A Dual Certificates Analysis of Compressive Off-the-Grid Recovery

Clarice Poon* C.M.H.S.POON@MATHS.CAM.AC.UK
Centre for Mathematical Sciences
University of Cambridge
Wilberforce Rd, Cambridge CB3 0WA, United Kingdom

Nicolas Keriven* NICOLAS.KERIVEN@ENS.FR
Gabriel Peyré GABRIEL.PEYRE@ENS.FR
Département de Mathématiques et Applications
École Normale Supérieure
45 rue d'Ulm, 75005 Paris, France

Abstract

Many problems in machine learning and imaging can be framed as an infinite dimensional Lasso problem to estimate a sparse measure. This includes for instance regression using a continuously parameterized dictionary, mixture model estimation and super-resolution of images. To make the problem tractable, one typically sketches the observations (often called compressive-sensing in imaging) using randomized projections. In this work, we provide a comprehensive treatment of the recovery performances of this class of approaches, proving that (up to log factors) a number of sketches proportional to the sparsity is enough to identify the sought after measure with robustness to noise. We prove both exact support stability (the number of recovered atoms matches that of the measure of interest) and approximate stability (localization of the atoms) by extending two classical proof techniques (minimal norm dual certificate and golfing scheme certificate).

1. Introduction

1.1 Compressive Recovery of Sparse Measures

In this work, we consider the general problem of estimating an unknown Radon measure $\mu_0 \in \mathcal{M}(\mathcal{X})$ defined over some metric space $\mathcal{X}$ (for instance $\mathcal{X} = \mathbb{R}^d$ for a possibly large d) from a few number m of randomized linear observations $y \in \mathbb{C}^m$

$$\forall\, k = 1, \ldots, m, \quad y_k \stackrel{\text{def.}}{=} \langle \varphi_{\omega_k},\, \mu_0 \rangle + \varepsilon_k \quad \text{where} \quad \langle \varphi,\, \mu \rangle \stackrel{\text{def.}}{=} \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mu(x) \in \mathbb{C}, \tag{1}$$

where $\varepsilon_k \in \mathbb{C}$ accounts for noise or modelling errors, $(\omega_1, \ldots, \omega_m)$ are identically and independently distributed according to some probability distribution $\Lambda(\omega)$ on $\omega \in \Omega$, and for $\omega \in \Omega$, $\varphi_\omega : \mathcal{X} \to \mathbb{C}$ is a continuous function, denoted $\varphi_\omega \in \mathscr{C}(\mathcal{X})$.

Some representative examples of this setting include:

– *Off-the-grid compressed sensing:* off-the-grid compressed sensing, initially introduced in the special case of 1-D Fourier measurements on $\mathcal{X} = \mathbb{T} = \mathbb{R}/\mathbb{Z}$ by (Tang et al., 2013),

*. C. Poon and N. Keriven contributed equally to this work.

corresponds exactly to measurements of the form (1). This is a "continuous" analogous of the celebrated compressed sensing line of works (Candès et al., 2006; Donoho, 2006).

- *Regression using an infinite dimensional dictionary:* given a set of m training samples $(\omega_k, y_k)_{k=1}^m$, one wants to predicts the values $y_k \in \mathbb{R}$ from the features $\omega_k \in \Omega$ using a continuous dictionary of functions $\omega \mapsto \varphi_\omega(x)$ (here $x \in \mathcal{X}$ parameterizes the dictionary), as $y_k \approx \int_{\mathcal{X}} \varphi_{\omega_k}(x)\mathrm{d}\mu(x)$. A typical example, studied for instance by Bach (2017) is the case of neural networks with a single hidden layer made of an infinite number of neurons, where $\Omega = \mathcal{X} = \mathbb{R}^p$ and one uses ridge functions of the form $\varphi_\omega(x) = \psi(\langle x,\, \omega\rangle)$, for instance using the ReLu non-linearity $\psi(u) = \max(u, 0)$.
- *Sketching mixtures:* the goal is estimate a (hopefully sparse) mixture of density probability distributions on some domain $\mathcal{T}$ of the form $\xi(t) = \sum_i a_i \xi_{x_i}(t)$ where the $(\xi_x)_{x\in\mathcal{X}}$ is a family of templates distribution, and $a_i \geqslant 0$, $\sum_i a_i = 1$. Introducing the measure $\mu_0 = \sum_i a_i \delta_{x_i}$, this mixture model is conveniently re-written as $\xi(t) = \int_{\mathcal{X}} \xi_x(t)\mathrm{d}\mu_0(x)$. The most studied example is the mixture of Gaussians, using (in 1-D for simplicity, $\mathcal{T} = \mathbb{R}$) as $\xi_x(t) \propto \sigma^{-1} e^{-\frac{(t-\tau)^2}{2\sigma^2}}$ where the parameter space is the mean and standard deviation $x = (\tau, \sigma) \in \mathcal{X} = \mathbb{R} \times \mathbb{R}^+$. In a typical machine learning scenario, one does not have direct access to ξ but rather to n i.i.d. samples $(t_1, \ldots, t_n) \in \mathcal{T}^n$ drawn from ξ_x. Instead of recording this (possibly huge, specially when $\mathcal{T}$ is high dimensional) set of data, following Gribonval et al. (2017) one computes "online" a small set $y \in \mathbb{C}^m$ of m sketches against sketching functions $\theta_\omega(t)$

$$\forall\, k = 1, \ldots, m, \quad y_k \overset{\text{def.}}{=} \frac{1}{n} \sum_{j=1}^{n} \theta_{\omega_k}(t_j) \approx \int_{\mathcal{T}} \theta_{\omega_k}(t)\xi(t)\mathrm{d}t = \int_{\mathcal{T}} \theta_{\omega_k}(t) \int_{\mathcal{X}} \xi_x(t)\mathrm{d}\mu_0(x)\mathrm{d}t.$$

  These sketches exactly have the form (1) when defining the functions $\varphi_\omega(x) \overset{\text{def.}}{=} \int_{\mathcal{T}} \theta_\omega(t)\xi_x(t)\mathrm{d}t$. A popular set of sketching functions, over $\mathcal{T} = \mathbb{R}^d$ are Fourier atoms $\theta_\omega(t) \overset{\text{def.}}{=} e^{\mathrm{i}\langle \omega,\, t\rangle}$, for which $\varphi_\cdot(x)$ is the characteristic functions of ξ_x, which can generally be computed in closed form.

In all these applications, and much more, one is actually interested in recovering a discrete and s-sparse measure μ_0 of the form $\mu_0 = \sum_{i=1}^s a_i \delta_{x_i}$ where $(x_i, a_i) \in \mathcal{X} \times \mathbb{R}$. Note that in this paper we consider real measures for simplicity, but our results could be extended to complex measures.

An increasingly popular (see Section 1.2) method to estimate such a sparse measure corresponds to solving a infinite-dimensional analogous of the Lasso regression problem

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} \frac{1}{2m} \sum_{k=1}^{m} |\langle \varphi_{\omega_k},\, \mu\rangle - y_k|^2 + \lambda |\mu|(\mathcal{X}). \qquad (\mathcal{P}_\lambda(y))$$

Following De Castro and Gamboa (2012), we call this method the BLASSO (for Beurling-Lasso). Here $|\mu|(\mathcal{X})$ is the so-called total variation of the measure μ, and is defined as

$$|\mu|(\mathcal{X}) \overset{\text{def.}}{=} \sup \left\{ \langle f,\, \mu\rangle \;;\; f \in \mathscr{C}(\mathcal{X}), \|f\|_\infty \leqslant 1 \right\}.$$

Note that on unbounded $\mathcal{X}$, one needs to impose that f vanishes at infinity. If $\mathcal{X} = \{x_i\}_i$ is a finite space, then this corresponds to the classical finite-dimensional Lasso problem (Tibshirani, 1996), because $|\mu|(\mathcal{X}) = \|a\|_1 \stackrel{\text{def.}}{=} \sum_i |a_i|$ where $a_i = \mu(\{x_i\})$. Similarly, if $\mathcal{X}$ is possibly infinite but $\mu = \sum_i a_i \delta_{x_i}$, one also has that $|\mu|(\mathcal{X}) = \|a\|_1$.

1.2 Previous Works

The BLASSO problem ($\mathcal{P}_\lambda(y)$) was initially proposed in (De Castro and Gamboa, 2012), see also (Bredies and Pikkarainen, 2013). The first sharp analysis of the solution of this problem is provided by Candes and Fernandez-Granda (Candès and Fernandez-Granda, 2014) in the case of Fourier measurement on $\mathbb{T}^d$. They show that if the spikes are separated enough, then μ_0 is the unique solution of ($\mathcal{P}_\lambda(y)$) when $\varepsilon = 0$ and $\lambda \to 0$. Robustness to noise under this separation condition is addressed in (Candès and Fernandez-Granda, 2013; Fernandez-Granda, 2013; Azais et al., 2015; Duval and Peyré, 2015), see Section 2 for more details. While this is not the topic of the present paper, note that for positive spikes, the separation condition is in some cases not needed, see for instance (Schiebinger et al., 2015; Denoyelle et al., 2017). These initial works have been extended by Tang et al. (2013) to the case of randomized compressive measurements of the form (1), when using Fourier sketching functions φ_ω.

It is not the focus of this paper, but it is important to note that efficient algorithms have been developed to solve ($\mathcal{P}_\lambda(y)$), among which SDP relaxations for Fourier measurements (Candès and Fernandez-Granda, 2013) and Frank-Wolfe (also known as conditional gradient) schemes (Bredies and Pikkarainen, 2013; Boyd et al., 2017). Note also that while we focus here on variational convex approaches, alternative methods exist, in particular greedy algorithms (Gribonval et al., 2017) and (for Fourier measurements) Prony-type approaches (Schmidt, 1986; Roy and Kailath, 1989). To the best of our knowledge, their theoretical analysis in the presence of noise is more involved, see however (Liao and Fannjiang, 2016) for an analysis of robustness to noise when a minimum separation holds.

1.3 Contributions

The theoretical analysis of the recovery performance of ($\mathcal{P}_\lambda(y)$) is classically achieved by constructing "dual certificates", which are Lagrange multipliers associated to the total variation regularization, and are detailed in Section 2. This paper presents the first comprehensive overview of methods to constructs these certificates for compressive measurement operators.

