim \Lambda$ satisfy the derivatives tail bounds assumption if: a) the kernel $K(\cdot,\cdot)$ defined by (1.12–$\mathbf{H}_{\psi_{\boldsymbol{\omega}},\Lambda}$) satisfies the local positive curvature assumption (Assumption 1) with some parameters s_0, Δ_0, r_0, $\bar{\varepsilon}_0$ and $\bar{\varepsilon}_2$; and b) there exists a constant $C > 0$ which depends on $K(\cdot,\cdot)$ and polynomially on d such that, for all $\alpha \in (0,1)$, there exist positive reals $L_0, L_1, L_2, L_3 > 0$ which may depend on α verifying:*

$$\sum_{j=0}^{3} \bar{\mathbb{P}}_j(L_j) \leq \frac{\min(\bar{\varepsilon}_0, \bar{\varepsilon}_2, \alpha)}{m_0} \tag{3.6a}$$

$$\max_{j=0}^{3} \left(L_j^2 \sum_{i=0}^{3} \bar{\mathbb{P}}_i(L_i) + 6 \int_{L_j}^{\infty} t\, \bar{\mathbb{P}}_j(t)\, \mathrm{d}t \right) \leq \frac{\min(\bar{\varepsilon}_0, \bar{\varepsilon}_2)}{m_0} \tag{3.6b}$$

$$m_0 := C s_0 \left(C_1 \log(s_0) \log\left(\frac{s_0}{\alpha}\right) + C_2 \log\left(\frac{(s_0 N)^d}{\alpha}\right)\right) \tag{3.6c}$$

where

$$\bar{\mathbb{P}}_j(t) := 1 - \mathbb{P}_{\boldsymbol{\omega}\sim\Lambda}\left(\sup_{\boldsymbol{x}\in\mathcal{X}} \|\psi_{\boldsymbol{\omega}}^{(j)}(\boldsymbol{x})\|_{\boldsymbol{x}} \leq t \right) \tag{3.6d}$$

$$\begin{aligned}
N &:= \frac{|\mathcal{X}| L_1}{\bar{\varepsilon}_0} + \frac{r_0 L_3 L_0 + L_2}{\bar{\varepsilon}_2} \\
C_1 &:= (L_0^2 + L_1^2)(B_0/\bar{\varepsilon}_0^2 + B_2/\bar{\varepsilon}_2^2) \\
C_2 &:= L_0^2/\bar{\varepsilon}_0^2 + L_{01} L_0/\bar{\varepsilon}_0 + L_2^2/\bar{\varepsilon}_2^2 + L_{01} L_2/\bar{\varepsilon}_2 \\
L_{ij} &:= \sqrt{L_i^2 + L_j^2} \\
|\mathcal{X}| &:= \sup_{\boldsymbol{s},\boldsymbol{t}\in\mathcal{X}} \mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t})
\end{aligned}$$

with $\psi_{\boldsymbol{\omega}}^{(j)}(\boldsymbol{x})$ the covariant derivative of order j at $\boldsymbol{x}$, $\mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t})$ the Fisher-Rao distance, $\|\cdot\|_{\boldsymbol{x}}$ the norm of the corresponding metric tensor given by $K(\cdot,\cdot)$ on the tangent space at $\boldsymbol{x}$.

3.2 Main theorem

Theorem 1 (Estimation error bounds for BLASSO)**.** *Let $\mathcal{X}$ be a compact set of $\mathbb{R}^d$, let $\mathcal{F}$ be a separable Hilbert space, and let $K_{\text{pivot}}(\cdot,\cdot)$ be a kernel of positive type which satisfies Assumption 1 with parameters s_0, Δ_0, r_0, $\bar{\varepsilon}_0$ and $\bar{\varepsilon}_2$, and which satisfies Assumption 2 with constant $C_{\text{switch}} < +\infty$.*

Let $F : \mathcal{M}(\mathcal{X}) \to \mathcal{F}$ be a linear map and let $\mu^0 \in \mathcal{M}(\mathcal{X})$, $\boldsymbol{y} \in \mathcal{F}$ and $\gamma \geq 0$ be such that

$$\mu^0 \in \mathcal{M}_{s_0,\Delta_0,\mathfrak{d}_{\mathfrak{g}}} \quad \text{and} \quad \gamma = \|\boldsymbol{y} - F\mu^0\|_{\mathcal{F}}\,,$$

with $\mathfrak{d}_{\mathfrak{g}}$ the Fisher-Rao distance associated to $K_{\text{pivot}}(\cdot,\cdot)$.

Consider J_κ given by (1.1–J_κ) where $\kappa = c_\kappa\gamma/\sqrt{s_0}$ with $c_\kappa > 0$ some constant, and let $\mu \in \mathcal{M}(\mathcal{X})$ be such that

$$J_\kappa(\mu) \leq J_\kappa(\mu^0)\,.$$

Then for each $r > 0$ such that $r < \min\left(r_0, \sqrt{\bar{\varepsilon}_0/\bar{\varepsilon}_2}\right)$, we have:

- *Control of the far region:*

$$|\mu|(\mathcal{F}^{\text{reg}}(r)) \leq \bar{c}_\kappa\Big(\frac{\gamma}{\bar{\varepsilon}_2 r^2}\Big)\sqrt{s_0}\,, \tag{3.7a}$$

- *Control of all the near regions: for all $k \in [s_0]$,*

$$|\mu(\mathcal{N}_k^{\text{reg}}(r)) - a_k^0| \leq \tilde{c}_\kappa\Big(\frac{\gamma}{\bar{\varepsilon}_2 r^2}\Big)\sqrt{s_0} + \hat{c}_\kappa\gamma\,, \tag{3.7b}$$

- *Detection level: for all Borelian $A \subset \mathcal{X}$ such that $|\mu|(A) > \bar{c}_\kappa\big(\frac{\gamma}{\bar{\varepsilon}_2 r^2}\big)\sqrt{s_0}$, there exists $\boldsymbol{x}_k^0$ such that*

$$\min_{\boldsymbol{t}\in A} \mathfrak{d}_{\mathfrak{g}}(t, \boldsymbol{x}_k^0) \leq r\,, \tag{3.7c}$$

where the far and near regions are defined by Definition 3 and

$$\bar{c}_\kappa := \frac{(1+\sqrt{2}\,C_{\text{switch}}c_\kappa)^2}{2c_\kappa} \geq 2\sqrt{2}\,C_{\text{switch}} \tag{3.7d}$$

$$\tilde{c}_\kappa = \bar{c}_\kappa \max(1, \bar{\varepsilon}_0) \tag{3.7e}$$

$$\hat{c}_\kappa = 2\sqrt{2}\,C_{\text{switch}}(1+\sqrt{2}C_{\text{switch}}c_\kappa) \tag{3.7f}$$

The proof of is given in Appendix A.5. In the statement above, one can see that the choice of the regularisation parameter is $\kappa \propto \gamma/\sqrt{s_0}$ up to a free constant c_κ. The selection of c_κ influences the constants in the error bounds and their dependency on s_0. We discuss several choices below:

Remark 3.2 (Tuning the regularisation parameter $\kappa = c_\kappa\gamma/\sqrt{s_0}$)**.** *The constant c_κ in the definition of the regularisation parameter κ affects the pre-factors in the error bounds (3.7a), (3.7b), and (3.7c).*

1. ***Optimal scaling with $\sqrt{s_0}$**: Choosing $c_\kappa = 1/(\sqrt{2}C_{\text{switch}})$ minimizes the constant $\bar{c}_\kappa$ to its lower bound $2\sqrt{2}C_{\text{switch}}$. This yields error bounds that scale as $\mathcal{O}(\gamma\sqrt{s_0}/(\bar{\varepsilon}_2 r^2))$ for the far region control and the leading term in the near region control. This is generally the preferred choice for optimal statistical rates.*
2. ***Alternative scaling with s_0**: If one chooses c_κ such that κ becomes independent of s_0 (e.g., by setting $c_\kappa = \sqrt{s_0}$), then $\kappa \propto \gamma$. This leads to $\bar{c}_\kappa \propto s_0$, degrading the error bounds to scale as $\mathcal{O}(\gamma s_0/(\bar{\varepsilon}_2 r^2))$. While this might simplify the expression for κ, it thus results in suboptimal statistical guarantees.*
3. ***Caution on noise level dependency**: It is crucial that κ maintains its proportionality to the noise level γ. For instance, attempting to make κ independent of γ (e.g., by setting $c_\kappa \propto 1/\gamma$) would lead to error bounds with a constant bias term, such as $\mathcal{O}((\text{constant} + C_{\text{switch}})\sqrt{s_0})$, which does not vanish as $\gamma \to 0$. This is undesirable in statistical applications where γ typically decreases with increasing sample size (e.g., mixture model estimation, see Section 2 and Proposition 2.1).*

3.3 New guarantees for sketched BLASSO

We now consider a sketched BLASSO given by a forward operator F as in (1.10a) with sketching function $\varphi_{\boldsymbol{\omega}}(\cdot)$, whose model kernel is given by (1.11b).

Theorem 2 (Estimation error bounds for sketched BLASSO)**.** *Let $\mathcal{X}$ be a compact set of $\mathbb{R}^d$. Let $\varphi_{\boldsymbol{\omega}}, \psi_{\boldsymbol{\omega}} : \mathcal{X} \to \mathbb{C}$ be functions and let Λ be a law on $\mathbb{R}^d$. Consider*

$$\begin{aligned} K_{\text{mod}}(\boldsymbol{s},\boldsymbol{t}) &= \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[\varphi_{\boldsymbol{\omega}}(s)\overline{\varphi}_{\boldsymbol{\omega}}(t)\right], \\ K_{\text{pivot}}(\boldsymbol{s},\boldsymbol{t}) &= \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[\psi_{\boldsymbol{\omega}}(s)\overline{\psi}_{\boldsymbol{\omega}}(t)\right]. \end{aligned} \tag{$\mathbf{H}_{\psi_{\boldsymbol{\omega}},\Lambda}$}$$

Assume that $K_{\text{pivot}}(\cdot,\cdot)$ satisfies Assumption 1 (with parameters s_0, Δ_0, r_0, $\bar{\varepsilon}_0$ and $\bar{\varepsilon}_2$) and Assumption 2 with constant $C_{\text{switch}} > 0$. Assume that the sketching function $\psi_{\boldsymbol{\omega}}$ satisfies Assumption 3 and that its derivatives are Λ-almost surely bounded up to order 3. Denote $C_{\text{sketch}} := 2C\max(C_1, C_2) > 0$.

Let $\alpha \in (0,1)$ and m be such that

$$m \geq C_{\text{sketch}}\ \max(d, \log(s_0))\ s_0\ \log\Big(\frac{\max(1,|\mathcal{X}|)\ s_0}{\alpha}\Big)\,, \tag{3.8}$$

where $|\mathcal{X}| = \sup_{\boldsymbol{s},\boldsymbol{t}\in\mathcal{X}} \mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t})$ is the diameter of $\mathcal{X}$ with respect to the Fisher-Rao distance $\mathfrak{d}_{\mathfrak{g}}$ given by $K_{\text{pivot}}(\cdot,\cdot)$.

Let $F : \mathcal{M}(\mathcal{X}) \to \mathbb{C}^m$ be given by

$$(F\mu)_i = \frac{1}{\sqrt{m}} \int_{\mathcal{X}} \varphi_{\boldsymbol{\omega}_i}(t)\mathrm{d}\mu(t)\,, \quad i = 1,\dots,m\,.$$

where the sequence of i.i.d. random vectors $(\boldsymbol{\omega}_i)_i$ is drawn with respect to Λ. Let $\mu^0 \in \mathcal{M}(\mathcal{X})$, $\boldsymbol{y}_{\text{sketch}} \in \mathbb{C}^m$ and $\gamma \geq 0$ be such that

$$\mu^0 \in \mathcal{M}_{s_0,\Delta_0,\mathfrak{d}_{\mathfrak{g}}} \quad \text{and} \quad \gamma = \|\boldsymbol{y}_{\text{sketch}} - F\mu^0\|_{\mathbb{C}^m}\,,$$

Consider J_κ given by (1.1–J_κ) where $\kappa = c_\kappa\gamma/\sqrt{s_0}$ with $c_\kappa > 0$ some constant, and let $\mu \in \mathcal{M}(\mathcal{X})$ be such that

$$J_\kappa(\mu) \leq J_\kappa(\mu^0)\,.$$

Then, there exists a constant $C'_{\text{pivot}} > 0$ which depends only on the kernel K_{pivot} such that, with probability at least $1-\alpha$ on the draw of $(\boldsymbol{\omega}_1,\dots,\boldsymbol{\omega}_m)$, for any $r > 0$ such that $r < \min\big(r_0, \sqrt{\bar{\varepsilon}_0/6\bar{\varepsilon}_2}\big)$, we have:

- *Control of the far region:*

$$|\mu|(\mathcal{F}^{\text{reg}}(r)) \leq \bar{c}_\kappa\Big(\frac{2\gamma}{3\bar{\varepsilon}_2 r^2}\Big)\sqrt{s_0}\,, \tag{3.9a}$$

- *Control of all the near regions: for all $k \in [s_0]$,*

$$|\mu(\mathcal{N}_k^{\text{reg}}(r)) - a_k^0| \leq \tilde{c}_\kappa\Big(\frac{2\gamma}{3\bar{\varepsilon}_2 r^2}\Big)\sqrt{s_0} + \hat{c}_\kappa\gamma\,, \tag{3.9b}$$

- *Detection level: for all borelian $A \subset \mathcal{X}$ such that $|\mu|(A) > \bar{c}_\kappa\big(\frac{2\gamma}{3\bar{\varepsilon}_2 r^2}\big)\sqrt{s_0}$, there exists $\boldsymbol{x}_k^0$ such that*

$$\min_{\boldsymbol{t}\in A} \mathfrak{d}_{\mathfrak{g}}(t, \boldsymbol{x}_k^0) \leq r\,, \tag{3.9c}$$

where the far and near regions are defined by Definition 3 and

$$\bar{c}_\kappa := \frac{(1 + C'_{\text{pivot}}C_{\text{switch}}c_\kappa)^2}{2c_\kappa} \geq 2C'_{\text{pivot}}C_{\text{switch}} \tag{3.9d}$$

$$\tilde{c}_\kappa = \bar{c}_\kappa \max(1, \bar{\varepsilon}_0/4) \tag{3.9e}$$

$$\hat{c}_\kappa = 2C'_{\text{pivot}}C_{\text{switch}}(1 + C'_{\text{pivot}}C_{\text{switch}}c_\kappa) \tag{3.9f}$$

The proof of is given in Appendix A.6.

Remark 3.3. *The sketch size (3.8) presented in the statement of Theorem 2 is given for pivot sketching functions $\psi_{\boldsymbol{\omega}}$ with Λ-almost surely bounded derivatives up to the order 3. One could weaken this hypothesis and only ask that the survival function $\bar{\mathbb{P}}_j(t)$ defined in Assumption 3 decays exponentially, which would incur additional logarithmic terms in s_0 and d in the sketch size (3.8).*

Remark 3.4. *For any $m \geq 1$, one can set*

$$\alpha = \max(1,|\mathcal{X}|)\ s_0\ \exp\left(-\frac{m}{C_{\text{sketch}}\ \max(d,\log(s_0))\ s_0}\right)$$

and the result of Theorem 2 holds with probability $1-\min(1,\alpha)$. We uncover that $m = \mathcal{O}(s_0)$ (up to multiplicative log factors in s_0 and multiplicative polynomial factors in d) are sufficient to get accurate statistical estimation error bounds.

3.4 The sinc-4 pivot: LPC and sketching

We consider the sinc-4 pivot $\Psi_\tau(\boldsymbol{x}-\boldsymbol{y}) = \mathrm{sinc}^4(\frac{\boldsymbol{x}-\boldsymbol{y}}{4\tau})$ defined in Equation (1.9d). This kernel has been studied in De Castro et al. 2021 without proving that the **LPC** holds. Since the theoretical analysis of the sketched BLASSO is built upon **LPC** (Poon et al. 2023), the statistical error bounds of sketched BLASSO with the sinc-4 pivot was an open question. The next theorem shows that the sinc-4 kernel Ψ_τ satisfies Assumption 1. Note that its Fisher-Rao distance is

$$\mathfrak{d}_{\mathfrak{g},\tau}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2\sqrt{3}\,\tau}\|\boldsymbol{x}-\boldsymbol{y}\|_2\,, \tag{3.10a}$$

which scales as τ^{-1}.

Theorem 3. *Given any $\tau > 0$ and $s_0 \geq 1$, the kernel $\Psi_\tau(\boldsymbol{x}-\boldsymbol{y}) = \mathrm{sinc}^4(\frac{\boldsymbol{x}-\boldsymbol{y}}{4\tau})$ on $\mathcal{X} = \mathbb{R}^d$ satisfies Assumption 1 with parameters s_0, $\Delta_0 = 42.66\, {s_0}^{1/4} d^{7/4}$, $r_0 = 1/(4\,d)$, $\bar{\varepsilon}_0 \geq 1/(32\,d^3)$ and $\bar{\varepsilon}_2 \geq 23/128$.*

The proof is given in Appendix A.3. Given a set of distinct points $\{\boldsymbol{x}_1^0, \ldots, \boldsymbol{x}_{s_0}^0\} \subset \mathbb{R}^d$, if

$$\tau \leq \frac{\min_{k,l} \|\boldsymbol{x}_k^0 - \boldsymbol{x}_l^0\|_2}{147.77\, {s_0}^{1/4} d^{7/4}} \tag{3.10b}$$

then any measure supported on these points is a sparse target measure (1.4) for which $\Delta_0 \leq 42.66\, {s_0}^{1/4} d^{7/4}$ and hence the **LPC** holds with $r_0 = 1/(4\,d)$, $\bar{\varepsilon}_0 \geq 1/(32\,d^3)$ and $\bar{\varepsilon}_2 \geq 23/128$. Condition (3.10b) is exactly (2.3–$\mathbf{H}_\tau$).

We now turn to the proof of Assumption 3 for the sinc-4 kernel. As a translation invariant kernel, we denote $f_\tau^{(4)}$ its spectral measure, detailed in Appendix C.5 and such that

$$\Psi_\tau(\boldsymbol{x}-\boldsymbol{y}) = \int e^{+\imath\omega^\top(\boldsymbol{x}-\boldsymbol{y})} f_\tau^{(4)}(\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{\omega}.$$

Since $\Psi_\tau(\boldsymbol{0}) = 1$ and Ψ_τ is a positive type kernel, $f_\tau^{(4)}$ is a probability density function. Thus, for any sketching distribution Λ, we may write the latter as an expectation in the form of Equation (1.12–$\mathbf{H}_{\psi_{\boldsymbol{\omega}},\Lambda}$)

$$\Psi_\tau(\boldsymbol{x}-\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[\psi_{\boldsymbol{\omega}}(\boldsymbol{x})\overline{\psi}_{\boldsymbol{\omega}}(\boldsymbol{y})\right], \quad \text{with sketching function } \psi_{\boldsymbol{\omega}}(\boldsymbol{t}) = e^{\imath\boldsymbol{\omega}^\top\boldsymbol{t}}\sqrt{\frac{f_\tau^{(4)}}{\Lambda}}(\boldsymbol{\omega}).$$

In order to prove Assumption 3 we need to exhibit $L_0, L_1, L_2, L_3 > 0$ such that Equations 3.6 is verified for the **LPC** constants of Theorem 3. This is the purpose of the following theorem.

Theorem 4. *For any bandwith τ, the sinc-4 kernel $\Psi_\tau(\boldsymbol{x}-\boldsymbol{y}) = \mathrm{sinc}^4(\frac{\boldsymbol{x}-\boldsymbol{y}}{4\tau})$ and its sketching function ψ_τ verify Assumption 3 with:*

$$m_0 = Cs_0\left(C_1\log(s_0)\log\left(\frac{s_0}{\alpha}\right) + C_2\log\left(\frac{(s_0N)^d}{\alpha}\right)\right) \tag{3.10c}$$

$$N \coloneqq |\mathcal{X}|32\sqrt{12}\sqrt{C_\Lambda}d^{7/2} + \frac{128}{23}(12\sqrt{12}C_\Lambda d^{1/2} + \sqrt{C_\Lambda}12d) \tag{3.10d}$$

$$C_1 \coloneqq (1+12d)C_\Lambda\left(1024d^6(2+\sqrt{12d}) + \frac{128^2}{23^2}(1+\sqrt{12d}+12d)\right) = \mathcal{O}\left(d^{15/2}C_\Lambda\right)$$

$$C_2 \coloneqq C_\Lambda\left(1024d^6 + 32d^3\sqrt{1+12d} + + \frac{128^2}{23^2}12^2d^2 + \frac{128}{23}12d\sqrt{1+12d}\right) = \mathcal{O}\left(d^6C_\Lambda\right)$$

$$C_\Lambda \coloneqq \sup_{\boldsymbol{\omega}\in[-\frac{1}{\tau},\frac{1}{\tau}]^d}\frac{f_\tau^{(4)}}{\Lambda}(\boldsymbol{\omega})$$

where $|\mathcal{X}| = \sup_{\boldsymbol{s},\boldsymbol{t}\in\mathcal{X}}\mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t})$ and C depends on Ψ_τ and polynomially on d. Moreover, derivatives of ψ_τ are almost surely bounded at any order.