Our first contribution (Theorem 18) is a detailed study of minimum-separation conditions between spikes to ensure that they can be interpolated by a well-behaved *kernel*, which can then be the basis to construct dual certificates. This is very much inspired by the original work of Candes and Fernandez-Granda (Candès and Fernandez-Granda, 2014), which is extended to general measurement operators. This preliminary study is the key to ensure that one can perform compressive sampling (as stated in the following contributions).

Our second contribution (Theorem 7) shows that, once one has constructed a well-behaved interpolation kernel that has a *random features expansion*, one can only use m proportional to s (up to log factors) such features and obtain a non-degenerate certificate, thus leading to a stable approximate recovery (in the sense of Section 2.1). To the best of our knowledge, there is no similar result in the literature, in particular the proof technics rely on an infinite dimensional "golfing scheme", which up to now has only been used for finite dimensional problems (e.g. on grids) (Gross, 2011; Candes and Plan, 2011).

Our third contribution (Theorem 8) shows that a similar contribution holds, but for a specific certificate η_V (see Section 2.2), which ensures exact recovery of the support. This stronger contribution comes however at the expense of introducing randomized signs for the coefficients $(a_i)_i$, or accepting a number of measurements that is quadratic in the number of Diracs that we want to recover instead of linear. This last theorem is inspired by (Tang et al., 2013), which studies off-the-grid compressed sensing when the sketching functions φ_ω are Fourier atoms on the torus $\mathbb{T}$. Our Theorem 8 extends this by considering quite general functions and by making explicit the dependency on the dimension d.

2. Dual Certificates

Unless stated otherwise, in the rest of the paper $\|\cdot\|$ designates the modulus for complex scalars, the ℓ_2 norm for complex vectors, and for complex symmetric matrices $\|\cdot\| = \left(\|\mathrm{Re}\,(\cdot)\|_{2\to2}^2 + \|\mathrm{Im}\,(\cdot)\|_{2\to2}^2\right)^{\frac{1}{2}}$ where $\|\cdot\|_{2\to2}$ is the spectral norm of matrices. The operator ∇^r is the identity for $r=0$, gradient for $r=1$ and Hessian matrix for $r=2$. For complex functions f, it is just $\nabla^r f = \nabla^r \mathrm{Re}\,(f) + \mathrm{i}\nabla^r \mathrm{Im}\,(f)$.

2.1 Generic Dual Certificate and Approximate Recovery

In this section, we discuss the use of dual certificates in establishing theoretical guarantees for solutions of $(\mathcal{P}_\lambda(y))$. For completeness, we discuss solutions to a more general form

$$\min_{\mu\in\mathcal{M}(X)} \lambda\,|\mu|\,(\mathcal{X}) + \frac{1}{2}\,\|\Phi\mu - y\|_{\mathcal{H}}^2 . \tag{2}$$

when $\Phi : \mathcal{M}(\mathcal{X}) \to \mathcal{H}$ for some Hilbert space $\mathcal{H}$, $y = \Phi\mu_0 + \varepsilon$ with $\|\varepsilon\|_{\mathcal{H}} \leqslant \delta$, so that $(\mathcal{P}_\lambda(y))$ corresponds to the special case $\mathcal{H} = \mathbb{C}^m$.

In order to study the stability to noise ε on the recovery performances of the BLASSO (2), it makes sense to consider the limit $(\varepsilon, \lambda) \to (0,0)$, which leads to consider the constrained problem

$$\min_{\mu}\ \{|\mu|(\mathcal{X})\ ;\ \Phi\mu = \Phi\mu_0\} . \tag{3}$$

The sought after measure μ_0 is solution of (3) if and only if the set $\mathcal{D}$ of Lagrange multipliers, often called "dual certificates" for this problem is non-empty, i.e. $\exists \eta \in \mathcal{D} \overset{\text{def.}}{=} \mathrm{Im}(\Phi^*) \cap \partial|\mu_0|(\mathcal{X})$, where $\partial|\mu_0|(\mathcal{X})$ is the sub-differential at μ_0 of the total variation norm. Assuming μ_0 is a sparse discrete measure of the form $\sum_i a_i\delta_{x_i}$, the set of dual certificates reads

$$\mathcal{D} = \{\eta = \Phi^* p\ ;\ \|\eta\|_\infty \leqslant 1, \forall\, i, \eta(x_i) = \mathrm{sign}(a_i)\} .$$

Constructing such a dual certificate thus amounts to solving an interpolation problem, where the interpolation function η should be bounded by 1 in magnitude.

Robustness from the existence of a nondegenerate dual certificate. Following Burger and Osher (2004), it is known that the existence of a dual certificate $\eta = \Phi^* p \in \mathcal{D}$ implies that solutions to (2) are stable with respect to the Bregman "distance" associated to $|.|(\mathcal{X})$. A direct consequence is that if $\lambda \sim \delta$, then we have the linear noise scaling $||\hat{\mu}|(\mathcal{X}) - |\mu_0|(\mathcal{X})| = \mathcal{O}(\delta + \delta \|p\|_{\mathcal{H}})$.

In order to guarantee stronger and more refined error bounds, such as localization of the recovered measure around true support points, and also the stability to model error, it is natural to impose additional control on how quickly a dual certificate decays away close to its saturation points. One such condition is given in the following result, which is a refinement of (Fernandez-Granda, 2013, Lemma 2.1) and (Azais et al., 2015, Theorem 2.1). In the following, we consider the recovery of an approximately sparse measure $\mu_0 = \sum_{i=1}^s a_i \delta_{x_i} + \tilde{\mu}_0$ for some $\tilde{\mu}_0 \in \mathcal{M}(X)$ such that $\tilde{\mu}_0 \perp \sum_{i=1}^s a_i \delta_{x_i}$, under the existence of a nondegenerate dual certificate for $\sum_{i=1}^s a_i \delta_{x_i}$. Given $\lambda > 0$, $\delta > 0$ and $p \in \mathcal{H}$, let $C(\lambda, \delta, v) \stackrel{\text{def.}}{=} 2\delta v + 2\lambda v^2 + \frac{\delta^2}{2\lambda}$.

Theorem 1 *For each $i = 1, \ldots, s$, let $\mathcal{X}_i^{\text{near}} \subset \mathcal{X}$ be a neighbourhood around the point x_i, and let $\mathcal{X}^{\text{far}} = \mathcal{X} \setminus \bigcup_{i=1}^s \mathcal{X}_i^{\text{near}}$. Suppose that there exists $C_a, C_b > 0$ and $\eta = \Phi^* p$ which satisfies the following conditions with $\sigma = (\text{sign}(a_i))_{i=1}^s$:*

(i) $\eta(x_i) = \sigma_i$ *for all* $i = 1, \ldots, s$,

(ii) $|\eta(x)| \leqslant 1 - C_a \|x - x_i\|^2$ *for all* $x \in \mathcal{X}_i^{\text{near}}$,

(iii) $|\eta(x)| < 1 - C_b$ *for all* $x \in \mathcal{X}^{\text{far}}$,

If $\hat{\mu}$ is a minimizer of (2), then

$$C_b |\hat{\mu} - \mu_0| (\mathcal{X}^{\text{far}}) + C_a \sum_{i=1}^s \int_{\mathcal{X}_i^{\text{near}}} \|x - x_i\|^2 \, \mathrm{d}\,|\nu|(x) \leqslant 2 |\tilde{\mu}_0|(\mathcal{X}) + C(\lambda, \delta, \|p\|_{\mathcal{H}}). \tag{4}$$

Suppose in addition that there exists $\eta_{j,\ell} = \Phi^* p_j$ for $j = 1, \ldots, s$ and $\ell = 1, 2$ such that $\eta_{j,1}$ satisfies (i), (ii), (iii) with $\sigma_i = 1$ for all i and $\eta_{j,2}$ satisfies (i), (ii), (iii) with $\sigma_j = 1$ and $\sigma_i = -1$ for all $i \neq j$, then

$$\forall\, j = 1, \ldots, s, \quad \Big| \int_{\mathcal{X}_j^{\text{near}}} \mathrm{d}(\hat{\mu} - \mu_0)(x) \Big| \leqslant 2 |\tilde{\mu}_0|(\mathcal{X}) + C(\lambda, \delta, \|p_j\|_{\mathcal{H}} + \|p\|_{\mathcal{H}}). \tag{5}$$

Remark 2 – *Note that if $\hat{\mu} = \sum_{j=1}^M \hat{a}_j \delta_{\hat{x}_j}$ and $\mu_0 = \sum_{i=1}^s a_i \delta_{x_i}$, then denoting $\Delta_i = \{j \,;\, \hat{x}_j \in \mathcal{X}_i^{\text{near}}\}$, the second term in (4) implies that $C_a \sum_{i=1}^s \sum_{\ell \in \Delta_i} \|\hat{x}_\ell - x_i\|^2 |\hat{a}_\ell| \leqslant C(\lambda, \delta, p)$, which suggests that spikes of "large" amplitudes cluster tightly around the true support $\{x_i\}_i$.*

– *In contrast to the error bound (4), the guarantee (5) is a local result and essentially says that the recovered mass within each neighbourhood of x_i corresponds roughly to the true mass a_i. This result requires the existence of $2N$ additional certificates, which naturally led to the localized error bound, since $\frac{1}{2}(\eta_{i,1} + \eta_{i,2})$ is a certificate which saturates only at x_i.*

2.2 Minimal Norm Certificate and Exact Support Recovery

In order to obtain sharper recovery property, it is necessary to look for more specific dual certificates. As exposed in (Duval and Peyré, 2015), to obtain exact support recovery (i.e.

for the solution of $(\mathcal{P}_\lambda(y))$ to have the correct number s of Diracs), one needs to consider the minimal norm certificate

$$\eta_0 \stackrel{\text{def.}}{=} \Phi^* p_0 \quad \text{where} \quad p_0 \stackrel{\text{def.}}{=} \underset{p \in \mathbb{C}^m}{\operatorname{argmin}} \; \{\|p\| \; ; \; \Phi^* p \in \partial|\mu_0|(\mathcal{X})\} \,.$$

This certificate is usually hard to compute, so that the way to analyze theoretically the problem is to introduce a proxy which can be computed in closed form by solving a linear system associated to the following least square

$$\eta_V \stackrel{\text{def.}}{=} \Phi^* p_V \quad \text{where} \quad p_V \stackrel{\text{def.}}{=} \underset{p \in \mathbb{C}^m}{\operatorname{argmin}} \; \{\|p\| \; ; \; \forall i, (\Phi^* p)(x_i) = \operatorname{sign}(a_i), \nabla(\Phi^* p)(x_i) = 0_d\} \,. \tag{6}$$

In the case where η_V is a valid certificate, so that $\|\eta_V\|_\infty \leqslant 1$ holds, then $\eta_V = \eta_0$. This is handy because η_V is simply expressed as $\eta_V(x) = \sum_i \alpha_i K(x_i, x) + \langle \beta_i, \, \nabla_1 K(x_i, x)\rangle$ where ∇_1 is the derivative with respect to the first variable, and $(\alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}^d)$ are $s(d+1)$ coefficients which solve a linear system of $s(d+1)$ equations that only depend on the empirical *covariance* as $K(x, x') \stackrel{\text{def.}}{=} \frac{1}{m} \sum_{k=1}^m \operatorname{Re}\left(\bar{\varphi}_{\omega_k}(x) \varphi_{\omega_k}(x')\right)$. The following theorem shows that controlling this minimum certificate ensure exact support recovery with a linear convergence rate on the positions and amplitudes.