The proof is given in Appendix A.4.

4 Conclusion and perspectives

This paper advances the theory of continuous sparse regularization on measures, providing novel theoretical guarantees for sketched problems, including mixture models. We introduce a key embedding constant, C_{switch}, which we explicitly characterize for translation-invariant kernels via a ratio of Fourier transforms for translation invariant kernels. Our analysis further reveals how the support of the solution belongs to so-called effective near regions that scale with the noise level and establishes formal error bounds for near-optimal solutions, validating them as reliable estimators in pratice.

These contributions open promising avenues for future research. Theoretically, a natural extension is to exhibit the kernel switch constant for non-translation invariant models, such as Gaussian mixtures with unknown covariance. Methodologically, our work motivates the development of practical conic particle gradient descent algorithms for the BLASSO. This pursuit will also require addressing the crucial challenge of selecting the regularization parameter κ and the sketch size in a data-driven manner, especially when the true number of components is unknown.

Finally, we note that our results are not limited to mixture models but can be applied to any continuous regression problem where some pivot kernel satisfies the **LPC**. This includes a wide range of applications in machine learning and statistics, such as density estimation, clustering, and regression tasks on continuous data (*e.g.,* tensor regression (Azaïs et al. 2024), shallow neural networks (Chizat 2019; Azaïs et al. 2024)).

Acknowledgments The authors would like to thank *Romane Giard* for precious comments and discussions on draft versions of this article.

Funding The work of R. Gribonval was partially supported by the AllegroAssai ANR project ANR-19-CHIA-0009 and the SHARP ANR Project ANR-23-PEIA-0008 of the PEPR IA, funded in the framework of the France 2030 program.

General notation

$[n]$	Set of integers $\{1, \dots, n\}$;

Continuous regression

$\mathcal{X}$	Compact set of $\mathbb{R}^d$;
$(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$	Continuous functions on $\mathcal{X}$;
$(\mathcal{M}(\mathcal{X}), , \|\cdot\|_{\mathrm{TV}})$	Radon measures on $\mathcal{X}$;
$(\mathcal{F}, \langle\cdot,\cdot\rangle_{\mathcal{F}})$	Separable Hilbert space;
κ	Regularisation parameter of (1.1–J_κ);
$\mathcal{M}_{s_0,\Delta_0,\mathfrak{d}_\mathfrak{g}}$	Class of models, Eq. (1.4);
Γ (resp. γ)	Noise (resp. noise level), Eq. (1.5);
$\mathcal{N}_k^{\mathrm{reg}}(r)$ (resp. $\mathcal{F}^{\mathrm{reg}}(r)$)	Near (resp. far) regions, Eq. (1.6);

Linear operators and kernels

K_{mod}	Model kernel, Eq. (1.2);
$(\mathcal{H}_{\mathsf{mod}}, \|\cdot\|_{\mathcal{H}_{\mathsf{mod}}})$	RKHS associated to K_{mod}, Eq. (B.2);
K_{pivot}	Pivot kernel, Eq. (3.5);
$(\mathcal{H}_{\mathsf{pivot}}, \|\cdot\|_{\mathcal{H}_{\mathsf{pivot}}})$	RKHS associated to K_{pivot};
$\varphi_{\boldsymbol{\omega}}$(resp. $\psi_{\boldsymbol{\omega}}$)	Model (resp. pivot) sketching functions;
Θ_Ω	Radial basis function kernel with covariance Ω, Eq. (1.9b);
Ψ_τ	Sinus cardinal kernel to the four, Eq. (1.9d);
Λ	Sketching distribution, Eq. (1.10b);
F	Forward operator, Eq. (1.10a);
$F^\star$	Adjoint operator of F;
$\mathfrak{F}[\cdot]$	Fourier transform;
$\mathfrak{F}^{-1}[\cdot]$	Inverse Fourier transform;

Differential geometry

$\nabla_1 K(\boldsymbol{s}, \boldsymbol{t})$	Gradient at point $\boldsymbol{s}$ of $\boldsymbol{u} \mapsto K(\boldsymbol{u}, \boldsymbol{t})$;
$\nabla_2 K(\boldsymbol{s}, \boldsymbol{t})$	Gradient at point $\boldsymbol{t}$ of $\boldsymbol{u} \mapsto K(\boldsymbol{s}, \boldsymbol{u})$;
$\mathfrak{g}_{\boldsymbol{x}}$	Metric tensor $\nabla_1 \nabla_2 K(\boldsymbol{x}, \boldsymbol{x})$ at point $\boldsymbol{x}$;
$\mathfrak{d}_\mathfrak{g}(\boldsymbol{s}, \boldsymbol{t})$	Fisher-Rao distance between $\boldsymbol{s}$ and $\boldsymbol{t}$ with respect to a kernel K with metric $\mathfrak{g}$, Eq. (3.1);
$\mathfrak{d}_{\mathfrak{g},\tau}(\boldsymbol{s}, \boldsymbol{t})$	Fisher-Rao distance between $\boldsymbol{s}$ and $\boldsymbol{t}$ with respect to the sinc-4 kernel Ψ_τ;

Table 2: List of notation

References

Arcones, Miguel A and Evarist Giné (1993). "Limit theorems for U-processes". In: *The Annals of Probability*, pp. 1494–1542 (cit. on pp. 14, 40).

Azais, Jean-Marc, Yohann De Castro, and Fabrice Gamboa (2015). "Spike detection from inaccurate samplings". In: *Applied and Computational Harmonic Analysis* 38.2, pp. 177–195 (cit. on pp. 5, 6, 10).

Azaïs, Jean-Marc, Federico Dalmao, and Yohann De Castro (2024). "Second Maximum of a Gaussian Random Field and Exact (t-) Spacing test". In: *arXiv preprint arXiv:2406.18397* (cit. on p. 24).

Bach, Francis (2017). "Breaking the curse of dimensionality with convex neural networks". In: *Journal of Machine Learning Research* 18.19, pp. 1–53 (cit. on p. 2).

Bach, Francis (2024). *Learning theory from first principles*. MIT press (cit. on p. 43).

Belhadji, Ayoub and Rémi Gribonval (2024). "Revisiting RIP Guarantees for Sketching Operators on Mixture Models". In: *Journal of Machine Learning Research* 25.55, pp. 1–68 (cit. on p. 8).

Bendory, Tamir, Shai Dekel, and Arie Feuer (2015). "Exact recovery of Dirac ensembles from the projection onto spaces of spherical harmonics". In: *Constructive Approximation* 42, pp. 183–207 (cit. on pp. 6, 7).

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518, pp. 859–877 (cit. on p. 11).

Boyd, Nicholas, Geoffrey Schiebinger, and Benjamin Recht (2017). "The alternating descent conditional gradient method for sparse inverse problems". In: *SIAM Journal on Optimization* 27.2, pp. 616–639 (cit. on p. 2).

Boyer, Claire, Yohann De Castro, and Joseph Salmon (2017). "Adapting to unknown noise level in sparse deconvolution". In: *Information and Inference: A Journal of the IMA* 6.3, pp. 310–348 (cit. on p. 2).

Boyer, Claire et al. (2019). "On representer theorems and convex regularization". In: *SIAM Journal on Optimization* 29.2, pp. 1260–1281 (cit. on p. 4).

Candès, Emmanuel J and Carlos Fernandez-Granda (2014). "Towards a mathematical theory of super-resolution". In: *Communications on pure and applied Mathematics* 67.6, pp. 906–956 (cit. on pp. 2, 5, 6, 10, 27).
Candès, Emmanuel J, Justin K Romberg, and Terence Tao (2006). "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on information theory* 52.2, pp. 489–509 (cit. on p. 2).
Celeux, Gilles, Sylvia Frühwirth-Schnatter, and Christian Robert (2019). "Model Selection for Mixture Models-Perspectives and Strategies". In: *Handbook of Mixture Analysis*. CRC Press (cit. on p. 11).
Chizat, Lenaic (2019). "Sparse optimization on measures with over-parameterized gradient descent". In: *arXiv preprint arXiv:1907.10300* (cit. on pp. 10, 11, 24).
De Castro, Yohann, Sébastien Gadat, and Clément Marteau (2023). "FastPart: Over-Parameterized Stochastic Gradient Descent for Sparse optimisation on Measures". In: *arXiv preprint arXiv:2312.05993* (cit. on pp. 4, 10, 11, 19).
De Castro, Yohann and Fabrice Gamboa (2012). "Exact reconstruction using Beurling minimal extrapolation". In: *Journal of Mathematical Analysis and applications* 395.1, pp. 336–354 (cit. on pp. 2, 6).
De Castro, Yohann et al. (2021). "SuperMix: Sparse regularization for mixtures". In: *The Annals of Statistics* 49.3, pp. 1779 –1809 (cit. on pp. 2, 3, 5, 7, 10–15, 23, 27, 40).
Denoyelle, Quentin et al. (2019). "The sliding frank-wolfe algorithm and its application to super-resolution microscopy". In: *Inverse Problems* (cit. on pp. 7, 10, 11).
Duval, Vincent and Gabriel Peyré (2015). "Exact Support Recovery for Sparse Spikes Deconvolution". In: *Foundations of Computational Mathematics* 15.5, pp. 1315–1355 (cit. on pp. 2, 3, 6, 10).
Duval, Vincent and Gabriel Peyré (2017). "Sparse spikes super-resolution on thin grids II: the continuous basis pursuit". In: *Inverse Problems* 33.9, p. 095008 (cit. on pp. 2, 6, 10).
Ekanadham, Chaitanya, Daniel Tranchina, and Eero P Simoncelli (2014). "A unified framework and method for automatic neural spike identification". In: *Journal of neuroscience methods* 222, pp. 47–55 (cit. on p. 2).
Foucart, Simon and Holger Rauhut (2013). *A Mathematical Introduction to Compressive Sensing* (cit. on pp. 2, 4, 10).
Fruhwirth-Schnatter, Sylvia, Gilles Celeux, and Christian P Robert (2019). *Handbook of mixture analysis*. Chapman and Hall/CRC (cit. on p. 11).
Gribonval, Rémi et al. (2021a). "Compressive statistical learning with random feature moments". In: *Mathematical Statistics and Learning* 3.2, pp. 113–164 (cit. on pp. 2, 8, 16).
Gribonval, Rémi et al. (2021b). "Statistical learning guarantees for compressive clustering and compressive mixture modeling". In: *Mathematical Statistics and Learning* 3.2, pp. 165–257 (cit. on p. 15).
Gribonval, Rémi et al. (2020). *Compressive Statistical Learning with Random Feature Moments* (cit. on p. 2).
Kingma, Diederik P (2014). "Auto-encoding variational bayes". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings* (cit. on p. 11).
McLachlan, Geoffrey J and Thriyambakam Krishnan (2007). *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons (cit. on p. 11).
McLachlan, Geoffrey J and David Peel (2004). *Finite mixture models*. John Wiley & Sons (cit. on p. 11).
Muandet, Krikamol et al. (2017). "Kernel mean embedding of distributions: A review and beyond". In: *Foundations and Trends® in Machine Learning* 10.1-2, pp. 1–141 (cit. on p. 12).
Poon, Clarice, Nicolas Keriven, and Gabriel Peyré (2019). "Support localization and the fisher metric for off-the-grid sparse regularization". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1341–1350 (cit. on pp. 4, 6–8).
Poon, Clarice, Nicolas Keriven, and Gabriel Peyré (2023). "The geometry of off-the-grid compressed sensing". In: *Foundations of Computational Mathematics* 23.1, pp. 241–327 (cit. on pp. 2–8, 10, 11, 13, 18, 19, 23, 27–31, 33, 41).
Rahimi, Ali and Benjamin Recht (2007). "Random features for large-scale kernel machines". In: *Advances in neural information processing systems* 20 (cit. on p. 15).
Steinwart, Ingo and Andreas Christmann (2008). *Support vector machines*. Springer Science & Business Media (cit. on pp. 28, 42).
Tang, Gongguo et al. (2013). "Compressed sensing off the grid". In: *IEEE transactions on information theory* 59.11, pp. 7465–7490 (cit. on p. 2).
Tomczak, Jakub M (2024). *Deep Generative Modeling*. Springer Cham (cit. on p. 11).
Wendland, Holger (2005). *Scattered data approximation*. Vol. 17. Cambridge university press (cit. on p. 43).
Zhang, Haizhang and Liang Zhao (2013). "On the inclusion relation of reproducing kernel Hilbert spaces". In: *Analysis and Applications* 11.02, p. 1350014 (cit. on p. 13).

A Proofs of the main results

A.1 Existence of pivot non-degenerate dual certificates

The crux of the statistical analysis lies in the construction of a *non-degenerate dual certificate* η^0, *i.e.* a sub-gradient of the TV-norm interpolating $\operatorname{sign}(a_k^0) = \pm 1$ at the support point t^k of μ^0, and with controlled behaviors in the near and far regions, see Figure 4. Both existence and construction of such objects have been the focus of a lot of attention in the continuous sparse recovery literature, see for example (Candès and Fernandez-Granda 2014; Poon et al. 2023; De Castro et al. 2021).

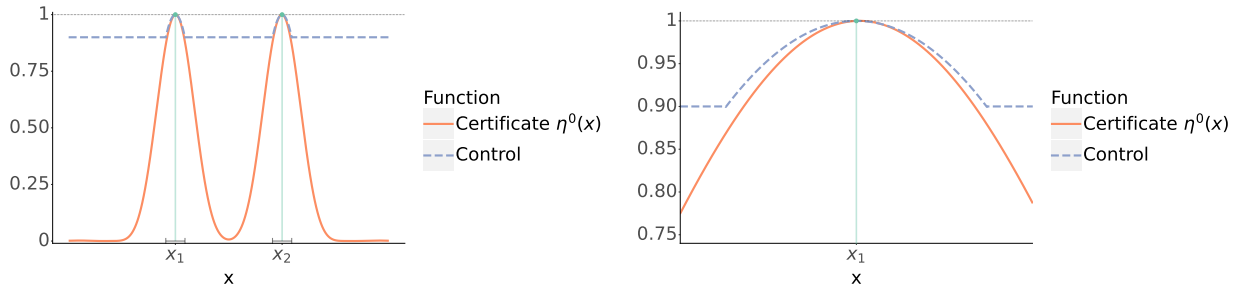

Figure 4: On the left, a one-dimensional illustration of the non-degenerate dual certificate η^0 in the near regions and far regions for some target μ^0 with two spikes x_1 and x_2, shown in light green. On the right, a zoomed location of the curvature control in the near region $\mathcal{N}_1^{\text{reg}}(r)$ around t_1. The certificates are drawn in orange while blue dotted lines illustrates the required control.

Definition 4 (Pivot non-degenerate certificates)**.** *Let $K_{\text{pivot}}(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a pivot kernel and let $\mu^0 \in \mathcal{M}_{s_0,\Delta_0,\mathfrak{d}_{\mathfrak{g}}}$ where the Fisher-Rao distance is given by $K_{\text{pivot}}(\cdot,\cdot)$.*

We say that η^0 is an $(\varepsilon_0, \varepsilon_2, r_0)$-pivot non-degenerate dual certificate if

- *There exists $c^0 \in \mathcal{F}$ such that $\eta^0 = F^\star c^0$,*
- *For all $i \in [s_0]$, $\eta^0(t_i^0) = \operatorname{sign}(a_i^0)$,*
- *For all $x \in \mathcal{F}^{\text{reg}}(r_0)$, $|\eta^0(x)| \leq 1 - \varepsilon_0$,*
- *For all $i \in [s_0]$, for all $x \in \mathcal{N}_i^{\text{reg}}(r_0)$, $|\eta^0(x)| \leq 1 - \varepsilon_2 \mathfrak{d}_{\mathfrak{g}}(x, t_i^0)^2$,*

where $\mathfrak{d}_{\mathfrak{g}}$ is the Fisher-Rao distance associated to $K_{\text{pivot}}(\cdot,\cdot)$, and the near and far regions are those of $(t_i^0)_{i=1}^{s_0}$.

For all $i \in [s_0]$, we say that η_i^0 is an $(\varepsilon_0, \varepsilon_2, r_0)$-pivot non-degenerate localizing certificate at point t_i^0 if

- *There exists $c_i^0 \in \mathcal{F}$ such that $\eta_i^0 = F^\star c_i^0$,*
- *For all $x \in \mathcal{F}^{\text{reg}}(r_0)$, $|\eta_i^0(x)| \leq 1 - \varepsilon_0$,*
- *For all $j \in [s_0]$, for all $x \in \mathcal{N}_j^{\text{reg}}(r_0)$, $|\delta_{ij} - \eta_i^0(x)| \leq \varepsilon_2 \mathfrak{d}_{\mathfrak{g}}(x, t_j^0)^2$,*

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

Theorem 5 (Existence of pivot non-degenerate dual certificates)**.** *Let $\mathcal{X}$ be a compact set of $\mathbb{R}^d$, let $\mathcal{F}$ be a separable Hilbert space, and let $K_{\text{pivot}}(\cdot,\cdot)$ be a positive type kernel that satisfies Assumption 1 with parameters s_0, Δ_0, r_0, $\bar{\varepsilon}_0$ and $\bar{\varepsilon}_2$, and Assumption 2 with constant $C_{\text{switch}} > 0$ with respect to some model kernel $K_{\text{mod}}(\cdot,\cdot)$ which satisfies (1.3–$\mathbf{H}_{\text{mod}}$).*

Then there exist an $(\bar{\varepsilon}_0, \bar{\varepsilon}_2, r_0)$-pivot non-degenerate dual certificate such that $\|c^0\|_{\mathcal{F}} \leq \sqrt{2} C_{\text{switch}} \sqrt{s_0}$ and $(\bar{\varepsilon}_0, \bar{\varepsilon}_2, r_0)$-pivot non-degenerate localizing certificates at points t_i^0 such that $\|c_i^0\|_{\mathcal{F}} \leq \sqrt{2} C_{\text{switch}}$.

Proof. This proof is an adaptation of the proof of Poon et al. 2023, Theorem 2, Page 266 for the pivot kernel. Denote by $f_{\text{pivot}}(x) \coloneqq K_{\text{pivot}}(x, \cdot)$ the canonical feature map of $\mathbb{H}_{\text{pivot}}$.

Consider some coefficients $(\alpha_1, \alpha_2) \in \mathbb{R}^{s_0} \times \mathbb{R}^{ds_0}$ such that

$$\eta^0(x) \coloneqq \sum_{j=1}^{s_0} \alpha_{1,j} K_{\text{pivot}}(\boldsymbol{x}_j^0, x) + \langle \alpha_{2,j}, \nabla_1 K_{\text{pivot}}(\boldsymbol{x}_j^0, x) \rangle_{\mathbb{R}^d} \tag{A.1}$$

satisfies $\nabla \eta^0(\boldsymbol{x}_j^0) = 0$ and $\eta^0(\boldsymbol{x}_j^0) = \operatorname{sign}(a_j^0)$, for $j \in [s_0]$. These $s_0(d+1)$ constraints can be written as a linear system

$$\Upsilon \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} (\operatorname{sign}(a_j^0))_{j=1}^{s_0} \\ \mathbf{0}_{s_0 d} \end{pmatrix} =: \mathbf{u},$$

where $\Upsilon \in \mathbb{R}^{ds_0 \times ds_0}$ is a real symmetric matrix defined as

$$\Upsilon = \begin{pmatrix} (K_{\text{pivot}}(\boldsymbol{x}_i^0, \boldsymbol{x}_j^0))_{1 \le i,j \le s_0} & (\nabla_2 K_{\text{pivot}}(\boldsymbol{x}_i^0, \boldsymbol{x}_j^0)^\top)_{1 \le i,j \le s_0} \\ (\nabla_1 K_{\text{pivot}}(\boldsymbol{x}_i^0, \boldsymbol{x}_j^0))_{1 \le i,j \le s_0} & (\nabla_1 \nabla_2 K_{\text{pivot}}(\boldsymbol{x}_i^0, \boldsymbol{x}_j^0))_{1 \le i,j \le s_0} \end{pmatrix}.$$

Then, define

$$\mathbf{f}(x) \coloneqq \left((K_{\text{pivot}}(\boldsymbol{x}_i^0, x))_{i=1}^{s_0}, \left(\nabla_1 K_{\text{pivot}}(\boldsymbol{x}_i^0, x)^\top\right)_{i=1}^{s_0} \right)^\top \in \mathbb{R}^{s_0(d+1)}$$

Since the mixed partial derivative $\partial_i \partial_j K_{\text{pivot}}$ exists and are continuous, Steinwart and Christmann (2008, Lemma 4.34) ensures that the feature map $f_{\text{pivot}} : \mathbb{R}^d \to \mathcal{H}_{\text{pivot}}$ is continuously differentiable and we denote by $\nabla f_{\text{pivot}}(\boldsymbol{s}) = (\partial_i f_{\text{pivot}}(\boldsymbol{s}))_{i=1}^d \in \mathcal{H}_{\text{pivot}}^d$ the vector of partial derivatives, the latter verifying: $\langle \partial_i f_{\text{pivot}}(\boldsymbol{s}), \partial_j f_{\text{pivot}}(\boldsymbol{t}) \rangle_{\mathcal{H}_{\text{pivot}}} = \partial_i \partial_j K_{\text{pivot}}(\boldsymbol{s}, \boldsymbol{t})$.

We can thus define $\mathbf{f} = \left((f_{\text{pivot}}(\boldsymbol{x}_i^0))_{i=1}^{s_0}, \left(\nabla f_{\text{pivot}}(\boldsymbol{x}_i^0)^\top\right)_{i=1}^{s_0} \right)^\top \in (\mathbb{H}_{\text{pivot}})^{s_0(d+1)}$ and we introduce the following notation:

$$\forall p \ge 1, \forall q \ge 1, \forall (a_i)_{i=1}^p \in (\mathbb{H}_{\text{pivot}})^p, \forall (b_j)_{j=1}^q \in (\mathbb{H}_{\text{pivot}})^q, \quad \langle a \otimes b \rangle_{\mathbb{H}_{\text{pivot}}} \coloneqq (\langle a_i, b_j \rangle_{\mathbb{H}_{\text{pivot}}})_{i \in [p], j \in [q]}.$$

Note that

$$\Upsilon = \begin{pmatrix} (\langle f_{\text{pivot}}(\boldsymbol{x}_i^0), f_{\text{pivot}}(\boldsymbol{x}_j^0) \rangle_{\mathbb{H}_{\text{pivot}}})_{1 \le i,j \le s_0} & (\langle f_{\text{pivot}}(\boldsymbol{x}_i^0) \otimes \nabla f_{\text{pivot}}(\boldsymbol{x}_j^0) \rangle_{\mathbb{H}_{\text{pivot}}})_{1 \le i,j \le s_0} \\ (\langle \nabla f_{\text{pivot}}(\boldsymbol{x}_i^0) \otimes f_{\text{pivot}}(\boldsymbol{x}_j^0) \rangle_{\mathbb{H}_{\text{pivot}}})_{1 \le i,j \le s_0} & (\langle \nabla f_{\text{pivot}}(\boldsymbol{x}_i^0) \otimes \nabla f_{\text{pivot}}(\boldsymbol{x}_j^0) \rangle_{\mathbb{H}_{\text{pivot}}})_{1 \le i,j \le s_0} \end{pmatrix}$$
$$\mathbf{f}(x) = \langle \mathbf{f}, f_{\text{pivot}}(x) \rangle_{\mathbb{H}_{\text{pivot}}}.$$

We deduce that Υ is the Gram matrix of $\mathbf{f}$, with respect to the dot product of $\mathbb{H}_{\text{pivot}}$, namely $\Upsilon = \langle \mathbf{f} \otimes \mathbf{f} \rangle_{\mathbb{H}_{\text{pivot}}}$.

Assuming that Υ is invertible, we can rewrite (A.1) as

$$\eta^0(x) = \left(\Upsilon^{-1}\mathbf{u}\right)^\top \mathbf{f}(x) \tag{A.2a}$$
$$\eta^0(x) = \langle \mathbf{f}^\top \Upsilon^{-1} \mathbf{u}, f_{\text{pivot}}(x) \rangle_{\mathbb{H}_{\text{pivot}}} \tag{A.2b}$$

and we deduce that $\eta = \mathbf{f}^\top \Upsilon^{-1} \mathbf{u}$ where the equality holds in $\mathbb{H}_{\text{pivot}}$, by the reproducing property.

Hence

$$\|\eta^0\|_{\mathbb{H}_{\text{pivot}}}^2 = \langle \mathbf{f}^\top \Upsilon^{-1} \mathbf{u}, \mathbf{f}^\top \Upsilon^{-1} \mathbf{u} \rangle_{\mathbb{H}_{\text{pivot}}} = \mathbf{u}^\top \Upsilon^{-1} \mathbf{u},$$

using that $\langle \mathbf{f} \otimes \mathbf{f} \rangle_{\mathbb{H}_{\text{pivot}}} = \Upsilon$.

We also define the block diagonal normalisation matrix $D_{\mathfrak{g}} \in \mathbb{R}^{s_0(d+1) \times s_0(d+1)}$ as

$$D_{\mathfrak{g}} \coloneqq \begin{pmatrix} \mathrm{Id}_s & & & \\ & \mathfrak{g}_{\boldsymbol{x}_1}^{-\frac{1}{2}} & & \\ & & \ddots & \\ & & & \mathfrak{g}_{\boldsymbol{x}_s}^{-\frac{1}{2}} \end{pmatrix} \tag{A.2c}$$

and $\tilde{\Upsilon} \coloneqq D_{\mathfrak{g}} \Upsilon D_{\mathfrak{g}}$ which has constant value 1 along its diagonal.

We deduce that

$$\|\eta^0\|^2_{\mathbb{H}_{\text{pivot}}} \leq \left\|\tilde{\Upsilon}^{-1}\right\| \|D_{\mathfrak{g}}\mathbf{u}\|_2^2 \leq 2s_0\,. \tag{A.2d}$$

using that $\|\tilde{\Upsilon}^{-1}\| \leq 2$ by Poon et al. 2023, Lemma 3.

Invoke Proposition 3.1, (A.2d) and (3.5b) to get that

$$\|c_0\|_{\mathcal{F}} = \|\eta^0\|_{\mathbb{H}_{\text{mod}}} \leq C_{\text{switch}} \|\eta^0\|_{\mathbb{H}_{\text{pivot}}} \leq C_{\text{switch}} \sqrt{2s_0} \tag{A.2e}$$

By Poon et al. 2023, Theorem 2, η^0 is a $(\varepsilon_0, \varepsilon_2, r_0)$-pivot non-degenerate dual certificate (note that we slightly changed the values of $\varepsilon_0, \varepsilon_2$ of Poon et al. 2023, Assumption in Assumption 1).

The existence of $(\varepsilon_0, \varepsilon_2, r_0)$-pivot non-degenerate localizing certificates follow the same line. In this case $\mathbf{u}^\top = ((1_{\{i=j\}}\text{sign}(a_j^0))_{j\in[s_0]}^\top, \mathbf{0}_{s_0 d}^T)$ and hence $\|D_{\mathfrak{g}}\mathbf{u}\|_2 = 1$. □

A.2 Existence of sketch pivot non-degenerate dual certificates

Definition 5 (Sketch pivot non-degenerate certificates)**.** *Let $\psi_{\boldsymbol{\omega}} : \mathcal{X} \to \mathbb{C}$ be a function and let Λ be a law on $\mathbb{R}^d$. Consider*

$$K_{\text{pivot}}(\boldsymbol{s}, \boldsymbol{t}) = \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[\psi_{\boldsymbol{\omega}}(s)\overline{\psi}_{\boldsymbol{\omega}}(t)\right], \tag{$\mathcal{H}_{\psi_{\boldsymbol{\omega}},\Lambda}$}$$

and let $\mu^0 \in \mathcal{M}_{s_0,\Delta_0,\mathfrak{d}_{\mathfrak{g}}}$ where the Fisher-Rao distance is given by $K_{\text{pivot}}(\cdot,\cdot)$. Let $m \geq 1$ and let $(\boldsymbol{\omega}_i)_i$ be i.i.d. *vectors drawn with respect to Λ.*

We say that η^0 is an $(\varepsilon_0, \varepsilon_2, r_0)$-sketch pivot non-degenerate dual certificate if

- *There exists $c^0 \in \mathbb{C}^m$ such that $\eta^0 = \dfrac{1}{\sqrt{m}} \displaystyle\sum_{j=1}^m c_j^0 \varphi_{\boldsymbol{\omega}_j}$,*
- *For all $i \in [s_0]$, $\eta^0(t_i^0) = \text{sign}(a_i^0)$,*
- *For all $x \in \mathcal{F}^{\text{reg}}(r_0)$, $|\eta^0(x)| \leq 1 - \varepsilon_0$,*
- *For all $i \in [s_0]$, for all $x \in \mathcal{N}_i^{\text{reg}}(r_0)$, $|\eta^0(x)| \leq 1 - \varepsilon_2 \mathfrak{d}_{\mathfrak{g}}(x, t_i^0)^2$,*

where $\mathfrak{d}_{\mathfrak{g}}$ is the Fisher-Rao distance associated to $K_{\text{pivot}}(\cdot,\cdot)$, and the near and far regions are those of $(t_i^0)_{i=1}^{s_0}$.

For all $i \in [s_0]$, we say that η_i^0 is an $(\varepsilon_0, \varepsilon_2, r_0)$-pivot non-degenerate localizing certificate at point t_i^0 if

- *There exists $c_i^0 \in \mathbb{C}^m$ such that $\eta^0 = \dfrac{1}{\sqrt{m}} \displaystyle\sum_{j=1}^m c_{i,j}^0 \varphi_{\boldsymbol{\omega}_j}$,*
- *For all $x \in \mathcal{F}^{\text{reg}}(r_0)$, $|\eta_i^0(x)| \leq 1 - \varepsilon_0$,*
- *For all $j \in [s_0]$, for all $x \in \mathcal{N}_j^{\text{reg}}(r_0)$, $|1_{\{i=j\}} - \eta_i^0(x)| \leq \varepsilon_2 \mathfrak{d}_{\mathfrak{g}}(x, t_j^0)^2$,*

where $1_{\{i=j\}}$ equals one if $i = j$ and 0 otherwise.

Theorem 6 (Existence of sketch pivot non-degenerate dual certificates)**.** *Let $\mathcal{X}$ be a compact set of $\mathbb{R}^d$. Let $\varphi_{\boldsymbol{\omega}}, \psi_{\boldsymbol{\omega}} : \mathcal{X} \to \mathbb{C}$ be functions and let Λ be a law on $\mathbb{R}^d$. Consider*

$$\begin{aligned} K_{\text{mod}}(\boldsymbol{s}, \boldsymbol{t}) &= \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[\varphi_{\boldsymbol{\omega}}(s)\overline{\varphi}_{\boldsymbol{\omega}}(t)\right], \\ K_{\text{pivot}}(\boldsymbol{s}, \boldsymbol{t}) &= \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[\psi_{\boldsymbol{\omega}}(s)\overline{\psi}_{\boldsymbol{\omega}}(t)\right]. \end{aligned} \tag{$\mathcal{H}_{\psi_{\boldsymbol{\omega}},\Lambda}$}$$

Assume that $\psi_{\boldsymbol{\omega}}$ satisfies Assumption 3, and assume that $K_{\text{pivot}}(\cdot,\cdot)$ satisfies Assumption 1 (with parameters s_0, Δ_0, r_0, $\bar{\varepsilon}_0$ and $\bar{\varepsilon}_2$) and Assumption 2 with constant $C_{\text{switch}} > 0$.

Let $\alpha \in (0,1)$ and m be such that

$$m \geq C_{\text{sketch}}\, \max(d, \log(s_0))\, s_0\, \log\left(\frac{\max(1, |\mathcal{X}|)\, s_0}{\alpha}\right), \tag{A.3}$$

where $C_{\text{sketch}} > 0$ *depends on* K_{pivot} *and polynomially on* d, $|\mathcal{X}| = \sup_{\boldsymbol{s},\boldsymbol{t}\in\mathcal{X}} \mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t})$ *and the Fisher-Rao distance* $\mathfrak{d}_{\mathfrak{g}}$ *is given by* $K_{\text{pivot}}(\cdot,\cdot)$.

Then, there exists a constant $C'_{\text{pivot}} > 0$ *which depends only on the kernel* K_{pivot} *such that, with probability at least* $1-\alpha$, *for all* $m \geq m_0$, *there exist a* $(\bar{\varepsilon}_0/4, 3\bar{\varepsilon}_2/2, r_0)$*-sketch pivot non-degenerate dual certificate such that* $\|c^0\|_{\mathbb{C}^m} \leq C'_{\text{pivot}} C_{\text{switch}} \sqrt{s_0}$ *and* $(\bar{\varepsilon}_0/4, 3\bar{\varepsilon}_2/2, r_0)$*-sketch pivot non-degenerate localizing certificates at points* t_i^0 *such that* $\|c_i^0\|_{\mathbb{C}^m} \leq C'_{\text{pivot}} C_{\text{switch}}$.

Proof. This result is a consequence of Poon et al. 2023, Theorem 4, Page 274 and Theorem 5. The only subtlety relies in Poon et al. 2023, Section 6.5, Step 3, Page 284 where one has to invoke Proposition 3.1 and (3.5b) to get the bound $\|c^0\|_{\mathbb{C}^m} \leq C'_{\text{pivot}} C_{\text{switch}} \sqrt{s_0}$. Except this latter point, the reasoning is identical. □

A.3 Proof of Theorem 3

Consider the sinc-4 kernel Ψ_τ as defined by (1.9d). We want to prove that this kernel satisfies the local positive curvature assumption, Assumption 1. We split the proof into 5 parts.

A.3.1 Useful geometric object

For $\tau > 0$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we define:

- The random features $\psi_{\boldsymbol{\omega}}(\boldsymbol{x}) = \sqrt{\frac{f_\tau^{(4)}}{\Lambda}(\boldsymbol{\omega})} e^{-\imath \boldsymbol{\omega}^\top \boldsymbol{x}}$ with the spectral measure $f_\tau^{(4)} = (2\pi)^{-d}\, \mathfrak{F}[\Psi_\tau]$ detailed in Appendix C.5
- The kernel $K_\tau(\boldsymbol{x},\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda_\tau}\left[\overline{\psi_{\boldsymbol{\omega}}(\boldsymbol{x})}\psi_{\boldsymbol{\omega}}(\boldsymbol{y})\right] = \Psi_\tau(\boldsymbol{x}-\boldsymbol{y}) = \Psi_1(\frac{\boldsymbol{x}-\boldsymbol{y}}{\tau}) = K_1(\frac{\boldsymbol{x}}{\tau}, \frac{\boldsymbol{y}}{\tau})$.
- The metric $\mathfrak{g}_{\boldsymbol{x},\tau} = \nabla_1\nabla_2 K_\tau(\boldsymbol{x},\boldsymbol{x}) = -\nabla^2\Psi_\tau(\boldsymbol{0}) = \frac{1}{12\tau^2}\,\mathrm{Id}$. It is independent of the point $\boldsymbol{x}$ in the case of translation invariant kernel.
- The norm at point $\boldsymbol{x}$: $\|\boldsymbol{z}\|^2_{\boldsymbol{x},\tau} = \boldsymbol{z}^\top \mathfrak{g}_{\boldsymbol{x},\tau}\boldsymbol{z} = \frac{1}{12\tau^2}\|\boldsymbol{z}\|_2^2$
- The geodesic distance from $\boldsymbol{x}$ to $\boldsymbol{y}$ which is proportional to the euclidean norm since the kernel is translation invariant:

$$\mathfrak{d}_{\mathfrak{g},\tau}(\boldsymbol{x},\boldsymbol{y}) := \inf\left\{\int_0^1 \|\gamma'(t)\|_{\gamma(t)} \;:\; \gamma(0)=\boldsymbol{x},\ \gamma(1)=\boldsymbol{y}\right\} = \frac{1}{2\sqrt{3}\tau}\|\boldsymbol{x}-\boldsymbol{y}\|_2$$

Moreover, the metric have the following re-scaling properties

$$\mathfrak{g}_{\boldsymbol{x},\tau} = \frac{1}{\tau^2}\mathfrak{g}_{\boldsymbol{u},1},\ \forall \boldsymbol{x},\boldsymbol{u}\in\mathbb{R}^d, \tag{A.4}$$

$$\mathfrak{d}_{\mathfrak{g},\tau}(\boldsymbol{x},\boldsymbol{y}) = \mathfrak{d}_{\mathfrak{g},1}\left(\frac{\boldsymbol{x}}{\tau},\frac{\boldsymbol{y}}{\tau}\right). \tag{A.5}$$

Finally, notice that this choice of K_τ defines a flat euclidean geometry so that euclidean and Riemannian gradients (and Hessians) coincide. With a slight abuse of notations, we will use $\Psi_\tau(\boldsymbol{x},\boldsymbol{y})$ in place of $K_\tau(\boldsymbol{x},\boldsymbol{y}) = \Psi_\tau(\boldsymbol{x}-\boldsymbol{y})$.

A.3.2 Reduction to bandwidth one

• **Covariant derivatives:** First, we begin by proving that it suffices to prove the result for $\tau = 1$. The controls we look after, see Equations (3.3), involve the covariant derivatives of the kernel of order (i,j) seen as multi-linear maps $\Psi^{(ij)}(\boldsymbol{x},\boldsymbol{y}) : \left(\mathbb{C}^d\right)^{i+j} \to \mathbb{C}$:

$$\Psi^{(ij)}(\boldsymbol{x},\boldsymbol{y})[\boldsymbol{q}_1,\ldots,\boldsymbol{q}_{i+j}] := \mathbb{E}_{\boldsymbol{\omega}}\left[\overline{D_i[\psi_{\boldsymbol{\omega}}](\boldsymbol{x})[\boldsymbol{q}_1,\ldots,\boldsymbol{q}_i]} D_j[\psi_{\boldsymbol{\omega}}](\boldsymbol{y})[\boldsymbol{q}_{i+1},\ldots,\boldsymbol{q}_{i+j}]\right] \tag{A.6a}$$

In the particular case of $\psi_{\boldsymbol{\omega}}$ yielding the Ψ_τ kernel, we have simple expression of the partial derivatives of order r as $\partial_{i_1,\ldots,i_r}\psi_{\boldsymbol{\omega}}(\boldsymbol{x}) = (-\imath)^r \omega_{i_1}\ldots\omega_{i_r}\psi_{\boldsymbol{\omega}}(\boldsymbol{x})$ so that the differential of order r can be written as a rank-1 tensor in the canonical basis of $(\mathbb{C}^d)^r$

$$D_r[\psi_{\boldsymbol{\omega}}](\boldsymbol{x})[\cdot] = (-\imath)^r\psi_{\boldsymbol{\omega}}(\boldsymbol{x})(\boldsymbol{\omega}\otimes\ldots\otimes\boldsymbol{\omega})[\cdot] = (-\imath)^r\psi_{\boldsymbol{\omega}}(\boldsymbol{x})\,\boldsymbol{\omega}^{\otimes r}[\cdot]. \tag{A.6b}$$

This leads to the following simplification of Equation (A.6a) putting $Q = (\boldsymbol{q}_1, \dots, \boldsymbol{q}_{i+j})$

$$\Psi^{(ij)}(\boldsymbol{x}, \boldsymbol{y})[\boldsymbol{q}_1, \dots, \boldsymbol{q}_{i+j}] = (-1)^j \mathbb{E}_{\boldsymbol{\omega}} \left[\overline{D_{i+j}[\psi_{\boldsymbol{\omega}}](\boldsymbol{x})[Q]} \psi_{\boldsymbol{\omega}}(\boldsymbol{y}) \right] \tag{A.6c}$$

The operator norm at $\boldsymbol{x}, \boldsymbol{y}$ is then defined as Poon et al. (2023, Eq. (26))

$$\|\Psi^{(ij)}(\boldsymbol{x}, \boldsymbol{y})\|_{\boldsymbol{x},\boldsymbol{y}} \coloneqq \sup_{\substack{\|\boldsymbol{q}_l\|_{\boldsymbol{x}} \le 1,\, l=1,\dots,i \\ \|\boldsymbol{q}_{l+i}\|_{\boldsymbol{y}} \le 1,\, l=1,\dots,j}} \left| \mathbb{E}_{\boldsymbol{\omega}} \left[\overline{D_i[\psi_{\boldsymbol{\omega}}](\boldsymbol{x})[\boldsymbol{q}_1, \dots, \boldsymbol{q}_i]} D_j[\psi_{\boldsymbol{\omega}}](\boldsymbol{y})[\boldsymbol{q}_{i+1}, \dots, \boldsymbol{q}_{i+j}] \right] \right| . \tag{A.6d}$$

The first fact is that Ψ_τ and Ψ_1 (and their covariant derivatives) have the same operator norm provided we rescale the coordinate by $1/\tau$, as stated by the following lemma.