A key assumption of the following theorem is that η_V is nondegenerate, that is

$$\forall\, x \notin \{x_1, \ldots, x_s\}, \quad |\eta(x)| < 1 \quad \text{and} \quad \forall\, i = 1, \ldots, s, \quad \operatorname{sign}(a_i) \partial^2 \eta(x_i) \succ 0 \qquad (\mathrm{ND}(\eta))$$

where $\partial^2 \eta(x)$ is the Hessian and $A \succ 0$ means that A is a symmetric positive definite matrix. Note that a nondegenerate certificate would automatically satisfy the conditions of Theorem 1.

In (Duval and Peyré, 2015), the authors show that if $\mathrm{ND}(\eta_V)$ holds, then for $\|\varepsilon\|$ sufficiently small and $(\|\varepsilon\|/\lambda) = O(1)$ the solution of $(\mathcal{P}_\lambda(y))$ is unique and can be written as $\sum_i a_{i,\lambda} \delta_{x_{i,\lambda}}$ where $\|x - x_\lambda\| + \|a - a_\lambda\| = O(\|\varepsilon\|)$. This result is somewhat "stronger" than the previous guarantees in the sense that, for sufficiently small noise, the recovered measure *has exactly s spikes*, whose positions converge to the true ones when the noise goes to 0. However, it assumes the existence of an s-sparse true measure μ_0 and does not allow for inexact sparsity.

3. Main Contributions

3.1 Acceptable kernels

The deterministic limit as $m \to +\infty$ of the (random) empirical covariance is denoted $\overline{K}$, and is of the form

$$\overline{K}(x, x') \stackrel{\text{def.}}{=} \int_\Omega \operatorname{Re}\left(\bar{\varphi}_\omega(x) \varphi_\omega(x')\right) \mathrm{d}\Lambda(\omega), \tag{7}$$

where the convergence $K \to \overline{K}$ holds in probability and almost surely under moment condition on ρ. Note that many covariance kernels can be written under the form (7). By Bochner's theorem, this includes all translation-invariant kernels, for which possible features are $\varphi_\omega(x) = e^{\mathrm{i}\omega^\top x}$.

Our strategy consists in studying the properties of the limit covariance $\overline{K}$, then sample ω_i i.i.d. from Λ, and bound the deviation from the limit case. Our analysis of the function $\overline{K}$ is centered on a *separation condition*, and on the fact that, for properly decreasing kernels, sufficiently separated Diracs have a negligible influence on each other.We therefore introduce a norm $\|\cdot\|_{\text{sep.}}$ to measure the separation of Diracs. Depending on the kernel, carefully choosing this norm will lead to sharper estimates (see our examples below). We introduce the notion of *acceptable kernel* of which we give a summarized description below, and full details in Appendix E.

Definition 3 (Acceptable kernel) *We say that $\overline{K}$ is an acceptable kernel for a maximum number of Diracs $s_{\max} \in \mathbb{N}_+$, constants $\varepsilon_{near}, \Delta, \varepsilon_\eta, \lambda_\eta > 0$ and norm $\|\cdot\|_{sep.}$ if it satisfies the bounds in Table 1 and equations* (31) *to* (34)*, in Appendix E.*

This definition basically states that for $\|x - x'\|_{\text{sep.}} \leqslant \varepsilon_{\text{near}}$, the second derivative of the kernel must not cancel, and for $\|x - x'\|_{\text{sep.}} \geqslant \Delta$, the kernel and all its derivatives must be sufficiently small. Equations (31) to (34) are then used in Theorem 18 in Appendix E, which is of independent interest. It proves that at most $s_{\max}$ signs of Δ-separated Diracs at x_i can be interpolated with a function $\eta \in \text{span}\{\overline{K}(x_i, \cdot),\ \partial_j \overline{K}(x_i, \cdot)\}_{i,j}$ that satisfies $\|\eta\|_\infty \leqslant 1$, with a curvature of at least λ_η when $\varepsilon_{\text{near}}$ close to the x_i and an amplitude of at most $1 - \varepsilon_\eta$ otherwise.

Many usual kernels are acceptable kernels. We limit ourselves to two examples for brevity: the Fejér kernel, which is the kernel usually considered for ideal low-pass filter on the torus $\mathbb{T}^d$, and the Gaussian kernel. The proofs of the following proposition are in Appendix I and H.

Proposition 4 (Multi-dimensional Fejér kernel) *Consider the multidimensional Fejér kernel*

$$\overline{K}(x, x') = \prod_{i=1}^{d} \kappa(x_i - x_i') \quad \textit{where} \quad \kappa(t) = \left(\frac{\sin\left(\left(\frac{f_c}{2} + 1\right)\pi t\right)}{\left(\frac{f_c}{2} + 1\right)\sin(\pi t)} \right)^4$$

on $\mathcal{X} = \mathbb{T}^d$. Take $\varepsilon_{near} = \frac{0.1}{\sqrt{d} f_c}$ and the minimal separation as $\Delta = 5 f_c^{-1} \sqrt{d} \sqrt[4]{s_{\max}}$ for the infinity norm $\|\cdot\|_{sep.} = \|\cdot\|_\infty$, assume f_c is large enough so that $\Delta \leqslant \frac{5}{128}$ for simplicity. Then the kernel $\overline{K}$ is acceptable with $\varepsilon_\eta \geqslant 0.0056/d$ and $\lambda_\eta \geqslant 0.0318 f_c^2$

Proposition 5 (Gaussian kernel) *Consider the Gaussian kernel $\overline{K}(x, y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right)$ on $\mathcal{X} = \mathbb{R}^d$. Take $\varepsilon_{near} = \sigma/\sqrt{2}$ and the minimal separation as $\Delta = \sigma\sqrt{10\log(s_{\max}) + 4\log(d) + 24}$ for the Euclidean norm $\|\cdot\|_{sep.} = \|\cdot\|_2$. Then the kernel $\overline{K}$ is acceptable with $\varepsilon_\eta \geqslant 0.1712$ and $\lambda_\eta \geqslant 0.0800/\sigma^2$*

Remark 6 (Summability) *The separation Δ depends on the maximum number of Diracs that we authorize $s_{\max}$, which is not the case for traditional low dimensional super-resolution (Candès and Fernandez-Granda, 2014). Indeed, in the proof of Theorem 18, we have to bound $\mathcal{S}_s = \sum_{i=1}^{s} \left|\overline{K}(x_0, x_i)\right|$, where the x_i's are Δ-separated Diracs. In dimension one, in the worst case, there is at most 2 Diracs that are at distance Δ from x_0, then 2 Diracs at* *distance 2Δ, and so on. If $|K(x,x')| \leqslant |x-x'|^{-2}$, one can bound $\mathcal{S}_s < \mathcal{S}_\infty \lesssim \frac{1}{\Delta^2}$, and the result does not depend on s. In the multidimensional case however, there is* an exponential *(in d) number of Diracs that can be at distance Δ from x_0, and so on. Applying the same strategy, we get $\mathcal{S}_s < \mathcal{S}_\infty \lesssim \frac{C^d}{\Delta^2}$, which would require a separation that in exponential in d, which is unacceptable. We therefore choosed to let the bound depend on s.*

3.2 Main results

In this section, we assume that $\mathcal{X}$ is compact. We suppose that we have an acceptable kernel $\overline{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. For a given set $x_1, \ldots, x_s$, we define $\mathcal{X}_j^{\text{near}} = \left\{x \;;\; \|x - x_j\|_{\text{sep.}} \leqslant \varepsilon_{\text{near}}\right\}$ and let $\mathcal{X}^{\text{near}}$ and $\mathcal{X}^{\text{far}}$ be a partition of the domain as in Theorem 1.

We recall that we consider covariance kernels $\overline{K}$ that can be written as (7), and we assume that the features $\varphi_\omega(x)$ are in $\mathcal{C}^2(X)$, with uniformly bounded derivatives, and a Lipschitz second derivative. Namely, for $r \in \{0,1,2\}$, we denote by L_r positive constants such that:

$$\sup_{\omega\in\Omega}\sup_{x\in\mathcal{X}} \|\nabla^r \varphi_\omega(x)\| \leqslant L_r \tag{8}$$

Furthermore, we assume

$$\sup_{\omega\in\Omega} \left\|\nabla^2\varphi_\omega(x) - \nabla^2\varphi_\omega(x')\right\| \leqslant L_3 \left\|x - x'\right\|$$

Finally, we define $L_{01} \stackrel{\text{def.}}{=} \sqrt{L_0^2 + L_1^2/v}$ (where v is defined in Table 1).

Our goal is to show that, by sampling a reduced number of parameters $\omega_1, \ldots, \omega_m$ iid from Λ, we obtain functions $(\varphi_{\omega_k})_k$ such that there exists a dual certificate with high probability. This is done in the following Theorem.

Theorem 7 *Assume the number of measurements satisfies*

$$m \gtrsim s \cdot \left[d^2 \left(\tfrac{L_{01}B_0}{h}\right)^2 \log(sd)\log\left(\tfrac{sd}{\rho}\right) + \sum_{i\in\{0,2\}} \bar{\mathcal{L}}_i \left(\log\left(\tfrac{(sN_i)^d}{\rho}\right) + \log\left(\tfrac{1}{\rho}\right)\tfrac{\log((sN_i)^d)}{\log(sd)}\right)\right] \tag{9}$$

where $N_0 = 1 + \frac{dL_{01}L_1B_\mathcal{X}}{\varepsilon_\eta}$, $N_2 = 1 + \frac{dL_{01}L_3\varepsilon_{near}}{\lambda_\eta}$, $h = \min\{\varepsilon_\eta, \frac{B_0\lambda_\eta}{B_2}, 1\}$ and, denoting by $\alpha_0 = \varepsilon_\eta$ and $\alpha_2 = \lambda_\eta$ and defining constants B_0 and B_2 that only depends on $\overline{K}$ (see Appendix E), we have $\bar{\mathcal{L}}_i = \left(\max\{\frac{dL_i^2}{\alpha_i^2}, \frac{\sqrt{d}L_iL_{01}}{\alpha_i}\} + \max\{\frac{L_i^2}{B_i^2}, \frac{L_iL_{01}}{B_i}\}\right)\log\left(\frac{L_0}{\varepsilon_\eta} + \frac{L_2}{\lambda_\eta}\right)$.