Lemma A.1 (Operator norm invariance by scaling $\boldsymbol{x}$ onto $\boldsymbol{x}/\tau$)**.** *Let $\tau > 0$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$, then, putting $\boldsymbol{u} = \boldsymbol{x}/\tau$ and $\boldsymbol{v} = \boldsymbol{y}/\tau$ for all $0 \le i, j \le 2$ we have*

$$\|K_\tau^{(ij)}(\boldsymbol{x}, \boldsymbol{y})\|_{\boldsymbol{x},\boldsymbol{y}} = \|K_1^{(ij)}(\boldsymbol{u}, \boldsymbol{v})\|_{\boldsymbol{u},\boldsymbol{v}}$$

Proof. We begin by noting that the norm $\|\boldsymbol{z}\|_{\boldsymbol{x}} = \|\boldsymbol{z}\|_{\boldsymbol{y}} \propto \frac{1}{\tau}\|\boldsymbol{z}\|_2$ is a rescaled euclidean metric, and is thus independent of the points $\boldsymbol{x}$ or $\boldsymbol{y}$. Thus, the operator norm defined in Equation (A.6d) can be written as

$$\begin{aligned} \|K_\tau^{(ij)}(\boldsymbol{x}, \boldsymbol{y})\|_{\boldsymbol{x},\boldsymbol{y}} &= \sup_{\substack{\|\boldsymbol{q}_l\|_\tau \le 1, \\ l=1,\dots,i+j}} \left| \mathbb{E}_{\boldsymbol{\omega}} \left[\overline{D_i[\psi_{\boldsymbol{\omega}}](\boldsymbol{x})[\boldsymbol{q}_1, \dots, \boldsymbol{q}_i]} D_j[\psi_{\boldsymbol{\omega}}](\boldsymbol{y})[\boldsymbol{q}_{i+1}, \dots, \boldsymbol{q}_{i+j}] \right] \right| , \\ &= \sup_{\substack{\|\boldsymbol{q}_l\|_\tau \le 1, \\ l=1,\dots,i+j}} \left| (-1)^j \mathbb{E}_{\boldsymbol{\omega}} \left[\overline{D_{i+j}[\psi_{\boldsymbol{\omega}}](\boldsymbol{x})[Q]} \psi_{\boldsymbol{\omega}}(\boldsymbol{y}) \right] \right| \end{aligned}$$

Using this symmetry, it suffice to prove the desired identity for $j = 0$ and $i = 0, \dots, 4$

- $i = j = 0$, $\|K_\tau^{(00)}(\boldsymbol{x}, \boldsymbol{y})\|_{\boldsymbol{x}} = K_\tau(\boldsymbol{x}, \boldsymbol{y}) = K_1(\frac{\boldsymbol{x}}{\tau}, \frac{\boldsymbol{y}}{\tau}) = K_1(\boldsymbol{u}, \boldsymbol{v})$.
- $i = 1, j = 0$: using that $\tau^2 \mathfrak{g}_{\boldsymbol{x},\tau} = \mathfrak{g}_{\boldsymbol{u},1}$ when $\boldsymbol{u} = \boldsymbol{x}/\tau$

$$\begin{aligned} \|K_\tau^{(10)}(\boldsymbol{x}, \boldsymbol{y})\|_{\boldsymbol{x}} &= \|\mathfrak{g}_{\boldsymbol{x},\tau}^{-1/2} \nabla_1 K_\tau(\boldsymbol{x}, \boldsymbol{y})\|_2 \\ &= \|\mathfrak{g}_{\boldsymbol{x},\tau}^{-1/2} \frac{1}{\tau} \nabla_1 K_1(\boldsymbol{u}, \boldsymbol{v})\|_2 \\ &= \|\mathfrak{g}_{\boldsymbol{u},1}^{-1/2} \nabla_1 K_1(\boldsymbol{u}, \boldsymbol{v})\|_2 \\ &= \|K_1^{(10)}(\boldsymbol{u}, \boldsymbol{v})\|_{\boldsymbol{u}} . \end{aligned}$$

- $i = 2, j = 0$: similarly

$$\begin{aligned} \|K_\tau^{(20)}(\boldsymbol{x}, \boldsymbol{y})\|_{\boldsymbol{x}} &= \|\mathfrak{g}_{\boldsymbol{x},\tau}^{-1/2} \nabla^2 \Psi_\tau(\boldsymbol{x} - \boldsymbol{y}) \mathfrak{g}_{\boldsymbol{x},\tau}^{-1/2}\|_2, \\ &= \|\mathfrak{g}_{\boldsymbol{x},\tau}^{-1/2} \frac{1}{\tau^2} \nabla^2 \Psi_1(\boldsymbol{u} - \boldsymbol{v}) \mathfrak{g}_{\boldsymbol{x},\tau}^{-1/2}\|_2, \\ &= \|\mathfrak{g}_{\boldsymbol{u},1}^{-1/2} \nabla^2 \Psi_1(\boldsymbol{u} - \boldsymbol{v}) \mathfrak{g}_{\boldsymbol{u},1}^{-1/2}\|_2, \\ &= \|K_1^{(20)}(\boldsymbol{u}, \boldsymbol{v})\|_{\boldsymbol{u}}. \end{aligned}$$

The same follows for higher order derivatives. □

• **Local curvature:** Poon et al. (2023) define geometric quantities related to the local curvature of the kernel K_τ (and hence of η^0) on near and far regions.

Definition 6 ($\bar{\varepsilon}_0$ and $\bar{\varepsilon_2}$)**.** *Let $r \ge 0$,*

$$\begin{aligned} \bar{\varepsilon}_0(r, \tau) &\coloneqq \sup \{ \varepsilon \,;\, K_\tau(\boldsymbol{x}, \boldsymbol{y}) \le 1 - \varepsilon,\ \forall \boldsymbol{x}, \boldsymbol{y} \text{ s.t. } \mathfrak{d}_{\mathfrak{g},\tau}(\boldsymbol{x}, \boldsymbol{y}) \ge r \} , \\ \bar{\varepsilon}_2(r, \tau) &\coloneqq \sup \left\{ \varepsilon \,;\, \Psi_\tau^{(02)}(\boldsymbol{x}, \boldsymbol{y})[\boldsymbol{z}, \boldsymbol{z}] \ge \varepsilon \|\boldsymbol{z}\|_{\boldsymbol{x},\tau}^2,\ \forall z \in \mathbb{R}^d,\ \forall \boldsymbol{x}, \boldsymbol{y} \text{ s.t. } \mathfrak{d}_{\mathfrak{g},\tau}(\boldsymbol{x}, \boldsymbol{y}) < r \right\} . \end{aligned}$$

These controls depend both on the kernel and the radius $r = r_{near}$ of the regions considered. However, the following Lemmas show that this controls do not depend on the bandwidth τ.

Lemma A.2 (Isometry between K_τ and K_1). *Let $\tau > 0$ and*

$$\mathcal{H}_\tau := \left\{ g \in L^2(\mathbb{R}^d) \,:\, \|g\|_{\mathcal{H}_\tau}^2 := \frac{1}{(2\pi)^d} \int \frac{|\mathfrak{F}[g]|^2}{\mathfrak{F}[\Psi_\tau]} < +\infty \right\},$$

then the rescaling map $iso_\tau : g \in \mathcal{H}_1 \to g(\frac{\cdot}{\tau}) \in \mathcal{H}_\tau$ is an isometry between $(\mathcal{H}_1, \|\cdot\|_{\tau=1})$ and $(\mathcal{H}_\tau, \|\cdot\|_\tau)$. Moreover, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we have:

(i) $K_1(\frac{\boldsymbol{x}}{\tau}, \frac{\boldsymbol{x}}{\tau}) = K_\tau(\boldsymbol{x}, \boldsymbol{y})$

(ii) $\nabla_1 K_1(\frac{\boldsymbol{x}}{\tau}, \frac{\boldsymbol{y}}{\tau}) = \tau \nabla_1 K_\tau(\boldsymbol{x}, \boldsymbol{y})$

(iii) $\nabla_{12} K_1(\frac{\boldsymbol{x}}{\tau}, \frac{\boldsymbol{y}}{\tau}) = \tau^2 \nabla_{12} K_\tau(\boldsymbol{x}, \boldsymbol{y})$

Proof. Let $\tau > 0$,

(i) By definition of K_τ and Ψ_τ.

(ii) $\nabla_1 K_1(\frac{\boldsymbol{x}}{\tau}, \frac{\boldsymbol{y}}{\tau}) = \nabla \Psi_1(\frac{\boldsymbol{x}-\boldsymbol{y}}{\tau}) = \tau \frac{1}{\tau} \nabla \Psi_1(\frac{\boldsymbol{x}-\boldsymbol{y}}{\tau}) = \tau \nabla \Psi_\tau(\boldsymbol{x}-\boldsymbol{y}) = \tau \nabla_1 K_\tau(\boldsymbol{x}, \boldsymbol{y})$

(iii) $\nabla_{12} K_1(\frac{\boldsymbol{x}}{\tau}, \frac{\boldsymbol{y}}{\tau}) = -\nabla^2 \Psi_1(\frac{\boldsymbol{x}-\boldsymbol{y}}{\tau}) = -\tau^2 \frac{1}{\tau^2} \nabla^2 \Psi_1(\frac{\boldsymbol{x}-\boldsymbol{y}}{\tau}) = -\tau^2 \nabla^2 \Psi_\tau(\boldsymbol{x}-\boldsymbol{y}) = +\tau^2 \nabla_{12} K_\tau(\boldsymbol{x}, \boldsymbol{y})$

For the isometry result, we simply use the change of variable $\boldsymbol{\omega}' = \boldsymbol{\omega}/\tau$ in the RKHS norm definition, with diagonal Jacobian yielding $\mathrm{d}\boldsymbol{\omega}' = \tau^{-d}\,\mathrm{d}\boldsymbol{\omega}$.

$$\begin{aligned}
\|g\|_{\mathcal{H}_1} &= \frac{1}{(2\pi)^d} \int \frac{|\mathfrak{F}[g](\boldsymbol{\omega})|^2}{\mathfrak{F}[\Psi_1](\boldsymbol{\omega})}\,\mathrm{d}\boldsymbol{\omega}, \\
&= \frac{1}{(2\pi)^d} \int \frac{|\mathfrak{F}[g](\tau\boldsymbol{\omega}')|^2}{\mathfrak{F}[\Psi_1](\tau\boldsymbol{\omega}')} \tau^d\,\mathrm{d}\boldsymbol{\omega}', && \text{(change of variables : } \boldsymbol{\omega} = \tau\boldsymbol{\omega}') \\
&= \frac{1}{(2\pi)^d} \int \frac{\tau^{-2d}\,|\mathfrak{F}[g(\cdot/\tau)](\boldsymbol{\omega}')|^2}{\mathfrak{F}[\Psi_1](\boldsymbol{\omega}')} \tau^d\,\mathrm{d}\boldsymbol{\omega}', && (\mathfrak{F}[g](\tau\boldsymbol{\omega}') = \tau^{-d}\,\mathfrak{F}[g(\cdot/\tau)](\boldsymbol{\omega}')), \\
&= \frac{1}{(2\pi)^d} \int \frac{|\mathfrak{F}[g(\cdot/\tau)](\boldsymbol{\omega}')|^2}{\tau^d\,\mathfrak{F}[\Psi_1](\tau\boldsymbol{\omega}')}\,\mathrm{d}\boldsymbol{\omega}', \\
&= \frac{1}{(2\pi)^d} \int \frac{|\mathfrak{F}[g(\cdot/\tau)](\boldsymbol{\omega}')|^2}{\mathfrak{F}[\Psi_\tau](\boldsymbol{\omega}')}\,\mathrm{d}\boldsymbol{\omega}', && \text{(same argument on } g = \Psi_1), \\
&= \|\mathrm{iso}_\tau(g)\|_{\mathcal{H}_\tau}\,,
\end{aligned}$$

which gives the result. □

Lemma A.3 ($\bar{\varepsilon}_0$ and $\bar{\varepsilon}_2$). *Let $\tau > 0$ and $r > 0$, then for*

$$j \in \{0, 2\}, \quad \bar{\varepsilon}_j(r, \tau) = \bar{\varepsilon}_j(r, 1)$$

Proof. Let $r > 0$. First, notice that due to the scaled euclidean metric induced by K_τ,

$$\mathfrak{d}_{\mathfrak{g},\tau}(\boldsymbol{x}, \boldsymbol{y}) \geq r \iff \mathfrak{d}_{\mathfrak{g},1}\left(\frac{\boldsymbol{x}}{\tau}, \frac{\boldsymbol{y}}{\tau}\right) \geq r$$

For $j = 0$, it suffices to use Lemma A.2 (i) together with the above equivalence to show that

$$\begin{aligned}
\bar{\varepsilon}_0(r, \tau) &= \sup\{\varepsilon \,:\, K_\tau(\boldsymbol{x}, \boldsymbol{y}) \leq 1 - \varepsilon,\ \forall \boldsymbol{x}, \boldsymbol{y} \text{ s.t. } \mathfrak{d}_{\mathfrak{g},\tau}(\boldsymbol{x}, \boldsymbol{y}) \geq r\}\,, \\
&= \sup\left\{\varepsilon \,:\, K_1\left(\frac{\boldsymbol{x}}{\tau}, \frac{\boldsymbol{y}}{\tau}\right) \leq 1 - \varepsilon,\ \forall \boldsymbol{x}, \boldsymbol{y} \text{ s.t. } \mathfrak{d}_{\mathfrak{g},1}\left(\frac{\boldsymbol{x}}{\tau}, \frac{\boldsymbol{y}}{\tau}\right) \geq r\right\}, \\
&= \sup\{\varepsilon \,:\, K_1(\boldsymbol{u}, \boldsymbol{v}) \leq 1 - \varepsilon,\ \forall \boldsymbol{u}, \boldsymbol{v} \text{ s.t. } \mathfrak{d}_{\mathfrak{g},1}(\boldsymbol{u}, \boldsymbol{v}) \geq r\}\,, \\
&= \bar{\varepsilon}_0(r, 1)\,.
\end{aligned}$$

For $j=2$, we begin by noting that $\psi_{\boldsymbol{\omega}}(\boldsymbol{x}) = \sqrt{f_\tau^{(4)}}(\boldsymbol{\omega})e^{-\imath\boldsymbol{\omega}^\top\boldsymbol{x}}$ so that, in the same lines as Lemma A.2 (iii):

$$\begin{aligned}
K_\tau^{(02)}(\boldsymbol{x},\boldsymbol{y})[\boldsymbol{z},\boldsymbol{z}] &= \mathbb{E}_{\boldsymbol{\omega}}\left[\overline{\psi_{\boldsymbol{\omega}}(\frac{\boldsymbol{x}}{\tau})}\boldsymbol{z}^\top\nabla^2\psi_{\boldsymbol{\omega}}(\frac{\boldsymbol{y}}{\tau})\boldsymbol{z}\right], \\
&= \mathbb{E}_{\boldsymbol{\omega}}\left[\overline{\psi_{\boldsymbol{\omega}}(\frac{\boldsymbol{x}}{\tau})}\psi_{\boldsymbol{\omega}}(\frac{\boldsymbol{y}}{\tau})\boldsymbol{z}^\top(-\imath)^2\boldsymbol{\omega}\boldsymbol{\omega}^\top\boldsymbol{z}\right], \\
&= -\boldsymbol{z}^\top\,\mathbb{E}_{\boldsymbol{\omega}}\left[\boldsymbol{\omega}\boldsymbol{\omega}^\top\overline{\psi_{\boldsymbol{\omega}}(\frac{\boldsymbol{x}}{\tau})}\psi_{\boldsymbol{\omega}}(\frac{\boldsymbol{y}}{\tau})\right]\boldsymbol{z}, \\
&= -\boldsymbol{z}^\top\nabla^2 K_\tau(\boldsymbol{x}-\boldsymbol{y})\boldsymbol{z}, \\
&= \boldsymbol{z}^\top\frac{1}{\tau^2}(-\tau^2)\nabla^2\Psi_\tau(\boldsymbol{x}-\boldsymbol{y})\boldsymbol{z}, \\
&= \frac{1}{\tau^2}\boldsymbol{z}^\top\left[-\nabla^2\Psi_1\left(\frac{\boldsymbol{x}-\boldsymbol{y}}{\tau}\right)\right]\boldsymbol{z}, \\
&= \frac{1}{\tau^2}\Psi_1^{(02)}\left(\frac{\boldsymbol{x}}{\tau},\frac{\boldsymbol{y}}{\tau}\right)[\boldsymbol{z},\boldsymbol{z}].
\end{aligned}$$

Moreover, for all $\boldsymbol{z}$ and $\boldsymbol{x}$, $\|\boldsymbol{z}\|^2_{\boldsymbol{x},\tau} = \frac{1}{\tau^2}\|\boldsymbol{z}\|^2_{\boldsymbol{x}/\tau,1}$. Thus, we again have equality of the ε-sets over which the supremum is taken.

$$\begin{aligned}
\bar{\varepsilon}_2(r,\tau) &= \sup\left\{\varepsilon\,;\, K_\tau^{(02)}(\boldsymbol{x},\boldsymbol{y})[\boldsymbol{z},\boldsymbol{z}] \geq \varepsilon\|\boldsymbol{z}\|^2_{\boldsymbol{x},\tau},\ \forall z\in\mathbb{R}^d,\ \forall \boldsymbol{x},\boldsymbol{y}\text{ s.t. }\mathfrak{d}_{\mathfrak{g},\tau}(\boldsymbol{x},\boldsymbol{y})<r\right\}, \\
&= \sup\left\{\varepsilon\,;\, \frac{1}{\tau^2}K_1^{(02)}\left(\frac{\boldsymbol{x}}{\tau},\frac{\boldsymbol{y}}{\tau}\right)[\boldsymbol{z},\boldsymbol{z}] \geq \varepsilon\frac{1}{\tau^2}\|\boldsymbol{z}\|^2_{\boldsymbol{x}/\tau,1},\ \forall z\in\mathbb{R}^d,\ \forall \boldsymbol{x},\boldsymbol{y}\text{ s.t. }\mathfrak{d}_{\mathfrak{g},1}\left(\frac{\boldsymbol{x}}{\tau},\frac{\boldsymbol{y}}{\tau}\right)<r\right\}, \\
&= \sup\left\{\varepsilon\,;\, K_1^{(02)}\left(\frac{\boldsymbol{x}}{\tau},\frac{\boldsymbol{y}}{\tau}\right)[\boldsymbol{z},\boldsymbol{z}] \geq \varepsilon\|\boldsymbol{z}\|^2_{\boldsymbol{x}/\tau,1},\ \forall z\in\mathbb{R}^d,\ \forall \boldsymbol{x},\boldsymbol{y}\text{ s.t. }\mathfrak{d}_{\mathfrak{g},1}\left(\frac{\boldsymbol{x}}{\tau},\frac{\boldsymbol{y}}{\tau}\right)<r\right\}, \\
&= \sup\left\{\varepsilon\,;\, K_1^{(02)}(\boldsymbol{u},\boldsymbol{v})[\boldsymbol{z},\boldsymbol{z}] \geq \varepsilon\|\boldsymbol{z}\|^2_{\boldsymbol{u},1},\ \forall z\in\mathbb{R}^d,\ \forall \boldsymbol{u},\boldsymbol{v}\text{ s.t. }\mathfrak{d}_{\mathfrak{g},1}(\boldsymbol{u},\boldsymbol{v})<r\right\}, \\
&= \bar{\varepsilon}_2(r,1).
\end{aligned}$$

□

• **Separation:** Then, for a number of components $s\in\mathbb{N}$ and $h>0$, the kernel width of K_τ is defined by Poon et al. (2023) as

$$\Delta(h,s) \coloneqq \inf\left\{\Delta\,;\, \sum_{k=2}^{s}\left\|K_\tau^{(ij)}(\boldsymbol{x}_1^0,\boldsymbol{x}_k^0)\right\|_{\boldsymbol{x}_1^0,\boldsymbol{x}_k^0} \leq h,\quad (i,j)\in\{0,1\}\times\{0,2\},\quad \{\boldsymbol{x}_k^0\}_{k=1}^s\in\mathcal{S}_{\Delta,\tau}\right\}, \tag{A.7}$$

where $\mathcal{S}_{\Delta,\tau} \coloneqq \{(\boldsymbol{x}_k)_{k=1}^s\in\mathcal{X}\,;\,\mathfrak{d}_\tau(\boldsymbol{x}_k,\boldsymbol{x}_l)\geq\Delta,\forall k\neq l\}$ is the set of $\boldsymbol{s}$-tuples of Δ-separated points. This notion is somewhat related to the coherence of the kernel K meaning that the interference is small enough so that functions $K_\tau(\boldsymbol{x}_k^0,\cdot)$ can interpolate between parameters.

Remark A.1. *The particular form of the metric $\mathfrak{d}_\tau$ gives $S_{\Delta,\tau} = \tau\cdot S_{\Delta,1}$. Together with Lemma A.1, this implies that $\Delta(h,s)$ does not depend on τ.*

A.3.3 First part of LPC

This lead to tracking uniform bound on the Kernel and its derivative

$$B_{ij} \coloneqq \sup_{\boldsymbol{x},\boldsymbol{y}}\|\Psi_\tau^{(ij)}(\boldsymbol{x}-\boldsymbol{y})\|_{\boldsymbol{x},\boldsymbol{y}} \qquad\qquad B_i \coloneqq 1+B_{0i}+B_{1i}, \tag{A.8}$$

which regimes w.r.t. to the dimension d are given by the following lemma.

Proposition A.1. *One has*

$$\forall i,j,\quad B_{ij} \leq (12d)^{\frac{i+j}{2}},$$

and hence $1/\sqrt{B_{02}} \geq 1/\sqrt{12d}$ and $B_2 \leq 1+\sqrt{12d}+12d$.

Proof. Again, the rate does not depend on τ as Lemma A.1 gives

$$\sup_{\boldsymbol{x},\boldsymbol{y}} \|K_\tau^{(ij)}(\boldsymbol{x},\boldsymbol{y})\|_{\boldsymbol{x},\boldsymbol{y}} = \sup_{\boldsymbol{u},\boldsymbol{v}} \|K_1^{(ij)}(\boldsymbol{u},\boldsymbol{v})\|_{\boldsymbol{u},\boldsymbol{v}}\,.$$

For the rest of the proof, τ is fixed equal to 1.

Trivially, for $i=j=0$, a triangle inequality yield a uniform control on $(\boldsymbol{u},\boldsymbol{v})$ as:

$$\sup_{\boldsymbol{u},\boldsymbol{v}} |K_1(\boldsymbol{u},\boldsymbol{v})| \le \sup_{\boldsymbol{u},\boldsymbol{v}} \int \left|e^{\imath\boldsymbol{\omega}^\top(\boldsymbol{u}-\boldsymbol{v})}\right| f_\tau^{(4)}(\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{\omega} = \int f_\tau^{(4)}(\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{\omega} = 1\,. \tag{A.9}$$

For higher order derivative, the proof relies on Appendix B.1 and the fact that the covariant derivatives of the kernel involves differential $\nabla^r\psi_{\boldsymbol{\omega}}(\boldsymbol{u}) : (\mathbb{C}^d)^r \to \mathbb{C}$ for which we can control the supremum uniformly over all $\boldsymbol{u}\in\mathcal{X}$:

$$\sup_{\|\boldsymbol{q}_l\|_2^2\le 12,\, l=1,\dots,r} |\nabla^r\psi_{\boldsymbol{\omega}}(\boldsymbol{u})[\boldsymbol{q}_1,\dots,\boldsymbol{q}_r]| = \sqrt{\frac{f_\tau^{(4)}}{\Lambda}(\boldsymbol{\omega})} \times \sup_{\|\boldsymbol{q}_l\|_2^2\le 12,\, l=1,\dots,r} \prod_{l=1}^{r} |\langle\boldsymbol{\omega},\boldsymbol{q}_l\rangle|\,. \tag{A.10}$$

Hence, the covariant derivative of the kernel can be uniformly upper bounded for all $\boldsymbol{u},\boldsymbol{v}\in\mathcal{X}$ by

$$\begin{aligned}
\left|K_1^{(ij)}(\boldsymbol{u},\boldsymbol{v})[\boldsymbol{q}_1,\dots,\boldsymbol{q}_{i+j}]\right| &\le \int_{[-1,1]^d} \left(\prod_{l=1}^{i+j} \sup_{\|\boldsymbol{q}_l\|_2^2\le 12} |<\boldsymbol{\omega},\boldsymbol{q}_l>|\right) f_\tau^{(4)}(\boldsymbol{\omega}) |e^{-\imath\boldsymbol{\omega}^T(\boldsymbol{u}-\boldsymbol{v})}|\,\mathrm{d}\boldsymbol{\omega},\\
&\le \sup_{\boldsymbol{\omega}\in[-1,1]^d} \prod_{l=1}^{i+j} \sup_{\|\boldsymbol{q}_l\|_2^2\le 12} |<\boldsymbol{\omega},\boldsymbol{q}_l>|, \qquad (f_\tau^{(4)} \text{ p.d.f on the unit hypercube})\\
&\le \sup_{\boldsymbol{\omega}\in[-1,1]^d} (2\sqrt{3}\|\boldsymbol{\omega}\|_2)^{i+j}, \qquad \text{(Cauchy-Schwarz)}\\
&\le \sqrt{12d}^{\,i+j}\,.
\end{aligned}$$

Taking the sup of the latter on $\boldsymbol{u},\boldsymbol{v}$ terminates the proof. □

A.3.4 Second part of LPC

Then, we focus on the triplets r_0, $\bar\varepsilon_0$ and $\bar\varepsilon_2$ defined in Equation (3.3b) which, in turn, define the minimal separation Δ_0 required between the parameters. Specifically, we need to exhibit some radius $r_0\in(0,1/\sqrt{B_{02}})$ for the near regions such that the kernel K_τ has positive curvature constants $\bar\varepsilon_0$ and $\bar\varepsilon_2$.

$$\bar\varepsilon_0 := \frac{1}{2}\sup_{\varepsilon\ge 0}\left\{\varepsilon \,:\, K_1(\boldsymbol{s},\boldsymbol{t})\le 1-\varepsilon\,,\ \forall \boldsymbol{s},\boldsymbol{t}\in\mathcal{X} \text{ s.t. } \mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t})\ge r_0\right\}, \tag{A.11a}$$

$$\bar\varepsilon_2 := \frac{1}{4}\sup_{\varepsilon\ge 0}\left\{\varepsilon \,:\, -K_1^{(0,2)}(\boldsymbol{s},\boldsymbol{t})[\boldsymbol{v},\boldsymbol{v}]\ge \varepsilon\|\boldsymbol{v}\|_{\boldsymbol{t}}^2\,,\ \forall \boldsymbol{v}\in\mathbb{T}_{\boldsymbol{t}}\,,\ \forall \boldsymbol{s},\boldsymbol{t}\in\mathcal{X} \text{ s.t. } \mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t})< r_0\right\}, \tag{A.11b}$$

$$\Delta_0 := \inf\left\{\Delta \,:\, 32\sum_{l=2}^{s_0}\|K_1^{(i,j)}(\boldsymbol{x}_1,\boldsymbol{x}_l)\|_{\boldsymbol{x}_1,\boldsymbol{x}_l} \le \min\left(\frac{\bar\varepsilon_0}{B_0},\frac{2\bar\varepsilon_2}{B_2}\right)\,,\ (i,j)\in\{0,1\}\times\{0,2\}\,,\ \{\boldsymbol{x}_l\}_{l=1}^{s_0}\in\mathcal{S}_\Delta\right\}. \tag{A.11c}$$

This is the purpose of the following Lemma, which also exhibits the scaling of the last quantities with respect to the dimension d.

Proposition A.2. *Let r_0 be such that $r_0 = 1/(4\,d)$ then $\bar\varepsilon_0\ge 1/(32\,d^3)$ and $\bar\varepsilon_2\ge 23/128$ in Equations A.11.*

Proof. One can check the following calculus:

$$\forall z\in\mathbb{R}\,,\quad |\operatorname{sinc} z|\le (1-\frac{z^2}{12})\mathbb{1}_{|z|\le 2} + \frac{1}{2}\mathbb{1}_{|z|>2}\,. \tag{A.12a}$$

Now, let $c>0$ and $\boldsymbol{t}\in\mathbb{R}^d$ such that $(1/\sqrt{12})\|\boldsymbol{t}\|_2\ge c/d$, then there exists a coordinate z of the vector $\boldsymbol{t}$ such that $|z|\ge c\sqrt{12}d^{-3/2}$ and we deduce by (A.12a) that $\bar\varepsilon_0\ge c^2d^{-3}/2$. The result is given for $c=1/4$.

We turn to $\bar{\varepsilon}_2$. Let $\boldsymbol{s},\boldsymbol{t},\boldsymbol{v}\in\mathbb{R}^d$ be fixed. Writing the exact Taylor expansion of $\boldsymbol{u}\mapsto -K_1^{(0,2)}(\boldsymbol{s},\boldsymbol{u})[\boldsymbol{v},\boldsymbol{v}]$ at $\boldsymbol{s}$ and using (B.1e), there exists $\boldsymbol{\xi}\in[\boldsymbol{s},\boldsymbol{t}]$ (which is the geodesic between these points) such that

$$\begin{aligned}-K_1^{(0,2)}(\boldsymbol{s},\boldsymbol{t})[\boldsymbol{v},\boldsymbol{v}] &= -K_1^{(0,2)}(\boldsymbol{s},\boldsymbol{s})[\boldsymbol{v},\boldsymbol{v}] - K_1^{(0,3)}(\boldsymbol{s},\boldsymbol{s})[\boldsymbol{v},\boldsymbol{v},\boldsymbol{s}-\boldsymbol{t}] - \frac{1}{2}K_1^{(0,4)}(\boldsymbol{s},\boldsymbol{\xi})[\boldsymbol{v},\boldsymbol{v},\boldsymbol{s}-\boldsymbol{t},\boldsymbol{s}-\boldsymbol{t}]\\ &= \|v\|_{\boldsymbol{t}}^2 + 0 + \frac{1}{2}\boldsymbol{v}^\top\big(A(\boldsymbol{s},\boldsymbol{\xi})[\boldsymbol{s}-\boldsymbol{t},\boldsymbol{s}-\boldsymbol{t}]\big)\boldsymbol{v}\\ &\geq \big(1-6\rho(A(\boldsymbol{s},\boldsymbol{\xi})[\boldsymbol{s}-\boldsymbol{t},\boldsymbol{s}-\boldsymbol{t}])\big)\,\|\boldsymbol{v}\|_{\boldsymbol{t}}^2\end{aligned}$$

where the quadratic form $K_1^{(0,4)}(\boldsymbol{s},\boldsymbol{\xi})[\boldsymbol{s}-\boldsymbol{t},\boldsymbol{s}-\boldsymbol{t}]$ is identified to the $d\times d$ matrix in the canonical basis:

$$\forall i_1,i_2\in[d]\,,\quad (A(\boldsymbol{s},\boldsymbol{\xi})[\boldsymbol{s}-\boldsymbol{t},\boldsymbol{s}-\boldsymbol{t}])_{i_1,i_2} = \sum_{i_3=1}^{d}\sum_{i_4=1}^{d}(s-t)_{i_3}(s-t)_{i_4}\partial_{i_1,i_2,i_3,i_4}\Psi_1(\boldsymbol{s}-\boldsymbol{\xi})\,, \tag{A.12b}$$

and $\rho(A)$ denotes its spectral radius.

Then, notice that $f=\mathrm{sinc}$ has all derivatives upper bounded by $1/2$ in absolute value so that

$$\begin{aligned}\frac{d^4}{dx^4}f^4(x) &= 4f^3(x)\frac{d^4}{dx^4}f(x) + 48f^2(x)\frac{d}{dx}f(x)\frac{d^3}{dx^3}f(x)\\ &\quad + 36f^2(x)\left(\frac{d^2}{dx^2}f(x)\right)^2 + 144f(x)\left(\frac{d}{dx}f(x)\right)^2\frac{d^2}{dx^2}f(x) + 24\left(\frac{d}{dx}f(x)\right)^4\\ &\leq \frac{256}{16} = 16.\end{aligned}$$

Remembering that $\Psi_1=\mathrm{sinc}^4(\cdot/4)$, we deduce from the previous calculus that

$$\forall i_1,i_2,i_3,i_4\in[d]\,,\quad |\partial_{i_1,i_2,i_3,i_4}\Psi_1(\boldsymbol{s}-\boldsymbol{\xi})| \leq 16\times\frac{1}{4^4} = \frac{1}{16}\,, \tag{A.12c}$$

$$\forall i_1,i_2\in[d]\,,\quad |(A(\boldsymbol{s},\boldsymbol{\xi})[\boldsymbol{s}-\boldsymbol{t},\boldsymbol{s}-\boldsymbol{t}])_{i_1,i_2}| \leq \frac{d\|\boldsymbol{s}-\boldsymbol{t}\|_2^2}{16}\,. \tag{A.12d}$$

Finally, using that $\rho(A)\leq d\sup_{i_1,i_2}|A_{i_1,i_2}|$, one has

$$6\rho(A(\boldsymbol{\xi},\boldsymbol{t})) \leq \frac{6d^2\|\boldsymbol{s}-\boldsymbol{t}\|_2^2}{16} = \frac{9d^2}{2}\,\mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t})^2\,,$$

Hence for $r_0=1/(4d)$ one has $-K_1^{(0,2)}(\boldsymbol{s},\boldsymbol{t})[\boldsymbol{v},\boldsymbol{v}]\geq(1-9/32)\|\boldsymbol{v}\|_{\boldsymbol{t}}^2$. □

Remark A.2. *Note that $r_0<B_{02}^{-1/2}$ since $r_0=1/(4\,d)$ and $B_{02}^{-1/2}\geq 1/\sqrt{12d}$.*

A.3.5 Last part of LPC

Note that $B_0\leq 1+\sqrt{12d}+1=2+\sqrt{12d}$ and $\bar{\varepsilon}_0/B_0\geq 1/(32\,d^3(2+\sqrt{12d}))\geq 1/(175\,d^{7/2})$. Also, it holds $B_2\leq 12d+(12d)^{3/2}+1$ and $2\bar{\varepsilon}_2/B_2\geq 23/(64+768d+64\,(12d)^{3/2})\geq 23/(1088\,d^{3/2})$. Hence

$$\frac{1}{32}\min\left(\frac{\bar{\varepsilon}_0}{B_0},\frac{2\bar{\varepsilon}_2}{B_2}\right)\geq\frac{1}{5600\,d^{7/2}}\,.$$

Using Lemma A.4 below, we deduce that

$$\sum_{k=2}^{s}\|K_1^{(ij)}(\boldsymbol{u}_k,\boldsymbol{u}_1)\|_{\boldsymbol{u}_k,\boldsymbol{u}_1}\leq\left(\frac{4}{3}\right)^2\sqrt{48d}^{\,i+j}\,\frac{sd^2}{\Delta^4}\leq\frac{1}{5600\,d^{7/2}}\leq\frac{1}{32}\min\left(\frac{\bar{\varepsilon}_0}{B_0},\frac{2\bar{\varepsilon}_2}{B_2}\right)$$

as soon as $\Delta\geq\Delta_0:=42.66\,s^{1/4}d^{7/4}$.

Lemma A.4. *It holds that*

$$\|K_1^{(ij)}(\boldsymbol{s},\boldsymbol{t})\|_{\boldsymbol{s},\boldsymbol{t}}\leq\left(\frac{4}{3}\right)^2\sqrt{48d}^{\,i+j}\,\frac{d^2}{\mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t})^4}\,.$$

Proof. We bound each entry of $\Psi_1^{(i+j)}(\boldsymbol{s}-\boldsymbol{t})$ using that $|\operatorname{sinc}^{(i)}(x/4)| \leq (1+i)\,4/|x|$ for all $x \neq 0$. One has

$$\Psi_1(\boldsymbol{s}-\boldsymbol{t}) = \prod_{\ell=1}^{d} \operatorname{sinc}^4(\frac{s_\ell - t_\ell}{4}) \leq \frac{4^4}{(s_{\ell_0}-t_{\ell_0})^4} \prod_{\ell\neq\ell_0}^{d} \operatorname{sinc}^4(\frac{s_\ell - t_\ell}{4}) \leq \frac{4^4 d^2}{\|\boldsymbol{s}-\boldsymbol{t}\|_2^4} = \Big(\frac{4}{3}\Big)^2 \frac{d^2}{\mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t})^4}\,,$$

because one coordinate ℓ_0 exists such that $(s_{\ell_0}-t_{\ell_0})^2 \geq \|\boldsymbol{s}-\boldsymbol{t}\|_2^2/d$. Let $h := \boldsymbol{s}-\boldsymbol{t}$ and ∂_u the derivative with respect to coordinate u, one has

$$\partial_u \Psi_1(h) = \operatorname{sinc}^{(1)}\big(\frac{h_u}{4}\big) \operatorname{sinc}^3\big(\frac{h_u}{4}\big) \prod_{\ell\neq u} \operatorname{sinc}^4\big(\frac{h_\ell}{4}\big) \leq \frac{2\,4^4}{(h_{\ell_0})^4} \leq 2\Big(\frac{4}{3}\Big)^2 \frac{d^2}{\mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t})^4}\,.$$

because one coordinate ℓ_0 exists such that $h_{\ell_0}^2 \geq \|h\|_2^2/d$. For the second order derivatives, if $u=v$ one has

$$\begin{aligned}\partial_{u,v}^2 \Psi_1(h) &= \big(\frac{1}{4}\big) \operatorname{sinc}^{(2)}\big(\frac{h_u}{4}\big) \operatorname{sinc}^3\big(\frac{h_u}{4}\big) \prod_{\ell\neq u} \operatorname{sinc}^4\big(\frac{h_\ell}{4}\big) + \big(\frac{3}{4}\big)\big(\operatorname{sinc}^{(1)}\big(\frac{h_u}{4}\big)\big)^2 \operatorname{sinc}^2\big(\frac{h_u}{4}\big) \prod_{\ell\neq u} \operatorname{sinc}^4\big(\frac{h_\ell}{4}\big) \\ &\leq 4\Big(\frac{4}{3}\Big)^2 \frac{d^2}{\mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t})^4}\,,\end{aligned}$$

and otherwise if $u \neq v$

$$\partial_{u,v}^2 \Psi_1(h) = \operatorname{sinc}^{(1)}\big(\frac{h_u}{4}\big) \operatorname{sinc}^3\big(\frac{h_u}{4}\big) \operatorname{sinc}^{(1)}\big(\frac{h_v}{4}\big) \operatorname{sinc}^3\big(\frac{h_v}{4}\big) \prod_{\ell\neq u,v} \operatorname{sinc}^4\big(\frac{h_\ell}{4}\big) \leq 4\Big(\frac{4}{3}\Big)^2 \frac{d^2}{\mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t})^4}\,.$$

Pursuing the reasoning, one gets

$$|\partial^{i+j}\Psi_1(h)| \leq 2^{i+j}\Big(\frac{4}{3}\Big)^2 \frac{d^2}{\mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t})^4}\,.$$

Now, observe that

$$\|K_1^{(ij)}(\boldsymbol{s},\boldsymbol{t})\|_{\boldsymbol{s},\boldsymbol{t}} = \sqrt{12}^{i+j}\|K_1^{(ij)}(\boldsymbol{s},\boldsymbol{t})\|_2 \leq (12d)^{\frac{i+j}{2}} \max|\partial^{i+j}\Psi_1| \leq (48d)^{\frac{i+j}{2}}\Big(\frac{4}{3}\Big)^2 \frac{d^2}{\mathfrak{d}_{\mathfrak{g}}(\boldsymbol{s},\boldsymbol{t})^4}\,,$$

which gives the result. □

A.4 Proof of Theorem 4

We refer the reader to the definitions of N, C_1, C_2, C_3 and L_{ij} in Assumption 3. Since $\bar{\varepsilon}_0$ and $\bar{\varepsilon}_2$ are already treated in Theorem 3, we only need to find the scaling of L_1, L_2, and L_3. Concretely, for $r \in \{0,\dots,3\}$, we need to study the tail of the distribution of the random variable

$$Y_r(\boldsymbol{\omega}) = \sup_{\boldsymbol{x}\in\mathcal{X}} \|D_r[\psi_{\boldsymbol{\omega}}](\boldsymbol{x})\|_{\boldsymbol{x}}, \tag{A.13}$$

where $\psi_{\boldsymbol{\omega}}(\boldsymbol{x}) = e^{\imath\boldsymbol{\omega}^\top\boldsymbol{x}}\sqrt{\frac{f_\tau^{(4)}}{\Lambda}(\boldsymbol{\omega})}$ and the convention $\|D_0[\psi_{\boldsymbol{\omega}}(\boldsymbol{x})\|_{\boldsymbol{x}} = |\psi_{\boldsymbol{\omega}}(\boldsymbol{x})|$. Specifically, we want to exhibits $L_r \in \mathbb{R}$ such that

$$\sum_{r=0}^{3} 1 - \mathbb{P}_{\boldsymbol{\omega}\sim\Lambda}(Y_r(\boldsymbol{\omega}) \leq L_r) \leq \frac{\min(\bar{\varepsilon}_0,\bar{\varepsilon}_2,\alpha)}{m}.$$