Then with probability at least $1-\rho$, there exists a dual certificate $\eta \in \operatorname{Im}(\Phi^\star)$ that satisfies the conditions of Theorem 1 with $C_a = \lambda_\eta$ and $C_b = \varepsilon_\eta$.

The proof of Theorem 7 constructs explicitly a dual certificate using an infinite dimensional extension of the so-called golfing scheme and therefore, its existence directly provides recovery and stability guarantees described in Section 2.1. We mention also that as discussed in Remark 2, in order to leverage the second result of Theorem 1, one would need to construct $\mathcal{O}(s)$ certificates with different sign patterns. With a simple union bound, this comes only at a price of $\log s$ in the number of measurements (9).

Note however that the constructed certificate is not necessarily the minimal norm certificate and hence, one cannot guarantee the stronger property of *support stability.* To address this issue, for the following Theorem, we proceed in a different manner which leads to the more pessimistic sampling bound which is quadratic in s. Similarly to (Tang et al., 2013) this can be solved under the somewhat unrealistic assumption that the signs of the a_i are random.

Theorem 8 *Assume the number of measurements satisfies*

$$m \gtrsim s \cdot \left[s\left(\mathcal{L}_0 + \mathcal{L}_2\right) \log\left(\tfrac{sd}{\rho}\right) + \sum_{i \in \{0,2\}} \bar{\mathcal{L}}_i \log\left(\tfrac{(sN_i)^d}{\rho}\right) \right] \tag{10}$$

where $\mathcal{L}_i \stackrel{\text{def.}}{=} L_{01}^2 \frac{B_i^2}{\alpha_i^2}$ *and* $\bar{\mathcal{L}}_i \stackrel{\text{def.}}{=} \frac{L_i}{\alpha_i} \cdot \left(\frac{L_i}{\alpha_i} + L_{01}\right)$*. Then, with probability at least* $1-\rho$*, the vanishing derivative pre-certificate* η_V *is non-degenerate.*

If the signs a_i *are drawn iid from a Rademacher distribution, the number of measurements* (10) *can be replaced by*

$$m \gtrsim s \cdot \log\left(\tfrac{sd}{\rho}\right) \cdot \left[\left(\mathcal{L}_0 + \bar{\mathcal{L}}_0\right) \log\left(\tfrac{(sN_0)^d}{\rho}\right) + \left(d\mathcal{L}_2 + \bar{\mathcal{L}}_2\right)\left(\tfrac{(sN_2)^d}{\rho}\right) \right] \tag{11}$$

to obtain the same result.

Polynomial dependency in d. In our examples of Section 3.4, often the bound on m are polynomial in the dimension d, even though, for instance, the bound (10) is at first glance linear in d. This will often come from the various Lipschitz constants L_i and may be refined in the future. In this paper, we focused on the rate in s.

3.3 Sketch of proof

Given φ_{ω_k}, a_i and x_i, define:

$$\gamma(\omega) \stackrel{\text{def.}}{=} D\left(\overline{\varphi_\omega(x_1)}, ..., \overline{\varphi_\omega(x_s)}, \nabla\overline{\varphi_\omega(x_1)}^\top, ..., \nabla\overline{\varphi_\omega(x_s)}^\top\right)^\top \in \mathbb{C}^{s(d+1)}$$

where, denoting $v_{il} = 1/\sqrt{\partial_{1,i}\partial_{2,i}K(x_l, x_l)}$, $D = \operatorname{diag}(1, \ldots, 1, v_{11}, \ldots, v_{d1}, v_{12}, \ldots\ldots, v_{dk})$ is a diagonal matrix whose first s elements are 1's, which is here for normalization purpose. Let $\Upsilon \in \mathbb{R}^{s(d+1)\times s(d+1)}$, $\mathbf{f} : \mathcal{X} \to \mathbb{R}^{s(d+1)}$, and $\mathbf{u}_s \in \mathbb{R}^{s(d+1)}$ be defined by

$$\Upsilon \stackrel{\text{def.}}{=} \frac{1}{m}\sum_{k=1}^m \operatorname{Re}\left(\gamma(\omega_k)\gamma(\omega_k)^H\right), \quad \mathbf{f}(x) \stackrel{\text{def.}}{=} \frac{1}{m}\sum_{k=1}^m \operatorname{Re}\left(\gamma(\omega_k)\varphi_{\omega_k}(x)\right), \quad \mathbf{u}_s = \begin{pmatrix} \operatorname{sign}(a) \\ 0_{sd} \end{pmatrix}$$

Note that Υ and $\mathbf{f}$ are respectively the empirical versions of

$$\overline{\Upsilon} \stackrel{\text{def.}}{=} \mathbb{E}_\omega \operatorname{Re}\left(\gamma(\omega)\gamma(\omega)^H\right) \in \mathbb{R}^{s(d+1)\times s(d+1)}, \qquad \overline{\mathbf{f}}(x) \quad \stackrel{\text{def.}}{=} \mathbb{E}_\omega \operatorname{Re}\left(\gamma(\omega)\varphi_\omega(x)\right) \in \mathbb{R}^{s(d+1)}.$$

We first comment briefly on the proof of Theorem 8 which closely follows (Tang et al., 2013), before presenting a sketch of the proof of Theorem 7. First observe that the vanishing derivative pre-certificate has a closed-form:

$$\eta_V(x) = \mathbf{u}_s^\top \Upsilon^{-1}\mathbf{f}(x).$$

Our goal is to prove that this function is non-degenerate. To do this, we study the "limit" version $\overline{\eta}_V(x) = \mathbf{u}_s^\top \overline{\Upsilon}^{-1}\overline{\mathbf{f}}(x)$ when $m \to \infty$, where we observe that $\overline{\Upsilon}$ and $\overline{\mathbf{f}}$ *only depend on the limit kernel* $\overline{K}$. Hence we proceed in two steps:

1. We first show that if the kernel $\overline{K}$ is acceptable then $\overline{\eta}_V$ is nondegenerate. This is done in Appendix E with a dedicated Theorem.

2. Using Bernstein concentration inequalities, we prove that Υ and $\mathbf{f}$ are close to their limit versions, and thereby deduce that η_V and $\overline{\eta}_V$ and their second derivative are close on a properly defined grid. By using covering arguments, we conclude that η_V must be close to $\overline{\eta}_V$ on the entire domain $\mathcal{X}$ and is therefore nondegenerate. See Appendix F for details.

The difficulty with this approach is that the control in the distance between η_V and $\overline{\eta}_V$ in Step 2 involves $\|\mathbf{u}_s\|_2 = \sqrt{s}$, and this term is the source of the quadratic bottleneck of s^2 in (10). As shown in (Tang et al., 2013) and also in Theorem 8, this bottleneck can be alleviated under the somewhat unrealistic assumption that the signs are drawn iid from a Rademacher distribution, so that for a fixed vector v there is little chance of having the worst case $\left|\mathrm{sign}(a)^\top v\right| = \sqrt{s}\,\|v\|_2$.

The golfing scheme. To circumvent this quadratic bottleneck without imposing the the random signs assumption, we develop a proof based on an infinite dimensional generalization of the so-called golfing scheme (Gross, 2011; Candes and Plan, 2011) for Theorem 7. We simply outline the key ideas here, a detailed proof can be found in Appendix G.

Step I: Constructing an approximate dual certificate. Divide the indices $\{1,\ldots,m\}$ into L blocks B_l of size m_l for $l = 1,\ldots,L$. Let Υ_l and $\mathbf{f}_l$ be the respective empirical versions of $\overline{\Upsilon}$ and $\overline{\mathbf{f}}$ with respect to the indices in B_l. For $j = 1,\ldots,L$, define $q_j \in \mathbb{C}^{(d+1)s}$, $\eta^j \in \mathscr{C}(X)$ by $q_0 = \mathbf{u}_s$, $\eta^0 = 0$, and for $j \geqslant 1$:

$$\eta^j = \sum_{i=1}^{j-1} (\bar{q}_{i-1})^T \overline{\mathbf{f}}_i(\cdot) \quad \text{and} \quad q_j = \mathbf{u}_s - \sum_{i=1}^{j-1} \Upsilon_j \bar{q}_{i-1}, \quad \text{where} \quad \bar{q}_i = \Upsilon^{-1} q_i.$$

The idea is that by choosing the m_i's appropriately, one can ensure that $\eta^{\mathrm{app}} \stackrel{\text{def.}}{=} \eta^L$ is approximately a nondegenerate dual certificate when evaluated on a fine grid. More precisely, suppose that for each i, the following conditions hold for constants $c_i, t_i, b_i > 0$:

(I) $\left\|(\mathrm{Id} - \Upsilon_i \overline{\Upsilon}^{-1}) q_{i-1}\right\|_\infty \leqslant c_i \|q_{i-1}\|_\infty$

(II) $\forall x \in \mathcal{X}^{\mathrm{far}}_{\mathrm{grid}}$, $\left|\bar{q}_{i-1}^\top \mathbf{f}_i(x)\right| \leqslant t_i \|q_{i-1}\|_\infty$

(III) $\forall x \in \mathcal{X}^{\mathrm{near}}_{\mathrm{grid}}$, $\left\|\nabla^2 \left(\bar{q}_{i-1}^\top \mathbf{f}_i(x)\right)\right\| \leqslant b_i \|q_{i-1}\|_\infty$

where $\mathcal{X}^{\mathrm{far}}_{\mathrm{grid}} \subset \mathcal{X}^{\mathrm{far}}$ and $\mathcal{X}^{\mathrm{near}}_{\mathrm{grid}}(j) \subset \mathcal{X}^{\mathrm{near}}_j$, $\mathcal{X}^{\mathrm{near}}_{\mathrm{grid}} = \bigcup_j \mathcal{X}^{\mathrm{near}}_{\mathrm{grid}}(j)$ are appropriately dense finite grids, and if $i = 1$, then we replace (III) by $\mathrm{sign}(a_j)\nabla^2 \left((\overline{\Upsilon}^{-1}\mathbf{u}_s)^\top \mathbf{f}_1(x)\right) \preceq -b_1 \mathrm{Id}$ for all $x \in \mathcal{X}^{\mathrm{near}}_{\mathrm{grid}}(j)$, $j = 1,\ldots,s$. Then, for appropriately chosen c_i, t_i, b_i, one can verify that η^{app} satisfies for some appropriate constant $c > 0$,