In the sinc-4 case, this is given by the following Lemma, independently of α, $r_0, \bar{\varepsilon}_0, \bar{\varepsilon}_2$ and m

Lemma A.5. *For any $\tau > 0$, and any $r \in \{0,\dots,3\}$, the covariant derivatives are almost surely bounded and we have*

$$L_r = (\sqrt{12d})^r \sup_{\boldsymbol{\omega}} \sqrt{\frac{f_\tau^{(4)}}{\Lambda}(\boldsymbol{\omega})}. \tag{A.14}$$

Proof. Since the support of Λ is bounded and $\psi_{\boldsymbol{\omega}}(\cdot)$ is a weighted Fourier feature, we don't need refined probabilistic control on the tail bound of $L_r(\boldsymbol{\omega})$. Indeed, using Appendix B.1 we have

$$Y_0(\boldsymbol{\omega}) = \sup_{\boldsymbol{x}} |\psi_{\boldsymbol{\omega}}(\boldsymbol{x})| = \sqrt{\frac{f_\tau^{(4)}}{\Lambda}}(\boldsymbol{\omega}), \tag{A.15}$$

$$Y_1(\boldsymbol{\omega}) = \sup_{\boldsymbol{x}} \|\mathfrak{g}_{\boldsymbol{x},\tau}^{-1/2}\nabla\psi_{\boldsymbol{\omega}}(\boldsymbol{x})\|_2 = \sqrt{\frac{f_\tau^{(4)}}{\Lambda}}(\boldsymbol{\omega})\sqrt{12}\tau\|\boldsymbol{\omega}\|_2, \tag{A.16}$$

$$Y_2(\boldsymbol{\omega}) = \sup_{\boldsymbol{x}} \|\mathfrak{g}_{\boldsymbol{x},\tau}^{-1/2}\nabla^2\psi_{\boldsymbol{\omega}}(\boldsymbol{x})\mathfrak{g}_{\boldsymbol{x},\tau}^{-1/2}\|_{op} = \sqrt{\frac{f_\tau^{(4)}}{\Lambda}}(\boldsymbol{\omega})(\sqrt{12}\tau)^2\|\boldsymbol{\omega}\|_2^2, \tag{A.17}$$

$$Y_3(\boldsymbol{\omega}) = \sup_{\boldsymbol{x}} \|\nabla^3\psi_{\boldsymbol{\omega}}(\boldsymbol{x})[\mathfrak{g}_{\boldsymbol{x},\tau}^{-1/2}\cdot, \mathfrak{g}_{\boldsymbol{x},\tau}^{-1/2}\cdot, \mathfrak{g}_{\boldsymbol{x},\tau}^{-1/2}\cdot]\|_{op} = \sqrt{\frac{f_\tau^{(4)}}{\Lambda}}(\boldsymbol{\omega})(\sqrt{12}\tau)^3\|\boldsymbol{\omega}\|_2^3. \tag{A.18}$$

Thus, since Λ has a bounded support on $\Omega = [-1/\tau, 1/\tau]^d$, and using $\|\boldsymbol{\omega}\|_2 \leq \sqrt{d}\|\boldsymbol{\omega}\|_\infty$, the random variables $Y_r(\boldsymbol{\omega})$ are a.s. bounded by

$$Y_r(\boldsymbol{\omega}) \leq \sup_{\boldsymbol{\omega}\in\Omega} \sqrt{\frac{f_\tau^{(4)}}{\Lambda}}(\boldsymbol{\omega})(\sqrt{12}\tau)^r\left(\frac{\sqrt{d}}{\tau}\right)^r = \sup_{\boldsymbol{\omega}\in\Omega} \sqrt{\frac{f_\tau^{(4)}}{\Lambda}}(\boldsymbol{\omega})(\sqrt{12}\sqrt{d})^r.$$

Denoting $C_\Lambda = \sup_{\boldsymbol{\omega}\in\Omega} \frac{f_\tau^{(4)}}{\Lambda}(\boldsymbol{\omega})$, we get that

$$\mathbb{P}_{\boldsymbol{\omega}\sim\Lambda}\Big(Y_r(\boldsymbol{\omega}) \leq \sqrt{C_\Lambda}(\sqrt{12}\sqrt{d})^r\Big) = 1,$$

so that $\psi_{\boldsymbol{\omega}}$ satisfies Assumption 3 with $L_r := \sqrt{C_\Lambda}(\sqrt{12}\sqrt{d})^r$. The proof ends by plugging these values in N, C_1, C_2 and L_{ij}. □

A.5 Proof of Theorem 1

By Theorem 5, we known that there exists pivot non-degenerate certificates. Let η^0 be an $(\bar{\varepsilon}_0, \bar{\varepsilon}_2, r_0)$-pivot non-degenerate dual certificate. Note that

$$\frac{1}{2}\big\|\boldsymbol{y} - F\mu\big\|_{\mathcal{F}}^2 + \kappa\|\mu\|_{\mathrm{TV}} \leq \frac{1}{2}\big\|\boldsymbol{y} - F\mu^0\big\|_{\mathcal{F}}^2 + \kappa\|\mu^0\|_{\mathrm{TV}} = \frac{\gamma^2}{2} + \kappa\|\mu\|_{\mathrm{TV}} \tag{A.19}$$

Using $\eta^0 = F^\star c^0$ and (B.3), we obtain

$$\begin{aligned}
\kappa\mathcal{D}_{\eta^0}(\mu\,||\,\mu^0) + \kappa\langle\eta^0, \mu - \mu^0\rangle_{\mathcal{C}(\mathcal{X})\times\mathcal{M}(\mathcal{X})} + \frac{1}{2}\big\|\boldsymbol{y} - F\mu\big\|_{\mathcal{F}}^2 &\leq \frac{\gamma^2}{2} \\
\kappa\mathcal{D}_{\eta^0}(\mu\,||\,\mu^0) + \frac{1}{2}\big\|\kappa c^0 + F\mu - \boldsymbol{y}\big\|_{\mathcal{F}}^2 &\leq \frac{\gamma^2}{2} + \frac{\kappa^2\|c^0\|_{\mathcal{F}}^2}{2} - \langle\kappa c^0, \Gamma\rangle_{\mathcal{F}} \\
\mathcal{D}_{\eta^0}(\mu\,||\,\mu^0) &\leq \frac{(\gamma + \kappa\|c^0\|_{\mathcal{F}})^2}{2\kappa} \\
&\leq \frac{(\gamma + C_{\mathsf{switch}}\kappa\sqrt{2s_0})^2}{2\kappa}
\end{aligned} \tag{A.20}$$

where we recall that

$$\mathcal{D}_{\eta^0}(\mu\,||\,\mu^0) := \|\mu\|_{\mathrm{TV}} - \|\mu^0\|_{\mathrm{TV}} - \langle\eta^0, \mu - \mu^0\rangle_{\mathcal{C}(\mathcal{X})\times\mathcal{M}(\mathcal{X})}\,,$$

see (1.8), and $\Gamma := \boldsymbol{y} - F\mu^0$. Choosing $\kappa = c_\kappa\gamma/\sqrt{s_0}$ one gets

$$\mathcal{D}_{\eta^0}(\mu\,||\,\mu^0) \leq \bar{c}_\kappa\gamma\sqrt{s_0} \quad \text{where} \quad \bar{c}_\kappa := \frac{(1 + \sqrt{2}\,C_{\mathsf{switch}}c_\kappa)^2}{2c_\kappa} \geq 2\sqrt{2}\,C_{\mathsf{switch}}\,. \tag{A.21}$$

On the other hand, using that η^0 is an $(\bar{\varepsilon}_0, \bar{\varepsilon}_2, r_0)$-pivot non-degenerate dual certificate, for $r \in (0, r_0]$,

$$\begin{aligned}
\mathcal{D}_{\eta^0}(\mu\,||\,\mu^0) &= \|\mu\|_{\mathrm{TV}} - \langle\eta^0, \mu\rangle_{\mathcal{C}(\mathcal{X})\times\mathcal{M}(\mathcal{X})} \\
&\geq \|\mu\|_{\mathrm{TV}} - \sum_{l=1}^{s_0}\int_{\mathcal{N}_l^{\mathsf{reg}}(r)} |\eta^0|\,\mathrm{d}|\mu| - \int_{\mathcal{F}_l^{\mathsf{reg}}(r)} |\eta^0|\,\mathrm{d}|\mu|
\end{aligned}$$

Now observe that, for $r \in (0, r_0]$,

- For all $i \in [s_0]$, $\eta^0(t_i^0) = \mathrm{sign}(a_i^0)$,
- For all $x \in \mathcal{F}^{\mathrm{reg}}(r)$,
$$|\eta^0(x)| \leq 1 - \min(\bar{\varepsilon}_0, \bar{\varepsilon}_2 r^2) = 1 - \bar{\varepsilon}_2 r^2\,, \tag{A.22}$$
- For all $i \in [s_0]$, for all $x \in \mathcal{N}_i^{\mathrm{reg}}(r)$, $|\eta^0(x)| \leq 1 - \bar{\varepsilon}_2 \mathfrak{d}_{\mathfrak{g}}(x, t_i^0)^2$.

Hence,
$$\mathcal{D}_{\eta^0}(\mu \,||\, \mu^0) \geq \bar{\varepsilon}_2 r^2 |\mu|(\mathcal{F}^{\mathrm{reg}}(r)) + \bar{\varepsilon}_2 \sum_{l=1}^{s_0} \int_{\mathcal{N}_l^{\mathrm{reg}}(r)} \mathfrak{d}_{\mathfrak{g}}(x, t_i^0)^2 \,\mathrm{d}|\mu|(x)\,. \tag{A.23}$$
Using (A.21) and (A.23), one gets (3.7a) and, by way of contradiction, (3.7c).

It remains to prove (3.7b). Observe that, using (A.21) and (A.22) and (A.23),
$$\begin{aligned}
\big|\mu(\mathcal{N}_j^{\mathrm{reg}}(r)) - a_j^0\big| &= \Big| \int_{\mathcal{N}_j^{\mathrm{reg}}(r)} \mathrm{d}(\mu - \mu^0) \Big| \\
&= \Big| \int_{\mathcal{X}} \eta_j^0 \mathrm{d}(\mu - \mu^0) + \int_{\mathcal{N}_j^{\mathrm{reg}}(r)} (1 - \eta_j^0) \mathrm{d}(\mu - \mu^0) \\
&\qquad - \sum_{l \neq j} \int_{\mathcal{N}_l^{\mathrm{reg}}(r)} \eta_j^0 \mathrm{d}(\mu - \mu^0) - \int_{\mathcal{F}^{\mathrm{reg}}(r)} \eta_j^0 \mathrm{d}(\mu - \mu^0) \Big| \\
&\leq \Big| \int_{\mathcal{X}} \eta_j^0 \mathrm{d}(\mu - \mu^0) \Big| + \bar{\varepsilon}_2 \sum_{l=1}^{s_0} \int_{\mathcal{N}_l^{\mathrm{reg}}(r)} \mathfrak{d}_{\mathfrak{g}}(x, t_j^0)^2 \mathrm{d}|\mu|(x) + (1 - \bar{\varepsilon}_2 r^2) |\mu|(\mathcal{F}^{\mathrm{reg}}(r)) \\
&\leq \Big| \int_{\mathcal{X}} \eta_j^0 \mathrm{d}(\mu - \mu^0) \Big| + \frac{c_r}{r^2} \Big(\bar{\varepsilon}_2 r^2 |\mu|(\mathcal{F}^{\mathrm{reg}}(r)) + \bar{\varepsilon}_2 \sum_{l=1}^{s_0} \int_{\mathcal{N}_l^{\mathrm{reg}}(r)} \mathfrak{d}_{\mathfrak{g}}(x, t_i^0)^2 \,\mathrm{d}|\mu|(x) \Big) \\
&\leq \Big| \int_{\mathcal{X}} \eta_j^0 \mathrm{d}(\mu - \mu^0) \Big| + c_r \bar{c}_\kappa \Big(\frac{\gamma}{r^2} \Big) \sqrt{s_0}
\end{aligned} \tag{A.24}$$
where
$$c_r := \max(r^2, \frac{1 - \bar{\varepsilon}_2 r^2}{\bar{\varepsilon}_2}) \leq \frac{\max(1, \bar{\varepsilon}_0)}{\bar{\varepsilon}_2}\,.$$
Now, note that
$$\begin{aligned}
\Big| \int_{\mathcal{X}} \eta_j^0 \mathrm{d}(\mu - f^0) \Big| &= \Big| \langle \eta_j^0, \mu - \mu^0 \rangle_{\mathcal{C}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})} \Big| \\
&= \Big| \langle c_j^0, F\mu - F\mu^0 \rangle_{\mathcal{F}} \Big| \\
&\leq \|c_j^0\|_{\mathcal{F}} (\gamma + \|\boldsymbol{y} - F\mu\|_{\mathcal{F}})
\end{aligned} \tag{A.25}$$
and, using (A.20), one has
$$\big\|\kappa c^0 + F\mu - \boldsymbol{y}\big\|_{\mathcal{F}} \leq \gamma + \kappa \|c^0\|_{\mathcal{F}}\,,$$
which leads to
$$\big\|F\mu - \boldsymbol{y}\big\|_{\mathcal{F}} \leq \gamma + 2\kappa \|c^0\|_{\mathcal{F}}\,. \tag{A.26}$$
By (A.25) and (A.26), one has
$$\Big| \int_{\mathcal{X}} \eta_j^0 \mathrm{d}(\mu - \mu^0) \Big| \leq 2 \|c_j^0\|_{\mathcal{F}} (\gamma + \kappa \|c^0\|_{\mathcal{F}}) \leq 2\sqrt{2}\, C_{\mathsf{switch}} (1 + \sqrt{2}\, C_{\mathsf{switch}} c_\kappa) \gamma \tag{A.27}$$
using Theorem 5. Combining (A.24) and (A.27), one gets the result.

A.6 Proof of Theorem 2

The choice $C_{\mathsf{sketch}} = 2C \max(C_1, C_2)$ together with Equation (3.8) ensures that $m \geq m_0$ from Assumption 3. First invoke Theorem 6 to get that there exists a constant $C'_{\mathsf{pivot}} > 0$ which depends only on the kernel K_{pivot} such that, with probability at least $1 - \alpha$, for all $m \geq m_0$, there exist a $(\bar{\varepsilon}_0/4, 3\bar{\varepsilon}_2/2, r_0)$-sketch pivot non-degenerate dual certificate such that $\|c^0\|_{\mathbb{C}^m} \leq C'_{\mathsf{pivot}} C_{\mathsf{switch}} \sqrt{s_0}$ and $(\bar{\varepsilon}_0/4, 3\bar{\varepsilon}_2/2, r_0)$-sketch pivot non-degenerate localizing certificates at points t_i^0 such that $\|c_i^0\|_{\mathbb{C}^m} \leq C'_{\mathsf{pivot}} C_{\mathsf{switch}}$. Then the proof goes the same line as in Appendix A.5 substituting $\bar{\varepsilon}_0$ by $\bar{\varepsilon}_0/4$ and $\bar{\varepsilon}_2$ by $3\bar{\varepsilon}_2/2$.

A.7 Proof of Proposition 2.1

The proof is an application of Theorem 1 using the sinc-4 as pivot kernel. Let $\mu_0 \in \mathcal{M}(\mathcal{X})$ be a sparse measure of s_0 spikes $\{\boldsymbol{x}_1^0, \dots, \boldsymbol{x}_{s_0}^0\}$. First, recall Lemma 2.1 gives a control with probability $1-\alpha$

$$\gamma_n \leq C_\alpha \left(\frac{1}{\tau}\right)^{\frac{d}{2}} \frac{1}{\sqrt{n}}\,, \tag{A.28}$$

where $C_\alpha := 2\sqrt{1 + C_1 \log(C_2/\alpha)}$ is a constant only depending on α, while C_1 and C_2 are universal constants.

By Theorem 3, the sinc-4 pivot kernel $K_{\text{pivot}}(\boldsymbol{s},\boldsymbol{t}) = \Psi_\tau(\boldsymbol{s}-\boldsymbol{t})$ verifies the **LPC** assumption with radius $r_0 = 1/(4\,d)$, $\bar{\varepsilon}_0 \geq 1/(32\,d^3)$, $\bar{\varepsilon}_2 \geq 23/128$ and $\Delta_0 = 147.77\, {s_0}^{1/4} d^{7/4}$, which are independent of the bandwidth τ. Moreover, its associated Fisher-Rao distance is a rescaled euclidean $\mathfrak{d}_{\mathfrak{g}_\tau}(\boldsymbol{s},\boldsymbol{t}) = (2\sqrt{3}\tau)^{-1}\|\boldsymbol{s}-\boldsymbol{t}\|^2$. Thus, as discussed in Equation (3.10b), taking τ verifying Equation (2.3–$\mathbf{H}_\tau$) implies that the target belongs to the model set $\mathcal{M}_{s_0,\Delta_0,\mathfrak{d}_\tau}$. Then, we may invoke Theorem 1 with regularisation

$$\kappa = c_\kappa \gamma / \sqrt{s_0} = \frac{1}{\sqrt{2} C_{\text{switch}}} C_\alpha \tau^{-\frac{d}{2}} \frac{1}{\sqrt{n}} \frac{1}{\sqrt{s_0}},$$

where we use $c_\kappa = 1/\sqrt{2}C_{\text{switch}}$ as in Remark 3.2. Injecting this in Equations (3.7a) to (3.7c), one gets for any radius $r > 0$ such that $r < \min\left(r_0, \sqrt{\bar{\varepsilon}_0/\bar{\varepsilon}_2}\right) = \mathcal{O}(d^{-3/2}) =: c'_d$,

- Control of the far region:

$$|\mu|(\mathcal{F}^{\text{reg}}(r)) \leq 2\sqrt{2} C_{\text{switch}} \frac{128}{23} C_\alpha \tau^{-\frac{d}{2}} \left(\frac{1}{r^2\sqrt{n}}\right) \sqrt{s_0}\,, \tag{A.29}$$

- Control of all the near regions: for all $k \in [s_0]$,

$$|\mu(\mathcal{N}_k^{\text{reg}}(r)) - a_k^0| \leq 2\sqrt{2} C_{\text{switch}} \frac{128}{23} C_\alpha \tau^{-\frac{d}{2}} \left(\frac{1}{r^2\sqrt{n}}\right) \sqrt{s_0} + 4\sqrt{2} C_{\text{switch}} C_\alpha \tau^{-\frac{d}{2}} \frac{1}{\sqrt{n}}\,, \tag{A.30}$$

- Detection level: for all Borelian $A \subset \mathcal{X}$ such that $|\mu|(A) > 2\sqrt{2} C_{\text{switch}} \frac{128}{23} C_\alpha \tau^{-\frac{d}{2}} \left(\frac{1}{r^2\sqrt{n}}\right) \sqrt{s_0}$, there exists $\boldsymbol{x}_k^0$ such that

$$\min_{\boldsymbol{t} \in A} \mathfrak{d}_{\mathfrak{g}}(t, \boldsymbol{x}_k^0) \leq r\,, \tag{A.31}$$

The last part of the proof consists in realizing that we may take the radius $r = r_n$ to depend on n. The only constraint is that $r_n^2\sqrt{n}$ goes to $+\infty$ as n grows so that the controls asymptotically make sense, *i.e.* the BLASSO estimator concentrates around the true support. Thus, we may consider a slowly diverging sequence δ_n such that $r_n^2\sqrt{n} = \delta_n^2$, or equivalently $r_n = n^{-1/4}\delta_n$. For comments over the choice of δ_n, we refer to Remark 2.3. Note that, for large n, $\delta_n^2 = r_n^2\sqrt{n} < \sqrt{n}$ so that the first term on the right-hand side of Equation (A.30) is the dominant one. Hence, we obtain the final control in Equation (2.9b)

$$|\mu(\mathcal{N}_k^{\text{reg}}(r)) - a_k^0| \leq 6\sqrt{2} C_{\text{switch}} \frac{128}{23} C_\alpha \tau^{-\frac{d}{2}} \left(\frac{1}{r^2\sqrt{n}}\right) \sqrt{s_0}.$$

Finally, the condition $r_n < c'_d$ can be restated implicitly in n as $n \geq (c'_d)^{-4}\delta_n^4$, and the constant $c_d := (c'_d)^{-4} = \mathcal{O}(d^6)$ depends polynomially on the dimension. This concludes the proof.