- $\|D\Psi\eta^{\mathrm{app}} - \mathbf{u}_s\|_2 \leqslant c \min\{\varepsilon_\eta,\ \lambda_\eta\}$,
- for all $x \in \mathcal{X}^{\mathrm{far}}_{\mathrm{grid}}$, $|\eta(x)| \leqslant 1 - \frac{3\varepsilon_\eta}{8}$,

- for all $x \in \mathcal{X}^{\text{near}}_{\text{grid}}(j)$, $\text{sign}(a_j)\nabla^2\eta(x) \preceq -\frac{3\lambda_\eta}{8}\text{Id}$.

where, given $f \in \mathscr{C}^1(\mathcal{X})$, $\Psi f = \left[f(x_1),\ldots,f(x_s),\nabla f(x_1)^\top,\ldots,\nabla f(x_s)^\top\right]^\top$. In fact, the constant c can be made arbitrarily small at the cost of increasing the number of samples by a factor of $\log(1/c)$. Indeed, we have the following relation between q_i and η^i: $q_i = \mathbf{u}_s - D\Psi(\eta^i)$. Therefore q_i represents the *error* between the value of η^i at each x_l and $\text{sign}(a_l)$ and the deviation of its gradient from zero. At each golfing step, it is easy to check that by definition we have $q_j = \left(\text{Id} - \Upsilon_j\overline{\Upsilon}^{-1}\right) q_{j-1}$ which leads to a geometric progression of the error.

So, to prove the existence of an approximate dual certificate, it is sufficient to bound for each i, the probability that conditions (I), (II) and (III) are satisfied. We remark that following (Gross, 2011), in order to obtain sharper sampling estimates, in the proof, we actually carry out a more refined construction where we impose only that conditions (I)-(III) hold for a sufficiently large subset of the indices $\{1,\ldots,L\}$, and discarding the sample draws B_i which fail these conditions.

Step II: correcting the approximate dual certificate. Once η^{app} has been constructed, we can then construct a function η which exactly satisfy the nondegeneracy conditions when evaluated on a fine grid. Indeed, by defining $\eta^e = (\Upsilon^{-1}e)^\top \mathbf{f}(\cdot)$ where $e = D\Psi\eta^{\text{app}} - \mathbf{u}_s$, the approximate certificate can be "fixed" by putting the final certificate as $\eta = \eta^{\text{app}} - \eta_e$. One can verify that

$$\Psi\eta = \Psi(\eta^{\text{app}} - \eta^e) = \mathbf{u}_s$$

and thus we have indeed $\eta(x_i) = \text{sign}(a_i)$ and $\nabla\eta(x_i) = 0$. Furthermore, since one can show that $\left\|\Upsilon^{-1}\right\| \lesssim 1$ with high probability, the control on $\|e\|$ from the first step yields that for all $x \in \mathcal{X}^{\text{far}}_{\text{grid}}$, $|\eta(x)| \leqslant 1 - \frac{\varepsilon_\eta}{4}$ and for all $x \in \mathcal{X}^{\text{near}}_{\text{grid}}$, $\nabla^2\eta(x) \preceq -\frac{\lambda_\eta}{4}\text{Id}$. Therefore, we have constructed a function η which is nondegenerate when evaluated on a fine grid, and by covering arguments, these nondegeneracy properties can be extended to the entire domain $\mathcal{X}$.

3.4 Examples

We make precise our theoretical results for different kernels in this section. Numerical illustrations of their associated certificates can be found in Section A.

Discrete Fourier: the Fejér kernel. We recall that the multivariate Fejér kernel is written as $\overline{K}(x,x') = \prod_{i=1}^d \kappa(x_i - x_i')$ on $\mathcal{X} = \mathbb{T}^d$, where $\kappa(t) = \left(\frac{\sin\left(\left(\frac{f_c}{2}+1\right)\pi t\right)}{\left(\frac{f_c}{2}+1\right)\sin(\pi t)}\right)^4 = \sum_{f=-f_c}^{f_c} g(f)e^{\mathrm{i}2\pi f t}$ where $g(f) \geqslant 0$ are such that $\sum_{f=-f_c}^{f_c} g(f) = 1$ (see (Candès and Fernandez-Granda, 2014)). Hence the frequency domain is the discrete domain $\Omega = [\![-f_c\,;\,f_c]\!]^d$, the features are $\varphi_\omega(x) = e^{\mathrm{i}2\pi\omega^\top x}$ and the discrete probability distribution over Ω is $\Lambda(\omega) = \prod_{i=1}^d g(\omega_i)$. We have $L_i = \mathcal{O}\left(d^{\frac{i}{2}} f_c^i\right)$, $B_0 = \mathcal{O}\left(\sqrt{d}\right)$ and $B_2 = \mathcal{O}\left(d f_c^2\right)$ (see Appendix H). Applying our results, the bounds on the number of measurements are:

– the bound (9) reads $m \gtrsim s\left(d^6\log(sd)\log\left(\frac{sd}{\rho}\right) + d^4\log\left(\frac{1}{\rho}\right)\log(df_c)\right)$,

– the bound (10) reads $m \gtrsim sd^2 \left(sd^2 \log\left(\frac{sd}{\rho}\right) + d\log(df_c) + \log\left(\frac{1}{\rho}\right)\right)$,

– the bound (11) reads $m \gtrsim sd^4 \log\left(\frac{sd}{\rho}\right)\left(d\log(df_c) + \log\left(\frac{1}{\rho}\right)\right)$.

In particular, for $d = 1$, (11) recovers the result of (Tang et al., 2013) (with a different separation condition, see Remark 6).

Continuous Fourier: the Gaussian kernel. It is well known (Rahimi and Recht, 2007) that the Gaussian kernel $\overline{K}(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$ can be written with random features $\varphi_\omega(x) = e^{\mathrm{i}\omega^\top x}$ with Gaussian frequency distribution $\Lambda = \mathcal{N}(0, \sigma^{-2})$. However, our result cannot be readily applied in this case, since when ω is unbounded the derivatives of the features φ_ω are not uniformly bounded. A simple way to fix that is by *weighting* the features φ_ω by a function $f(\omega)$, and modifying the distribution Λ so as to keep the same kernel $\overline{K}$. For instance, in the case of the Gaussian kernel, denote by $\gamma_l = \mathbb{E}_{\omega\sim\mathcal{N}(0,\sigma^{-2})} \|\omega\|^l = \mathcal{O}\left(\frac{d^{\frac{l}{2}}}{\sigma^l}\right)$ the moments of the χ-distribution, and $f(\omega) = \frac{1}{2}\sqrt{\sum_{l=0}^3 \frac{\|\omega\|^{2l}}{\gamma_{2l}}}$. Then, it is easy to check that the features $\varphi_\omega(x) \stackrel{\text{def.}}{=} \frac{e^{\mathrm{i}\omega^\top x}}{f(\omega)}$ satisfy all the assumptions required by our analysis with Lipschitz constants $L_l \propto \sqrt{\gamma_{2l}}$, that $\Lambda(\omega) \stackrel{\text{def.}}{=} f(\omega)^2\mathcal{N}(0, \sigma^{-2})$ is a proper probability distribution (since it is positive and sum to one), and that we kept the property (7). For the Gaussian kernel we have $B_0 = \mathcal{O}(1)$ and $B_2 = \mathcal{O}(1/\sigma^2)$ (see Appendix I), and applying our results,

– the bound (9) reads $m \gtrsim sd^2 \left(d\log(sd)\log\left(\frac{sd}{\rho}\right) + \log\left(\frac{B_\mathcal{X}}{\sigma}\right) + d^2\log(sd)\log\left(\frac{1}{\rho}\right)\right)$,

– the bound (10) reads $m \gtrsim s\left(sd\log\left(\frac{sd}{\rho}\right) + d^{\frac{3}{2}}\log\left(\frac{B_\mathcal{X}}{\sigma}\right) + d^3\log(sd) + d^2\log\left(\frac{1}{\rho}\right)\right)$,

– the bound (11) reads $m \gtrsim s\log\left(\frac{sd}{\rho}\right)\left(d^2\log\left(\frac{B_\mathcal{X}}{\rho\sigma}\right) + d^3\log(sd)\right)$.

Mixture Model learning. We now illustrate our framework applied to a simple problem of learning Gaussian Mixture model, with identity covariance for simplicity (we leave for future the treatment of unknown covariances, which seems more involved). We summarize our results in the next proposition, and postpone the proof in Appendix J.

Proposition 9 *Assume data points $t_1, \ldots, t_n$ are drawn according to a mixture of Gaussians $\sum_{i=1}^s a_i\mathcal{N}(x_i, \mathrm{Id})$. Choose any $\sigma_K > 0$, and draw $\omega_1, \ldots, \omega_m$ iid from $\mathcal{N}(0, \sigma_K^2\mathrm{Id})$. Define $M_K = \left(1 + 2\sigma_K^2\right)^{d/2}$. Perform the following*

1. *for $k = 1, \ldots, m$ compute $y_k = \frac{M_K}{n}\sum_{j=1}^n e^{\mathrm{i}\omega_k^\top t_j}$*

2. *solve $(\mathcal{P}_\lambda(y))$ with $\varphi_\omega(x) = M_K e^{\mathrm{i}\omega^\top x} e^{-\frac{\|\omega\|^2}{2}}$.*

where M_K is here for consistency with the previous theorems and has no effect on the BLASSO.

We obtain a problem of the form (1), with a noise vector ε_n, such that with probability at least $1-\rho$ the noise level in Theorem 1 is as $\delta \stackrel{\text{def.}}{=} \left\|\frac{1}{m}\varepsilon_n\right\|_2 \leqslant \frac{M_K}{\sqrt{n}}\left(1 + \sqrt{2\log\left(\frac{2}{\rho}\right)}\right)$.

Assume that the x_i are Δ-separated, with $\Delta \propto (1 + \sigma_K^{-1})\sqrt{\log sd}$. Applying our results:

– *Theorem 7: if*

$$m \gtrsim sd^2 M_K^2 \left(\sigma^2 \log(sd) \log\left(\frac{sd}{\rho}\right) + \log\left(\frac{1}{\rho}\right) \left(\log\left(M_K B_{\mathcal{X}}\right) + \sigma^4 \log\left(sdM_K\right)\right)\right)$$

where $\sigma^2 = 1 + \sigma_K^{-2}$, *then we can apply Theorem 1 with* $C_a = \mathcal{O}(1), C_b = \mathcal{O}\left(1/\sigma^2\right)$ *and* $\varepsilon_{near} = \mathcal{O}(\sigma)$.

– *Theorem 8: if*

$$m \gtrsim sM_K^2 \left(s\sigma^2 \log\left(\frac{sd}{\rho}\right) + d\sigma \log M_K B_{\mathcal{X}} + \sigma^4 \left(d \log\left(sdM_K\right) + \log\left(\frac{1}{\rho}\right)\right)\right)$$

and the number of samples is sufficiently large (ie for small noise, see Section 2.2), the recovered measure has exactly s *components, whose positions and weights converge to the* x_i*'s and* a_i*'s as the number of samples increases.*

In the second case in particular, we obtain a convex optimization problem that is able to *exactly identify* the number of components of a mixture model, which removes the need for prior knowledge usually required by classical methods, even in this simple case. In particular, likelihood-based methods often require knowing the number of components in advance, and exhibit local minima (Jin et al., 2016). As a bonus, the proposed method enjoys the advantage of *sketching* (Gribonval et al., 2017), meaning that the computation of the y_k's can be done in a online or distributed setting, without having to store the whole data.

The user-picked parameter σ_K plays a significant role in the number of measurements m and separation Δ. At one end of the spectrum, choosing $\sigma_K = \mathcal{O}(1)$, both number of measurements m and sample complexity n are in $\mathcal{O}\left(e^d\right)$, but the BLASSO can distinguish Gaussians whose means are only separated by $\Delta = \mathcal{O}\left(\sqrt{\log(sd)}\right)$. At the other end of the spectrum, by choosing $\sigma_K^2 = \frac{1}{d}$, the number of measurements is polynomial in d and the sample complexity is linear in d (through C_b), but the required separation is $\Delta = \mathcal{O}\left(\sqrt{d \log(sd)}\right)$.

For these three examples, we summarize all quantities in Table 2 in Appendix J.

4. Conclusion

It is well known that the existence and properties of dual certificates provide various stability and recovery guarantees. However, there have been few works characterizing the conditions under which such certificates can be constructed in the compressive off-the-grid setting with random sampling. Furthermore, in existing works, optimal sampling bounds are often attained only under the random signs assumption.We address this problem with a comprehensive analysis of the conditions under which a well-behaved dual certificate can be constructed. Our assumptions on the sampling kernel cover many common cases , such as the Féjer kernel and the Gaussain kernel. Furthermore, up to log factors, the number of samples that we require are optimal with respect to the sparsity s of the underlying measure. Candidate kernels may be obtained with other random feature schemes (Vedaldi and Zisserman, 2012) and more general classes of non-linearities (Bach, 2017).

References

Ben Adcock, Anders C. Hansen, Clarice Poon, and Bogdan Roman. Breaking the coherence barrier: a new theory for compressed sensing. *Forum of Mathematics, Sigma*, 5:1–84, 2017. doi: 10.1017/fms.2016.32.

Jean-Marc Azais, Yohann De Castro, and Fabrice Gamboa. Spike detection from inaccurate samplings. *Applied and Computational Harmonic Analysis*, 38(2):177–195, 2015.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

Nicholas Boyd, Geoffrey Schiebinger, and Benjamin Recht. The alternating descent conditional gradient method for sparse inverse problems. *SIAM Journal on Optimization*, 27 (2):616–639, 2017.

Kristian Bredies and Hanna Katriina Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.

Martin Burger and Stanley Osher. Convergence rates of convex variational regularization. *Inverse problems*, 20(5):1411, 2004.

Emmanuel J. Candès and Carlos Fernandez-Granda. Super-resolution from noisy data. *Journal of Fourier Analysis and Applications*, 19(6):1229–1254, 2013. ISSN 1069-5869. doi: 10.1007/s00041-013-9292-3. URL `http://dx.doi.org/10.1007/s00041-013-9292-3`.

Emmanuel J. Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956, 2014. ISSN 1097-0312. doi: 10.1002/cpa.21455. URL `http://dx.doi.org/10.1002/cpa.21455`.

Emmanuel J Candes and Yaniv Plan. A probabilistic and RIPless theory of compressed sensing. *IEEE Transactions on Information Theory*, 57(11):7235–7254, 2011.

Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

Yohann De Castro and Fabrice Gamboa. Exact reconstruction using Beurling minimal extrapolation. *Journal of Mathematical Analysis and applications*, 395(1):336–354, 2012.

Quentin Denoyelle, Vincent Duval, and Gabriel Peyré. Support recovery for sparse super-resolution of positive measures. *to appear in Journal of Fourier Analysis and Applications*, 2017. doi: 10.1007/s00041-016-9502-x. URL `http://arxiv.org/abs/1506.08264`.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4): 1289–1306, 2006.

Vincent Duval and Gabriel Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.

C. Fernandez-Granda. Support detection in super-resolution. *Proc. Proceedings of the 10th International Conference on Sampling Theory and Applications*, pages 145–148, 2013. URL `http://arxiv.org/abs/1302.3921`.

Carlos Fernandez-Granda. Support detection in super-resolution. *arXiv preprint arXiv:1302.3921*, 2013.

Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, and Yann Traonmilin. Compressive statistical learning with random feature moments. *arXiv preprint arXiv:1706.07180*, 2017.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J Wainwright, and Michael I Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems*, pages 4116–4124, 2016.

Wenjing Liao and Albert Fannjiang. MUSIC for single-snapshot spectral estimation: Stability and super-resolution. *Applied and Computational Harmonic Analysis*, 40(1):33–67, 2016.

Ali Rahimi and Benjamin Recht. Random Features for Large Scale Kernel Machines. *Advances in Neural Information Processing Systems (NIPS)*, 2007.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in Neural Information Processing Systems (NIPS)*, 2008.

Richard Roy and Thomas Kailath. ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on acoustics, speech, and signal processing*, 37 (7):984–995, 1989.

Geoffrey Schiebinger, Elina Robeva, and Benjamin Recht. Superresolution without separation. *arXiv preprint arXiv:1506.03144*, 2015.

Ralph Schmidt. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation*, 34(3):276–280, 1986.

Karthik Sridharan. A Gentle Introduction to Concentration Inequalities. Technical report, 2002. URL `http://www.intertwingly.net/stories/2002/03/16/aGentleIntroductionToSoap.html`.

Gongguo Tang, Badri Narayan Bhaskar, Parikshit Shah, and Benjamin Recht. Compressed sensing off the grid. *IEEE transactions on information theory*, 59(11):7465–7490, 2013.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Joel A. Tropp. An introduction to matrix concentration inequalities. (December), 2015. ISSN 1096603X. doi: 10.1016/j.acha.2016.07.005.

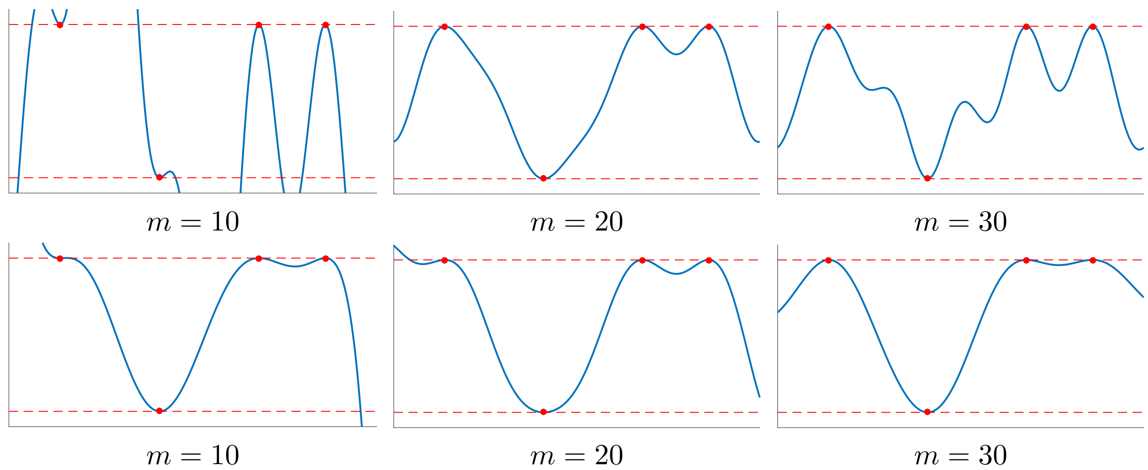

Figure 1: Display of η_V in dimension $d = 1$. The input measures μ_0 has $s = 4$ spikes (3 positive and 1 negative in the center). Top row: squared Fejér kernel with $f_c = 16$, displayed on $[0, 2\pi]$. Bottom row: Gaussian kernel with $\sigma = 0.08$, displayed on $[0, 1]$. The two red dashed lines indicate the $[-1, +1]$ acceptable range. Regions where η_V is not between these lines indicate that it is degenerate (not a valid dual certificate).

Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):480–492, 2012. ISSN 01628828. doi: 10.1109/TPAMI.2011.153.

Appendix A. Numerical Illustrations

Figures 1 and 2 display, in dimensions $d = 1$ and $d = 2$, the influence of the number of m of sketches on the pre-certificate $\eta_V(x)$, defined in (6). It shows the two cases considered in Section 3.4, namely the squared Fejér kernel on $\mathcal{X} = \mathbb{T}^d$ and the Gaussian kernel on $\mathcal{X} = \mathbb{R}^d$. These figures show that for small values of m, η_V oscillates and might fail to have values in the acceptable range $[-1, 1]$. Since in these examples the spikes are well separated, for m large enough, this does not happens and $\eta_V = \eta_0$ is a valid dual certificate.

Appendix B. Proof of Theorem 1

Throughout, we let $T = \{x_1, \ldots, x_s\}$ and given a finite measure ν, let ν_T be the restriction of this measure to the set T.