A.8 Proof of Proposition 2.2

The proof is an application of Theorem 2 which give the controls for sketched BLASSO problems. We first prove the control on the sketch noise level and then turn on proving the actual proposition.

A.8.1 Proof of Lemma 2.2

First, we recall and prove Lemma 2.2 giving the control over the sketched noise level with probability $1-\alpha$ w.r.t. the joint draws of the sample $\boldsymbol{Z}$ and the sketch. Precisely, we prove that with probability $1-\alpha$ it holds that

$$\|\Gamma_{\text{sketch}}\|_{\mathbb{C}^m} \leq C_{\alpha,m} \frac{1}{\sqrt{n}},$$

with

$$C_{\alpha,m} = 2\sqrt{\left[\left(\frac{1}{\tau}\right)^d + \frac{1}{2\sqrt{m}} \left\|\frac{U_\tau}{\Lambda}\right\|_\infty \log\left(\frac{2}{\alpha}\right)\right]\left[1 + C_1 \log\left(\frac{2C_2}{\alpha}\right)\right]},$$

and C_1, C_2 universal constants.

We follow the proof of De Castro et al. (2021, Lemma 3) and adapt it with a union bound argument to take into account the extra-randomness from the sketch. First, introduce the i.i.d. random variables $(\boldsymbol{Y}_j)_{j=1}^n$

$$\boldsymbol{Y}_j := \mathcal{A}\delta_{\boldsymbol{Z}_j} - \mathbb{E}_{\boldsymbol{Z}}\,\mathcal{A}\delta_{\boldsymbol{Z}},$$

where $\mathcal{A} : \nu \in \mathcal{M}(\mathbb{R}^d) \mapsto \frac{1}{\sqrt{m}}\left(W(\boldsymbol{\omega}_i)\,\mathfrak{F}[\nu](\boldsymbol{\omega}_i)\right)_{i=1}^m \in \mathbb{C}^m$ is the so-called *sketching* operator yielding the random Fourier features of a Radon measure ν. Then, we have

$$\left\|\Gamma_{\text{sketch}}\right\|_{\mathbb{C}^m}^2 = \left\|\frac{1}{n}\sum_{j=1}^n \boldsymbol{Y}_j\right\|_{\mathbb{C}^m}^2 = \frac{1}{n^2}\sum_{j=1}^n \|\boldsymbol{Y}_j\|_{\mathbb{C}^m}^2 + \frac{1}{n^2}\sum_{j\neq l}^n \langle \boldsymbol{Y}_j, \boldsymbol{Y}_l\rangle_{\mathbb{C}^m}, \tag{A.32}$$

Fix $\alpha \in (0,1)$, and let us assume that there exists a constant M such that $\|\boldsymbol{Y}_j\|_{\mathbb{C}^m}^2 \leq M$. Then, using the same arguments as in the proof of De Castro et al. (2021, Lemma 3), the first term of Equation (A.32) is bounded by M/n while the second term is a canonical U-process. Thus, Arcones and Giné (1993, Proposition 2.3) ensures that there exists two universal constants $C_1, C_2 > 0$ such that

$$\left\|\Gamma_{\text{sketch}}\right\|_{\mathbb{C}^m}^2 \leq M\Big(1 + C_1 \log\left(\frac{2C_2}{\alpha}\right)\Big)\frac{1}{n} \tag{A.33}$$

with probability at least $1 - \frac{\alpha}{2}$.

We then proceed to show that $\|\boldsymbol{Y}_j\|_{\mathbb{C}^m}^2$ is bounded with probability $1 - \alpha/2$ using Hoeffding's inequality. Indeed, for any $j \in [n]$ we may write

$$\|\boldsymbol{Y}_j\|_{\mathbb{C}^m}^2 = \frac{1}{m}\sum_{i=1}^m \left|W(\boldsymbol{\omega}_i)\left(e^{-\imath\boldsymbol{\omega}_i^\top \boldsymbol{Z}_j} - \mathfrak{F}[f^0](\boldsymbol{\omega}_i)\right)\right|^2 = \frac{1}{m}\sum_{i=1}^m X_i^j,$$

with $X_i^j = \left|W(\boldsymbol{\omega}_i)\left(e^{-\imath\boldsymbol{\omega}_i^\top \boldsymbol{Z}_j} - \mathfrak{F}[f^0](\boldsymbol{\omega}_i)\right)\right|^2$. Moreover, the expectation with respect to the draw of $\boldsymbol{\omega}_{1:m}$ is

$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}}\|\boldsymbol{Y}_j\|_{\mathbb{C}^m}^2 = \|L_\tau\delta_{\boldsymbol{Z}_j} - L_\tau f^0\|_{\mathcal{F}_\tau}^2,$$

where we recall the operator $L_\tau\nu = \lambda_\tau \star \nu$ is the smoothing kernel operator, which can be seen as the kernel mean embedding of ν in the RKHS $\mathcal{F}_\tau$. The $(X_i^j)_{i=1}^m$ are independent and bounded random variables with

$$0 \leq X_i^j \leq W^2(\boldsymbol{\omega}_i)\left(|e^{-\imath\boldsymbol{\omega}_i^\top \boldsymbol{Z}_j}| + |\,\mathfrak{F}[f^0](\boldsymbol{\omega}_i)|\right)^2 \leq 4\|W^2\|_\infty\,,$$

using Jensen's inequality to get that

$$|\,\mathfrak{F}[f^0](\boldsymbol{\omega}_i)| = |\,\mathbb{E}_{\boldsymbol{Z}}\,e^{-\imath\boldsymbol{\omega}_i^\top \boldsymbol{Z}}| \leq \mathbb{E}_{\boldsymbol{Z}}\,|e^{-\imath\boldsymbol{\omega}_i^\top \boldsymbol{Z}}| = 1\,.$$

Applying Hoeffding's inequality yields that with probability greater than $1 - \alpha/2$

$$\|\boldsymbol{Y}_j\|_{\mathbb{C}^m}^2 \leq \|L_\tau\delta_{\boldsymbol{Z}_j} - L_\tau f^0\|_{\mathcal{F}_\tau}^2 + \frac{2\|W^2\|_\infty}{\sqrt{m}}\log\left(\frac{2}{\alpha}\right)$$

Now observe that

$$\|L_\tau\delta_{\boldsymbol{Z}_j}\|_{\mathcal{F}_\tau}^2 = \lambda_\tau(\boldsymbol{Z}_j - \boldsymbol{Z}_j) = \lambda_\tau(\boldsymbol{0})$$
$$\|L_\tau f^0\|_{\mathcal{F}_\tau}^2 \leq \mathbb{E}_{\boldsymbol{Z}}\,\|L_\tau\delta_{\boldsymbol{Z}}\|_{\mathcal{F}_\tau}^2 = \lambda_\tau(\boldsymbol{0})$$

by the reproducing property of the RKHS $\mathcal{F}_\tau$ and Jensen's inequality. We deduce that

$$\|L_\tau\delta_{\boldsymbol{Z}_j} - L_\tau f^0\|_{\mathcal{F}_\tau}^2 \leq 4\lambda_\tau(\boldsymbol{0})\,,$$

so that we can conclude

$$\|\boldsymbol{Y}_j\|_{\mathbb{C}^m}^2 \leq M_{\alpha,m} := 4\lambda_\tau(0) + \frac{2\|W^2\|_\infty}{\sqrt{m}}\log\left(\frac{2}{\alpha}\right) \tag{A.34}$$

with probability greater than $1 - \alpha/2$. A union bound bound yields that the joint probability of Equations (A.33) and (A.34) is greater than $1 - \alpha$. The proof ends by plugging $\lambda_\tau(\boldsymbol{0}) = (\frac{1}{\tau})^d$ and $W^2 = \frac{U_\tau}{\Lambda}$.

A.8.2 Proof of the main proposition

The proof of Proposition 2.2 follows the same line as the proof of Proposition 2.1 given above in Appendix A.7. First, remark that Theorem 4 ensures the sinc-4 kernel verifies Assumption 3 and is thus amenable to sketching with the following constants:

$$
\begin{aligned}
m_0 &= Cs_0\left(C_1\log(s_0)\log\left(\frac{s_0}{\alpha}\right)+C_2\log\left(\frac{(s_0N)^d}{\alpha}\right)\right)\\
N &:= |\mathcal{X}|32\sqrt{12}\sqrt{C_\Lambda}d^{7/2}+\frac{128}{23}(12\sqrt{12}C_\Lambda d^{1/2}+\sqrt{C_\Lambda}12d)\\
C_1 &= \mathcal{O}\left(d^{15/2}C_\Lambda\right)\\
C_2 &= \mathcal{O}\left(d^6C_\Lambda\right)\\
C_\Lambda &:= \sup_{\boldsymbol{\omega}\in[-\frac{1}{\tau},\frac{1}{\tau}]^d}\frac{f_\tau^{(4)}}{\Lambda}(\boldsymbol{\omega})
\end{aligned}
$$

Where the diameter $|\mathcal{X}| = \sup_{\boldsymbol{s},\boldsymbol{t}\in\mathcal{X}}\mathfrak{d}_{\mathfrak{g}_\tau}(\boldsymbol{s},\boldsymbol{t})$, and C depends on Ψ_τ and polynomially on d.

With our hypothesis on the bandwidth τ Equation (2.3–$\mathbf{H}_\tau$), the target μ^0 belongs to the model set $\mathcal{M}_{s_0,\Delta_0,\mathfrak{d}_\tau}$ where $\mathfrak{d}_\tau$ is the Fisher-Rao metric of the sinc-4 kernel Ψ_τ (see Equation (3.10b)). We can then invoke Theorem 2, which ensures a positive constant C'_{pivot} exists and yields the sketched BLASSO controls with high-probability. Precisely, take the sketch size $m \geq m_0$, and use the control on the sketched noise level γ_{sketch} to tune the regularisation

$$
\kappa = \frac{c'_\kappa\gamma_{\text{sketch}}}{\sqrt{s_0}} = \frac{1}{C'_{\text{pivot}}C_{\text{switch}}}\frac{C_{\alpha,m}}{\sqrt{n}}\frac{1}{\sqrt{s_0}},
$$

where we used the optimal $c'_\kappa = \frac{1}{C'_{\text{pivot}}C_{\text{switch}}}$. Injecting this in Equations (3.9a) to (3.9c), one gets for any radius $r>0$ such that $r < \min\left(r_0, \sqrt{\bar{\varepsilon}_0/6\bar{\varepsilon}_2}\right) =: c'_d$,

- Control of the far region:

$$
|\mu|(\mathcal{F}^{\text{reg}}(r)) \leq 2C'_{\text{pivot}}C_{\text{switch}}\frac{256}{69}C_{\alpha,m}\left(\frac{1}{r^2\sqrt{n}}\right)\sqrt{s_0}\,, \tag{A.35}
$$

- Control of all the near regions: for all $k\in[s_0]$,

$$
|\mu(\mathcal{N}_k^{\text{reg}}(r)) - a_k^0| \leq 2C'_{\text{pivot}}C_{\text{switch}}\frac{256}{69}C_{\alpha,m}\left(\frac{1}{r^2\sqrt{n}}\right)\sqrt{s_0} + 4C'_{\text{pivot}}C_{\text{switch}}\frac{C_{\alpha,m}}{\sqrt{n}}\,, \tag{A.36}
$$

- Detection level: for all Borelian $A\subset\mathcal{X}$ such that $|\mu|(A) > 2C'_{\text{pivot}}C_{\text{switch}}\frac{256}{69}C_{\alpha,m}\left(\frac{1}{r^2\sqrt{n}}\right)\sqrt{s_0}$, there exists $\boldsymbol{x}_k^0$ such that

$$
\min_{\boldsymbol{t}\in A}\mathfrak{d}_{\mathfrak{g}}(t,\boldsymbol{x}_k^0) \leq r\,, \tag{A.37}
$$

Using a similar reasoning as in Appendix A.7, we finish the proof by taking the radius $r = r_n$ to depend on n. Let δ_n be a slowly diverging sequence such that $r_n^2\sqrt{n} = \delta_n^2$, or equivalently $r_n = n^{-1/4}\delta_n$. Again, the first term on the right-hand side of Equation (A.36) is the dominant one. Hence, we obtain the final control in Equation (2.9b)

$$
|\mu(\mathcal{N}_k^{\text{reg}}(r)) - a_k^0| \leq 6C'_{\text{pivot}}C_{\text{switch}}\frac{256}{69}C_{\alpha,m}\left(\frac{1}{r^2\sqrt{n}}\right)\sqrt{s_0}.
$$

Again, the condition on the radius $r_n < c'_d$ is met for any such that $n \geq (c'_d)^{-4}\delta_n^4$, with $c_d := (c'_d)^{-4} = \mathcal{O}(d^6)$ scaling polynomially in the dimension d. This concludes the proof.

B Technical results and remarks

B.1 Covariant derivatives: the translation invariant kernel case

The covariant derivative of the kernel $K(\boldsymbol{s},\boldsymbol{t}) = \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}[\psi_{\boldsymbol{\omega}}(\boldsymbol{s})\overline{\psi_{\boldsymbol{\omega}}(\boldsymbol{t})}]$ is properly introduced by Poon et al. (2023, Section 4.1) as a "bi"-multilinear map $K^{(ij)}(\boldsymbol{s},\boldsymbol{t}) : (\mathbb{C}^d)^i\times(\mathbb{C}^d)^j\to\mathbb{C}$, and it involves the covariant derivatives

of order i and j of the feature map $\psi_{\boldsymbol{\omega}} : \mathbb{R}^d \to \mathbb{C}$. In the case of translation invariant kernels $K(\boldsymbol{s}, \boldsymbol{t}) = \rho(\boldsymbol{s} - \boldsymbol{t})$, the Riemannian covariant derivatives $K^{(i,j)}(\boldsymbol{s}, \boldsymbol{t})[U, V]$ and its Riemannian operator norm $\|K^{(i,j)}(\boldsymbol{s}, \boldsymbol{t})\|_{\boldsymbol{s},\boldsymbol{t}}$ are given by the standard Euclidean derivatives $(-1)^j \nabla^{i+j} \rho(h)$ of ρ at point $\boldsymbol{h} = \boldsymbol{s} - \boldsymbol{t}$ and the Euclidean operator norm, namely

$$K^{(i,j)}(\boldsymbol{s}, \boldsymbol{t})[U, V] = (-1)^j \mathrm{Trace}(\nabla^{i+j} \rho(\boldsymbol{h}) \times U_1 \otimes \cdots \otimes U_i \otimes V_1 \otimes \cdots \otimes V_j)\,, \tag{B.1a}$$

$$\|K^{(i,j)}(\boldsymbol{s}, \boldsymbol{t})\|_{\boldsymbol{s},\boldsymbol{t}} = \Big\| \nabla^{i+j} \rho(\boldsymbol{h}) \bigtimes_{l=1}^{i+j} \mathfrak{g}^{-\frac{1}{2}} \Big\|_2\,. \tag{B.1b}$$

where $U \in (\mathbb{C}^d)^i$ (resp. $V \in (\mathbb{C}^d)^j$) are i (resp. j) tangent vectors at points $\boldsymbol{s}$ (resp. $\boldsymbol{t}$), $\otimes$ denotes the tensor product and

$$\forall T \in (\mathbb{C}^d)^{i+j}\,, \quad \big(\nabla^{i+j} \rho(h) \bigtimes_{l=1}^{i+j} \mathfrak{g}^{-\frac{1}{2}}\big)[T] \coloneqq \nabla^{i+j} \rho(h)[\mathfrak{g}^{-\frac{1}{2}} T_1, \ldots, \mathfrak{g}^{-\frac{1}{2}} T_{i+j}]\,. \tag{B.1c}$$

In particular, one has

$$T_1^\top \nabla \rho(h) = K^{(1,0)}(\boldsymbol{s}, \boldsymbol{t})[T_1] = -K^{(0,1)}(\boldsymbol{s}, \boldsymbol{t})[T_1]\,, \tag{B.1d}$$

$$T_1^\top \nabla^2 \rho(h) T_2 = K^{(2,0)}(\boldsymbol{s}, \boldsymbol{t})[T_1, T_2] = K^{(0,2)}(\boldsymbol{s}, \boldsymbol{t})[T_1, T_2] = -K^{(1,1)}(\boldsymbol{s}, \boldsymbol{t})[T_1, T_2]\,, \tag{B.1e}$$

$$\Big\| \mathfrak{g}^{-\frac{1}{2}} \nabla \rho(h) \Big\|_2 = \|K^{(1,0)}(\boldsymbol{s}, \boldsymbol{t})\|_s\,, \tag{B.1f}$$

$$\Big\| \mathfrak{g}^{-\frac{1}{2}} \nabla^2 \rho(h) \mathfrak{g}^{-\frac{1}{2}} \Big\|_2 = \|K^{(2,0)}(\boldsymbol{s}, \boldsymbol{t})\|_{\boldsymbol{s},\boldsymbol{t}}\,. \tag{B.1g}$$

B.2 Model RKHS and dual operator

Proposition B.1. *Assume* (1.3–$\mathbf{H}_{\mathsf{mod}}$) *holds true.*

- *The unique RKHS of* K_{mod} *is given by*

$$\mathcal{H}_{\mathsf{mod}} = \big\{\eta : \mathcal{X} \to \mathbb{R} \;\big|\; \exists c \in \mathcal{F}\,,\ \eta = \eta_c\big\} = \big\{\eta : \mathcal{X} \to \mathbb{R} \;\big|\; \exists! c \in \overline{\mathrm{Im}(F)}\,,\ \eta = \eta_c\big\}\,, \tag{B.2}$$

  and for all c *orthogonal to* $\overline{\mathrm{Im}(F)}$ *in* $\mathcal{F}$, $\eta_c = 0$.

- *The isometry is given by the mapping* $c \in \overline{\mathrm{Im}(F)} \mapsto \eta_c \in \mathcal{H}_{\mathsf{mod}}$.

- *The norms satisfy*

$$\forall c \in \overline{\mathrm{Im}(F)}\,, \quad \|\eta_c\|_{\mathcal{H}_{\mathsf{mod}}} = \|c\|_{\mathcal{F}}\,.$$

Proof. See Steinwart and Christmann 2008, Theorem 4.21. □

B.3 Proof of Proposition 3.1

By Assumption (1.3–$\mathbf{H}_{\mathsf{mod}}$), it holds that $\mathcal{H}_{\mathsf{mod}} \subset \mathcal{C}(\mathcal{X})$. Define $\mathrm{Id}_{\mathsf{mod}} : \mathcal{H}_{\mathsf{mod}} \to \mathcal{C}(\mathcal{X})$ the identity mapping. One has

$$F^\star c = \mathrm{Id}_{\mathsf{mod}}(\eta_c)\,, \text{ for all } c \in \overline{\mathrm{Im}(F)}\,, \tag{B.3a}$$

$$\langle \mathrm{Id}_{\mathsf{mod}}(\eta_c), \nu \rangle_{\mathcal{C}(\mathcal{X}),\mathcal{M}(\mathcal{X})} = \langle c, F\nu \rangle_{\mathcal{F}}\,, \text{ for all } c \in \mathcal{F},\, \nu \in \mathcal{M}(\mathcal{X})\,. \tag{B.3b}$$

The result follows from (B.3) and Proposition B.1. For the sake of readability, we denote $\mathrm{Id}_{\mathsf{mod}}(\eta_c)$ by η_c.

C Functional framework

C.1 Radon measures

Definition 7 (Set $\big(\mathcal{M}(\mathbb{R}^d), \|\cdot\|_{\mathrm{TV}}\big)$)**.** *We work in the vector space* $\mathcal{M}(\mathbb{R}^d)$ *of finite signed Radon measures on* $\mathbb{R}^d$*, endowed with the total variation norm:*

$$\|\mu\|_{\mathrm{TV}} \coloneqq \int_{\mathbb{R}^d} \mathrm{d}|\mu|,$$

where $|\mu| = \mu^+ + \mu^-$ *and* $\mu = \mu^+ - \mu^-$ *is the Jordan decomposition of* $\mu \in \mathcal{M}(\mathbb{R}^d)$.