Proof of (4)**:** From (Burger and Osher, 2004, Thm. 2),

$$\frac{1}{2} \|\Phi\hat{\mu} - y + \lambda p\|_{\mathcal{H}}^2 + \lambda \left(|\hat{\mu}| - |\mu_0| - \langle \eta, \, \hat{\mu} - \mu_0 \rangle\right) \leqslant \frac{1}{2} \left(\delta + \lambda \|p\|_{\mathcal{H}}\right)^2,$$

which implies that

$$\|\Phi\hat{\mu} - \Phi\mu_0\|_{\mathcal{H}} \leqslant 2(\delta + \lambda \|p\|_{\mathcal{H}}),$$

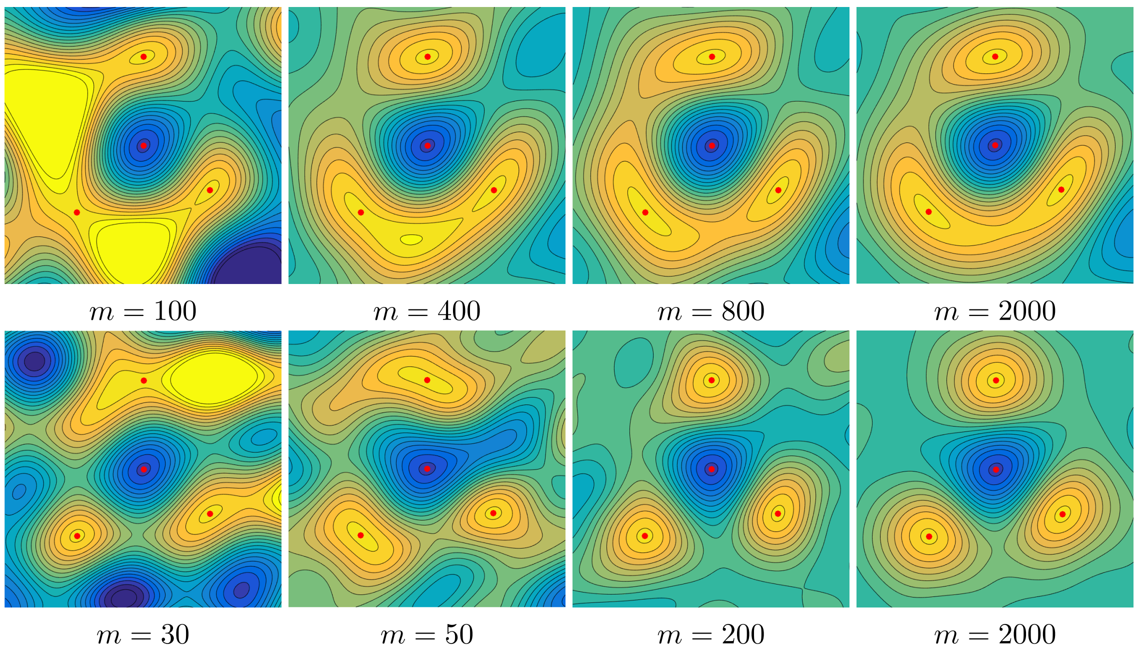

Figure 2: Display of η_V in dimension $d = 2$. The input measures μ_0 has $s = 4$ spikes (3 positive and 1 negative in the center). Top row: squared Fejér kernel with $f_c = 16$, displayed on $[0, \pi]^2$. Bottom row: Gaussian kernel with $\sigma = 0.13$, displayed on $[0, 1]^2$. The color code for admissible values ranges from light blue for -1 to orange for $+1$. Dark blue and yellow indicates regions where η_V is outside the range $[-1, 1]$, and is thus degenerate (not a valid dual certificate).

and

$$||\hat{\mu}|(\mathcal{X}) - |\mu_0|(\mathcal{X})| \leqslant \frac{1}{2\lambda}(\delta + \lambda \|p\|_{\mathcal{H}})^2 + 2\|p\|_{\mathcal{H}}(\delta + \lambda\|p\|_{\mathcal{H}}) = \frac{5\lambda\|p\|_{\mathcal{H}}^2}{2} + 3\|p\|_{\mathcal{H}}\delta + \frac{\delta^2}{2\lambda}.$$

Therefore, by letting $\nu = \hat{\mu} - \mu_0$,

$$|\langle \eta, \nu \rangle| = |\langle p, \Phi\nu \rangle| \leqslant 2\delta\|p\|_{\mathcal{H}} + 2\lambda\|p\|_{\mathcal{H}}^2 \tag{12}$$

Observe that

$$\begin{aligned}\left|\int \eta(x)\mathrm{d}\nu_T(x)\right| &\leqslant \left|\int \eta(x)\mathrm{d}\nu(x)\right| + \left|\int_{T^c} \eta(x)\mathrm{d}\nu(x)\right| \\ &\leqslant 2\delta\|p\|_{\mathcal{H}} + 2\lambda\|p\|_{\mathcal{H}}^2 + \sum_{j=1}^{s}\left|\int_{\mathcal{X}_j^{\text{near}}\setminus\{x_j\}} \eta(x)\mathrm{d}\nu(x)\right| + \left|\int_{\mathcal{X}^{\text{far}}} \eta(x)\mathrm{d}\nu(x)\right| \\ &\leqslant 2\delta\|p\|_{\mathcal{H}} + 2\lambda\|p\|_{\mathcal{H}}^2 + \sum_{j=1}^{s}\left|\int_{\mathcal{X}_j^{\text{near}}\setminus\{x_j\}} \eta(x)\mathrm{d}\nu(x)\right| + (1-C_b)|\nu|(\mathcal{X}^{\text{far}}).\end{aligned} \tag{13}$$

To bound $\sum_{j=1}^{s}\left|\int_{\mathcal{X}_j^{\text{near}}\setminus\{x_j\}} \eta(x)\mathrm{d}\nu(x)\right|$, observe that for each j,

$$\begin{aligned}&\left|\int_{\mathcal{X}_j^{\text{near}}\setminus\{x_j\}} \eta(x)\mathrm{d}\nu(x)\right| \leqslant \int_{\mathcal{X}_j^{\text{near}}\setminus\{x_j\}} |\eta(x)|\,\mathrm{d}|\nu|(x) \leqslant \int_{\mathcal{X}_j^{\text{near}}\setminus\{x_j\}} (1 - C_a\|t-x_j\|^2)\mathrm{d}|\nu|(x) \\ &= \int_{\mathcal{X}_j^{\text{near}}\setminus\{x_j\}} \mathrm{d}|\nu|(x) - C_a\int_{\mathcal{X}_j^{\text{near}}\setminus\{x_j\}} \|x-x_j\|^2\,\mathrm{d}|\nu|(x).\end{aligned}$$

Therefore, $\left|\int \eta(x)\mathrm{d}\nu_T(x)\right|$ is upper bounded by

$$2\delta\|p\|_{\mathcal{H}} + 2\lambda\|p\|_{\mathcal{H}}^2 + |\nu|(T^c) - C_b|\nu|(\mathcal{X}^{\text{far}}) - C_a\sum_{j=1}^{s}\int_{\mathcal{X}_j^{\text{near}}} \|x-x_j\|^2\,\mathrm{d}|\nu|(x). \tag{14}$$

So, by denoting $J_\lambda(\mu_0) = |\mu_0| + \frac{1}{2\lambda}\|\Phi\mu_0 - y\|_{\mathcal{H}}^2$,

$$\begin{aligned}&J_\lambda(\mu_0) \geqslant |\hat{\mu}| + \frac{1}{2\lambda}\|\Phi\hat{\mu} - y\|_{\mathcal{H}}^2 \\ &= |\mu_0 + \nu|(\mathcal{X}) + \frac{1}{2\lambda}\|\Phi\mu_0 - y\|_{\mathcal{H}}^2 + \frac{1}{2\lambda}\|\Phi\nu\|_{\mathcal{H}}^2 - \frac{1}{\lambda}\langle \Phi\nu, \Phi\mu_0 - y\rangle_{\mathcal{H}} \\ &\geqslant \frac{1}{\lambda}J_\lambda(\mu_0) - 2|\mu_0|(T^c) + \int \eta(t)\mathrm{d}\nu_T(t) + |\nu|(T^c) + \frac{1}{2\lambda}\|\Phi\nu\|_{\mathcal{H}}^2 - \frac{1}{\lambda}\langle \Phi\nu, \Phi\mu_0 - y\rangle_{\mathcal{H}} \\ &\geqslant \frac{1}{\lambda}J_\lambda(\mu_0) - 2|\mu_0|(T^c) + \int \eta(t)\mathrm{d}\nu_T(t) + |\nu|(T^c) - \frac{\delta^2}{2\lambda}.\end{aligned} \tag{15}$$

where the second inequality follows because

$$\begin{aligned}|\mu_0 + \nu|(\mathcal{X}) &\geqslant \int \eta(x)\mathrm{d}\mu_{0,T}(x) + \int \eta(x)\mathrm{d}\nu_T(x) + |\nu|(T^c) - |\mu_0|(T^c) \\ &= |\mu_0|(X) - 2|\mu_0|(T^c) + \int \eta(x)\mathrm{d}\nu_T(x) + |\nu|(T^c)\end{aligned}$$

and the last inequality follows because

$$\frac{1}{2\lambda}\left\|\Phi\nu\right\|_{\mathcal{H}}^2-\frac{1}{\lambda}\langle\Phi\nu,\,\Phi\mu_0-y\rangle_{\mathcal{H}}\geqslant\frac{1}{2\lambda}\left\|\Phi\nu\right\|_{\mathcal{H}}^2-\frac{\delta}{\lambda}\left\|\Phi\nu\right\|_{\mathcal{H}}\geqslant\frac{-\delta^2}{2\lambda}.$$

Finally, plugging in the bound (14) into (15) implies that

$$C_b\left|\nu\right|(\mathcal{X}^{\text{far}})+C_a\sum_{j=1}^{s}\int_{\mathcal{X}_j^{\text{near}}}\left\|x-x_j\right\|^2\mathrm{d}\left|\nu\right|(x)\leqslant 2\left|\mu_0\right|(T^c)+2\delta\left\|p\right\|_{\mathcal{H}}+2\lambda\left\|p\right\|_{\mathcal{H}}^2+\frac{\delta^2}{2\lambda}. \quad (16)$$

Proof of (5)**:**
First observe that for each j, $\eta_j\overset{\text{def.}}{=}\frac{1}{2}(\eta_{j,1}+\eta_{j,2})$ satisfies

- $\eta_j(x_\ell)=1$ when $\ell=j$ and $\eta_j(x_\ell)=0$ for all $\ell\neq j$,
- $|1-\eta_j(x)|\leqslant C_a\left\|x-x_j\right\|^2$ for all $x\in\mathcal{X}_j^{\text{near}}$,
- $|\eta_j(x)|\leqslant C_a\left\|x-x_\ell\right\|^2$ for all $x\in\mathcal{X}^{\text{near}}(\ell)$ and $l\neq j$,
- $|\eta_j(x)|\leqslant C_b$ for all $x\in\mathcal{X}^{\text{far}}$.