By the Riesz-Markov theorem, this space is the dual of real-valued continuous functions on $\mathbb{R}^d$ *vanishing at infinity*[3] equipped with the supremum norm, which we denote as $\left(C_0(\mathbb{R}^d, \mathbb{R}), \|\cdot\|_\infty\right)$ or $\left(C_0(\mathbb{R}^d), \|\cdot\|_\infty\right)$. The total variation norm then admit a variational formulation as:

$$\|\mu\|_{\mathrm{TV}} = \sup\left\{\int \eta \,\mathrm{d}\mu;\ \eta \in \mathcal{C}(\mathbb{R}^d),\ \|\eta\|_\infty \leq 1\right\}, \tag{C.1}$$

C.2 Fourier transform

Here, we detail the definition of the Fourier transform and its inverse used throughout the paper.

Definition 8 (Fourier transform)**.** *For any $g \in L^1(\mathbb{R}^d)$, we define its Fourier transform $\mathfrak{F}[g]$ and its inverse as*

$$\forall \boldsymbol{\omega} \in \mathbb{R}^d, \quad \mathfrak{F}[g](\boldsymbol{\omega}) = \int_{\mathbb{R}^d} g(\boldsymbol{z}) e^{-i\boldsymbol{\omega}^\top \boldsymbol{z}} \,\mathrm{d}\boldsymbol{z}. \tag{C.2a}$$

Furthermore, for g continuous and both g and $\mathfrak{F}[g]$ in $L^1(\mathbb{R}^d)$, we define the inverse transform

$$\forall \boldsymbol{z} \in \mathbb{R}^d, \quad \mathfrak{F}^{-1}\left[\mathfrak{F}[g]\right](\boldsymbol{z}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \mathfrak{F}[g](\boldsymbol{\omega}) e^{i\boldsymbol{\omega}^\top \boldsymbol{z}} \,\mathrm{d}\boldsymbol{\omega}. \tag{C.2b}$$

C.3 Reproducing kernel Hilbert space of translation invariant kernel

In this subsection, we remind particular results for translation invariant kernels $k(\boldsymbol{s}, \boldsymbol{t}) = \rho(\boldsymbol{t} - \boldsymbol{s})$ and their characterisation in the Fourier domain. We begin by stating Bochner's theorem, a well known result characterizing functions of the positive definite type as the characteristic function of a positive measure, coined the *spectral measure*.

Theorem 7 (Bochner)**.** *Let $k(\boldsymbol{s}, \boldsymbol{t}) = \rho(\boldsymbol{t} - \boldsymbol{s})$. Then k is a reproducing kernel if and only if there exists a positive measure ν such that*

$$\rho(\boldsymbol{t} - \boldsymbol{s}) = \int e^{+\imath \boldsymbol{\omega}^\top (\boldsymbol{t} - \boldsymbol{s})} \,\mathrm{d}\nu(\boldsymbol{\omega}).$$

Furthermore, if the kernel is normalized, i.e. $\rho(\mathbf{0}) = 1$, then ν is a probability distribution.

This allows to define the RKHS associated to the translation invariant kernel ρ as functions with sufficiently fast decay rate in the Fourier domain.

Proposition C.1 (RKHS of translation invariant kernels)**.** *The RKHS of the translation invariant kernel $k(\boldsymbol{s}, \boldsymbol{t}) = \rho(\boldsymbol{t} - \boldsymbol{s})$ is defined as*

$$\mathcal{H}_\rho := \left\{ g \in L^2(\mathcal{X}) \ :\ \|g\|_{\mathcal{H}_\rho}^2 = \frac{1}{(2\pi)^d} \int_{\mathcal{X}} \frac{|\mathfrak{F}[g](\boldsymbol{\omega})|^2}{\mathfrak{F}[\rho](\boldsymbol{\omega})} \,\mathrm{d}\boldsymbol{\omega} \right\}$$

with scalar product

$$\langle g_1, g_2 \rangle_{\mathcal{H}_\rho} = \frac{1}{(2\pi)^d} \int \frac{\mathfrak{F}[g_1](\boldsymbol{\omega}) \overline{\mathfrak{F}[g_2](\boldsymbol{\omega})}}{\mathfrak{F}[\rho](\boldsymbol{\omega})} \,\mathrm{d}\boldsymbol{\omega}$$

Element of proof. A formal proof is given in Wendland (2005, Theorem 10.12, p. 139), here we focus on proving the reproducing property. Moreover, we refer to Bach (2024, Section 7.3.3, p.192) for readers interested in an intuitive construction of the RKHS norm.

First, we prove that $(\mathcal{H}_\rho, \langle\cdot,\cdot\rangle_{\mathcal{H}_\rho})$ contains function of the form $k(\boldsymbol{s}, \cdot)$. To see this, one needs to remember that the Fourier transform is covariant by translation, meaning $\mathfrak{F}[f(\cdot - \boldsymbol{s})](\boldsymbol{\omega}) = e^{-\imath \boldsymbol{\omega}^\top \boldsymbol{s}} \mathfrak{F}[f](\boldsymbol{\omega})$. Hence, we have that

$$\|k(\boldsymbol{s}, \cdot)\|_{\mathcal{H}_\rho} = \|\rho(\cdot - \boldsymbol{s})\|_{\mathcal{H}_\rho} = \frac{1}{(2\pi)^d} \int \frac{|\mathfrak{F}[\rho](\boldsymbol{\omega})|^2}{\mathfrak{F}[\rho](\boldsymbol{\omega})} |e^{-\imath \boldsymbol{\omega}^\top \boldsymbol{s}}|^2 \,\mathrm{d}\boldsymbol{\omega} = \int \frac{\mathfrak{F}[\rho](\boldsymbol{\omega})}{(2\pi)^d} \,\mathrm{d}\boldsymbol{\omega} = \rho(\mathbf{0}) < +\infty.$$

Thus, $k(\boldsymbol{s}, \cdot)$ belongs to the RKHS. Following the same line, we can prove that the reproducing property holds. Let $g \in \mathcal{H}_\rho$ and $\boldsymbol{s} \in \mathbb{R}^d$

$$\langle g, \rho(\cdot - \boldsymbol{s}) \rangle_{\mathcal{H}_\rho} = \frac{1}{(2\pi)^d} \int \frac{\mathfrak{F}[g](\boldsymbol{\omega}) \overline{\mathfrak{F}[\rho](\boldsymbol{\omega})}}{\mathfrak{F}[\rho](\boldsymbol{\omega})} e^{+\imath \boldsymbol{\omega}^\top \boldsymbol{s}} \,\mathrm{d}\boldsymbol{\omega} = \mathfrak{F}^{-1}[\mathfrak{F}[g]](\boldsymbol{s}) = g(\boldsymbol{s}).$$

Completeness is treated in Wendland (2005, Theorem 10.12, p. 139). □

[3] f vanishes at infinity iff $\forall \varepsilon > 0$, there exists a compact $K_\varepsilon \subset \mathbb{R}^d$ such that $\forall x \in K_\varepsilon^c$, $|f(x)| < \varepsilon$

Remark C.1 (Link between the Fourier transform of ρ and its spectral measure ν)**.** *With the choice of convention introduced in Appendix C.2 for the inverse Fourier transform, the following identity holds*

$$\rho = (2\pi)^d \, \mathfrak{F}^{-1}[\nu] \iff \nu = \frac{\mathfrak{F}[\rho]}{(2\pi)^d}.$$

As a consequence we have

$$\begin{aligned}
\|\rho \star g\|^2_{\mathcal{H}_\rho} &= \frac{1}{(2\pi)^d} \int_{\mathcal{X}} \frac{|\mathfrak{F}[\rho](\boldsymbol{\omega})|^2 \, |\mathfrak{F}[g](\boldsymbol{\omega})|^2}{\mathfrak{F}[\rho](\boldsymbol{\omega})} \, \mathrm{d}\boldsymbol{\omega}, \\
&= \int_{\mathcal{X}} \frac{\mathfrak{F}[\rho](\boldsymbol{\omega})}{(2\pi)^d} \, |\mathfrak{F}[g](\boldsymbol{\omega})|^2 \, \mathrm{d}\boldsymbol{\omega}, \\
&= \int_{\mathcal{X}} \nu(\boldsymbol{\omega}) \, |\mathfrak{F}[g](\boldsymbol{\omega})|^2 \, \mathrm{d}\boldsymbol{\omega},
\end{aligned}$$

If the kernel is normalized, i.e. $\rho(0) = 1$, then ν is a probability distribution and the last line can be written as an expectation.

C.4 Fourier transform and spectral measure of the sinus cardinal smoothing kernel λ_τ

We recall our choice of smoothing kernel λ_τ which is the rescaled sinus cardinal

$$\lambda_\tau(\boldsymbol{x}) = \left(\frac{1}{\tau}\right)^d \operatorname{sinc}\left(\frac{\boldsymbol{x}}{\tau}\right),$$

with its spectral measure from Theorem 7 denoted as U_τ. Denoting $\operatorname{sinc}_\tau = \operatorname{sinc}(\cdot/\tau)$, we have with our choice of Fourier conventions that $\Pi_1(\boldsymbol{\omega}) \coloneqq \mathfrak{F}[\operatorname{sinc}_1] = \pi^d \mathbb{1}_{[-1,1]^d}$. Thus, the Fourier transform of its scaled version is the gate function

$$\Pi_\tau(\boldsymbol{\omega}) \coloneqq \mathfrak{F}[\operatorname{sinc}_\tau](\boldsymbol{\omega}) = \mathfrak{F}[\operatorname{sinc}_1(\cdot/\tau)](\boldsymbol{\omega}) = (\pi\tau)^d \mathbb{1}_{[-1/\tau,1/\tau]^d}(\boldsymbol{\omega}),$$

with a frequency cut-off as $1/\tau$. Finally, by linearity, the Fourier transform of the smoothing kernel is

$$\mathfrak{F}[\lambda_\tau](\boldsymbol{\omega}) = \pi^d \mathbb{1}_{[-1/\tau,1/\tau]^d}(\boldsymbol{\omega}), \tag{C.3}$$

ans its spectral measure is

$$U_\tau(\boldsymbol{\omega}) = \frac{\mathfrak{F}[\lambda_\tau]}{(2\pi)^d}(\boldsymbol{\omega}) = \frac{1}{2^d} \mathbb{1}_{[-1/\tau,1/\tau]^d}(\boldsymbol{\omega}). \tag{C.4}$$

C.5 Spectral measure of the sinc-4 kernel

The sinc-4 kernel $\Psi_\tau(\boldsymbol{x}-\boldsymbol{y}) = \operatorname{sinc}(\frac{\boldsymbol{x}-\boldsymbol{y}}{4\tau})^4 = \operatorname{sinc}_{4\tau}(\boldsymbol{x}-\boldsymbol{y})^4$ defined in Equation (1.9d) is a translation invariant kernel with associated RKHS given by

$$\mathcal{H}_{\operatorname{sinc}} \coloneqq \left\{ g : \mathbb{R}^d \to \mathbb{R} \text{ s.t. } \|g\|^2_{\mathcal{H}_{\operatorname{sinc}}} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{|\mathfrak{F}[g]|^2}{\mathfrak{F}[\Psi_\tau]} < +\infty \right\}. \tag{C.5}$$

We denote its spectral measure as $f^{(4)}_\tau = \mathfrak{F}[\Psi_\tau]/(2\pi)^d$, which verifies by Bochner's theorem (Theorem 7)

$$\Psi_\tau(\boldsymbol{x}) = \int_{\mathbb{R}^d} e^{+\imath \boldsymbol{x}^\top \boldsymbol{\omega}} f^{(4)}_\tau(\boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega}.$$

Since $\Psi_\tau(\mathbf{0}) = 1$, then $f^{(4)}_\tau$ is a normalized probability density function. Moreover, it has an explicit analytical form. Indeed, remember that with our choice of Fourier conventions, the function $\operatorname{sinc}_{4\tau} = \operatorname{sinc}(\cdot/4\tau)$ is the inverse Fourier transform of the rescaled low-pass filter $\Pi_{4\tau}(\boldsymbol{\omega}) = (4\pi\tau)^d \mathbb{1}_{[-1/4\tau,1/4\tau]^d}(\boldsymbol{\omega})$. Thus, $f^{(4)}_\tau$ is proportional to $\Pi^{(\star 4)}_{4\tau}$, which we define as the convolution of $\Pi_{4\tau}$ with itself taken four time. The proportionality constant is such that $f^{(4)}_\tau$ is a normalized probability density function. Precisely, we know that $f^{(4)}_\tau$ is even since Ψ_τ is even and real-valued, so that

$$f^{(4)}_\tau = \check{f}^{(4)}_\tau = \mathfrak{F}^{-1}[\Psi_\tau] = \mathfrak{F}^{-1}[\operatorname{sinc}^4_{4\tau}] = \mathfrak{F}^{-1}[\mathfrak{F}^{-1}[\Pi_{4\tau}]^4] = (2\pi)^{-4d} \, \mathfrak{F}^{-1}[\mathfrak{F}[\Pi_{4\tau}](-\cdot)^4] = (2\pi)^{-4d} \Pi^{(\star 4)}_{4\tau}$$

where we used that $g^{(\star n)} = \mathfrak{F}^{-1}[\mathfrak{F}[g]^n]$ and $\mathfrak{F}[\Pi_{4\tau}]$ is even.

Note that $f_\tau^{(4)}$ may also be interpreted as the density of the sum of independent uniform random variable on $[-1/4\tau, 1/4\tau]^d$, also known as the (rescaled) Irwin-Hall distribution of order 4. It is even by symmetricity around $\mathbf{0}$ and its expression is

$$\forall \boldsymbol{\omega} \in [1/\tau, 1/\tau]^d, \quad f_\tau^{(4)}(\boldsymbol{\omega}) = (2\tau)^d \prod_{j=1}^d \frac{1}{6} \sum_{k=1}^{\lfloor 2\tau\omega_j + 2 \rfloor} (-1)^k \binom{4}{k} \big(2\tau\omega_j + 2 - k\big)^4 \tag{C.6}$$

Moreover, its mode is in $\boldsymbol{\omega} = \mathbf{0}$ and its maximum value $(\frac{2}{3})^d (2\tau)^d$ so that

$$\sup_{\boldsymbol{\omega} \in [-1/\tau, 1/\tau]^d} \mathfrak{F}[\Psi_\tau](\boldsymbol{\omega}) = (2\pi)^d \sup_{\boldsymbol{\omega} \in [-1/\tau, 1/\tau]^d} f_\tau^{(4)}(\boldsymbol{\omega}) = (2\pi)^d \Big(\frac{2}{3}\Big)^d (2\tau)^d = \mathcal{O}(\tau^d)$$

C.6 Kernel switch between translation invariant RKHS

A direct consequence of Bochner's theorem and Proposition C.1 is that we can explicitly give the embedding constant C_{switch} between the RKHS of two translation invariant kernels. This is precisely the essential supremum of the ratio of their Fourier transform as introduced in Equation (2.6–$\mathbf{H}_\phi$). More generally, considering two continuous and translation-invariant kernels.

$$K_{\mathsf{pivot}}(\boldsymbol{s}, \boldsymbol{t}) = \rho_{\mathsf{pivot}}(\boldsymbol{t} - \boldsymbol{s}) = \frac{1}{(2\pi)^d} \int e^{+\imath \boldsymbol{\omega}^\top (\boldsymbol{t} - \boldsymbol{s})} \, \mathfrak{F}[\rho_{\mathsf{pivot}}](\boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega},$$
$$K_{\mathsf{mod}}(\boldsymbol{s}, \boldsymbol{t}) = \rho_{\mathsf{mod}}(\boldsymbol{t} - \boldsymbol{s}) = \frac{1}{(2\pi)^d} \int e^{+\imath \boldsymbol{\omega}^\top (\boldsymbol{t} - \boldsymbol{s})} \, \mathfrak{F}[\rho_{\mathsf{mod}}](\boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega}$$

Then, assuming the support of $\mathfrak{F}[\rho_{\mathsf{mod}}]$ contains that of $\mathfrak{F}[\rho_{\mathsf{pivot}}]$, we have for any $\eta \in \mathcal{H}_{\mathsf{pivot}}$ that

$$\|\eta\|_{\mathcal{H}_{\mathsf{mod}}}^2 = \frac{1}{(2\pi)^d} \int \frac{\big|\mathfrak{F}[\eta]\big|^2}{\mathfrak{F}[\rho_{\mathsf{mod}}]}(\boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega} = \frac{1}{(2\pi)^d} \int \frac{\mathfrak{F}[\rho_{\mathsf{pivot}}]}{\mathfrak{F}[\rho_{\mathsf{mod}}]}(\boldsymbol{\omega}) \frac{\big|\mathfrak{F}[\eta]\big|^2}{\mathfrak{F}[\rho_{\mathsf{pivot}}]}(\boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega} \leq \operatorname*{ess\,sup}_{\boldsymbol{\omega} \in \operatorname{Supp} \mathfrak{F}[\rho_{\mathsf{pivot}}]} \frac{\mathfrak{F}[\rho_{\mathsf{pivot}}]}{\mathfrak{F}[\rho_{\mathsf{mod}}]}(\boldsymbol{\omega}) \cdot \|\eta\|_{\mathcal{H}_{\mathsf{pivot}}}^2.$$

Thus,

$$\frac{\|\eta\|_{\mathcal{H}_{\mathsf{mod}}}}{\|\eta\|_{\mathcal{H}_{\mathsf{pivot}}}} \leq \operatorname*{ess\,sup}_{\boldsymbol{\omega} \in \operatorname{Supp} \mathfrak{F}[\rho_{\mathsf{pivot}}]} \sqrt{\frac{\mathfrak{F}[\rho_{\mathsf{pivot}}]}{\mathfrak{F}[\rho_{\mathsf{mod}}]}(\boldsymbol{\omega})} =: C_{\mathsf{switch}}. \tag{C.7}$$

Remark C.2 (Kernel switch constant for sinc-4 and mixture)**.** *Applying this results for the sinc-4 pivot $\rho_{\mathsf{pivot}} = \Psi_\tau$ and model kernel $\rho_{\mathsf{mod}} = \lambda_\tau \star \phi \star \check{\phi}$ directly yields the constant C_{switch} of Equation (2.6–$\mathbf{H}_\phi$)*

$$C_{\mathsf{switch}}(\tau, \phi) = \operatorname*{ess\,sup}_{\boldsymbol{\omega} \in [-1/\tau, 1/\tau]^d} \sqrt{\frac{\mathfrak{F}[\Psi_\tau]}{\mathfrak{F}[\lambda_\tau] \big|\mathfrak{F}[\phi]\big|^2}(\boldsymbol{\omega})} \leq \frac{\sup_{\boldsymbol{\omega} \in [-1/\tau, 1/\tau]^d} \sqrt{\mathfrak{F}[\Psi_\tau](\boldsymbol{\omega})}}{\inf_{\boldsymbol{\omega} \in [-1/\tau, 1/\tau]^d} \sqrt{\mathfrak{F}[\lambda_\tau] \big|\mathfrak{F}[\phi]\big|^2(\boldsymbol{\omega})}}, \tag{C.8}$$

since $\operatorname{Supp} \mathfrak{F}[\Psi_\tau] = [-1/\tau, 1/\tau]^d$. Moreover, the numerator $\sup \sqrt{\mathfrak{F}[\Psi_\tau]}$ scales as $\mathcal{O}(\tau^{d/2})$, thus C_{switch} is finite as soon as $\inf_{\boldsymbol{\omega} \in [-1/\tau, 1/\tau]^d} \mathfrak{F}[\lambda_\tau] \big|\mathfrak{F}[\phi]\big|^2(\boldsymbol{\omega}) < +\infty$. In this work, we use the rescaled sinus cardinal as λ_τ, which has a constant *Fourier transform $\mathfrak{F}[\lambda_\tau] = \pi^d \mathbb{1}_{[-1/\tau, 1/\tau]^d}(\boldsymbol{\omega})$. Thus, any template distribution ϕ such that $|\mathfrak{F}[\phi]|(\cdot)$ is lower bounded on $[-1/\tau, 1/\tau]^d$ is amenable to the kernel switch in Proposition 2.1, and the scaling of the constant $C_{\mathsf{switch}}(\tau, \phi)$ depends on the smoothness of $\mathfrak{F}[\phi]$.*