Similarly to (12), we have the bound

$$\left|\int\eta_j(x)\mathrm{d}\nu(x)\right|\leqslant 2\delta\left\|p_j\right\|_{\mathcal{H}}+2\lambda\left\|p_j\right\|_{\mathcal{H}}^2.$$

Observe that

$$\begin{aligned}\left|\int_{\mathcal{X}_j^{\text{near}}}\mathrm{d}\nu(x)\right|&=\left|\int_{\mathcal{X}}\eta_j(x)\mathrm{d}\nu(x)+\int_{\mathcal{X}_j^{\text{near}}}(\eta_j(x)-1)\mathrm{d}\nu(x)-\int_{(\mathcal{X}_j^{\text{near}})^c}\eta_j(x)\mathrm{d}\nu(x)\right|\\&\leqslant\left|\int_{\mathcal{X}}\eta_j(x)\mathrm{d}\nu(x)\right|+C_a\sum_{j=1}^{s}\left|\int_{\mathcal{X}_j^{\text{near}}}\left\|x-x_j\right\|^2\mathrm{d}\nu(x)\right|+C_b\left|\nu\right|(\mathcal{X}^{\text{far}}).\end{aligned}$$

The result follows by applying (16).

Appendix C. Concentration inequalities

Lemma 10 (Bernstein's inequality (Sridharan (2002), Thm. 6)) *Let $x_1,\ldots,x_n\in\mathbb{R}$ be i.i.d. bounded random variables such that $\mathbb{E}x_i=0$, $|x_i|\leqslant M$ and $Var(x_i)\leqslant\sigma^2$ for all i's.*

Then for all $t>0$ we have

$$\mathcal{X}\left(\frac{1}{n}\sum_{i=1}^{n}x_i\geqslant t\right)\leqslant\mathbb{E}\left(-\frac{nt^2}{2\sigma^2+2Mt/3}\right). \quad (17)$$

Lemma 11 (Matrix Bernstein (Tropp (2015), Theorem 6.1.1)) *Consider a finite sequence $Y_1,\ldots,Y_m$ of iid random matrices of size $d_1\times d_2$, assume that*

$$\mathbb{E}Y_j=0,\quad\left\|Y_j\right\|\leqslant L,\quad v(Y_j):=\max(\left\|\mathbb{E}Y_jY_j^\top\right\|,\left\|\mathbb{E}Y_j^\top Y_j\right\|)\leqslant M$$

for each index $1 \leqslant j \leqslant m$. Introduce the random matrix

$$Z = \frac{1}{m}\sum_j Y_j$$

Then

$$\mathbb{P}\left(\|Z\| \geqslant t\right) \leqslant (d_1 + d_2) e^{-\frac{mt^2/2}{M+Lt/3}} \tag{18}$$

Lemma 12 (Hoeffding's inequality ((Tang et al., 2013), Lemma G.1)) *Let the components of $u \in \mathbb{R}^k$ be drawn iid from a Rademacher distribution, consider a vector $w \in \mathbb{R}^k$. Then, with probability at least $1-\rho$, we have*

$$\mathbb{P}\left(\left|u^\top w\right| \geqslant t\right) \leqslant 4e^{-\frac{t^2}{4\|w\|^2}} \tag{19}$$

Appendix D. Notations and first properties.

We recall that $\|\cdot\|$ designates the modulus for complex scalars, the ℓ_2 norm for complex vectors, and for complex symmetric matrices $\|\cdot\| = \left(\|\mathrm{Re}\,(\cdot)\|_{2\to2}^2 + \|\mathrm{Im}\,(\cdot)\|_{2\to2}^2\right)^{\frac{1}{2}}$ where $\|\cdot\|_{2\to2}$ is the spectral norm of matrices. The operator ∇^r is the identity for $r=0$, gradient for $r=1$ and Hessian matrix for $r=2$. For complex functions f, it is just $\nabla^r f = \nabla^r \mathrm{Re}\,(f) + \mathrm{i}\nabla^r \mathrm{Im}\,(f)$. For a bivariate function $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, $\partial_{1,i}$ (resp. $\partial_{2,i}$) designates the derivative with respect to the i^{th} coordinate of the first variable (resp. second variable), and similarly for the gradient operator ∇ and Hessian operator ∇^2. The object $\nabla_1\nabla_2^2 K$ is a $d\times d\times d$ tensor whose "spectral" norm is defined as $\left\|\nabla_1\nabla_2^2 K\right\| = \sup_{\|u\|\leqslant 1}\left\|\sum_{i=1}^d u_i \partial_{1,i}\nabla_2^2 K\right\|$.

In the rest of the proof, we always consider some a_i and x_i that are clear from the context. We recall the definition of section 3.3:

$$\gamma(\omega) \overset{\text{def.}}{=} D\left(\overline{\varphi_\omega(x_1)}, ..., \overline{\varphi_\omega(x_s)}, \nabla\overline{\varphi_\omega(x_1)}^\top, ..., \nabla\overline{\varphi_\omega(x_s)}^\top\right)^\top \in \mathbb{C}^{s(d+1)}$$

where, denoting $v_{il} = 1/\sqrt{\partial_{1,i}\partial_{2,i}K(x_l,x_l)} \geqslant v^{-\frac{1}{2}}$, we define

$$D = \mathrm{diag}\left((1,\dots,1,v_{11},\dots,v_{d1},v_{12},\dots\dots,v_{dk})\right)$$

a diagonal matrix whose first s elements are 1's, which is here for normalization purpose.

Note that under the assumptions of Thm 18 we have $\|\gamma\| \leqslant \sqrt{s(L_0^2 + L_1^2/v)} = \sqrt{s}L_{01}$ with $L_{01} \overset{\text{def.}}{=} \sqrt{L_0^2 + L_1^2/v}$, where the L_j's are the constants defined in (8).

We define the following object that we are going to use in the proofs.

$$\overline{\Upsilon} \overset{\text{def.}}{=} \mathbb{E}_\omega \mathrm{Re}\left(\gamma(\omega)\gamma(\omega)^H\right) \in \mathbb{R}^{s(d+1)\times s(d+1)}$$

$$\overline{\mathbf{f}}(x) \overset{\text{def.}}{=} \mathbb{E}_\omega \mathrm{Re}\left(\gamma(\omega)\varphi_\omega(x)\right) \in \mathbb{R}^{s(d+1)}$$

$$\overline{\alpha} \overset{\text{def.}}{=} \overline{\Upsilon}^{-1}\mathbf{u}_s \quad \text{where} \quad \mathbf{u}_s \overset{\text{def.}}{=} \begin{pmatrix}\mathrm{sign}(a)\\ \mathbf{0}_{sd}\end{pmatrix}$$

Note that D has been chosen so that the diagonal of $\overline{\Upsilon}$ has only 1's (in fact, $\overline{\alpha}^\top\overline{\mathbf{f}}$ is independent of the matrix D as long as its diagonal's first s coefficients are 1's).

Next, for $\omega_1, \ldots, \omega_m$, we denote their empirical version:

$$\Upsilon \overset{\text{def.}}{=} \frac{1}{m}\sum_{k=1}^{m} \mathrm{Re}\left(\gamma(\omega_k)\gamma(\omega_k)^H\right), \quad \mathbf{f}(x) \overset{\text{def.}}{=} \frac{1}{m}\sum_{k=1}^{m} \mathrm{Re}\left(\gamma(\omega_k)\varphi_{\omega_k}(x)\right), \quad \alpha \overset{\text{def.}}{=} \Upsilon^{-1}\mathbf{u}_s$$

which will serve us to construct our certificate, using the properties of their respective limit version. Note that the vanishing derivative pre-certificate η_V is $\alpha^\top \mathbf{f}(\cdot)$.

In the course of the golfing scheme, we are also going to use independent blocks of (independent) random frequencies. For a block B_l containing m_l frequencies, we denote respectively Υ_l and $\mathbf{f}_l$ the empirical sum over these m_l frequencies (eg $\mathbf{f}_l = \frac{1}{m_l}\sum_{k=1}^{m_l} \mathrm{Re}\left(\gamma(\omega_k)\varphi_{\omega_k}(x)\right)$).

Effect of D. We claimed that by changing the diagonal matrix D and by keeping the first s elements on its diagonal fixed at 1 and the other non-zero, the certificate η_V do not change. Indeed, if we considered γ without D, the effect of adding the D would give the function:

$$\eta_V(x) = \left((D\Upsilon D)^{-1}\mathbf{u}_s\right)^\top D\mathbf{f}(x) = (D^{-1}\mathbf{u}_s)^\top \Upsilon^{-1}\mathbf{f}(x)$$

and since $D^{-1}\mathbf{u}_s = \mathbf{u}_s$, the normalization is indeed without effect on η_V.

Finally, we define the linear operator $\Psi : \mathscr{C}^1(\mathcal{X}) \to \mathbb{R}^{s(d+1)}$ as

$$\Psi f = \left[f(x_1), \ldots, f(x_s), \nabla f(x_1)^\top, \ldots, \nabla f(x_s)^\top\right]^\top \tag{20}$$

A useful property is that for any vector $q \in \mathbb{R}^{s(d+1)}$, one has

$$\Psi\left((\overline{\Upsilon}^{-1}q)^\top \overline{\mathbf{f}}(\cdot)\right) = D^{-1}q \tag{21}$$

and similarly if we replace $\overline{\Upsilon}$ and $\overline{\mathbf{f}}$ by their subsampled versions. Note that by putting $q = \mathbf{u}_s$, this expresses the fact that η_V interpolates the signs of a_i at x_i with a cancelling gradient. If one replaces $\overline{\mathbf{f}}$ by its subsampled version but not $\overline{\Upsilon}$, one obtains

$$\Psi\left((\overline{\Upsilon}^{-1}q)^\top \mathbf{f}(\cdot)\right) = D^{-1}\Upsilon\overline{\Upsilon}^{-1}q \tag{22}$$

Concentration inequalities. In the course of our proofs, we will frequently use the following probabilistic bounds.

Lemma 13 *For any $\varepsilon \leqslant 1/4$,*

$$\mathbb{P}\left(\left\|\overline{\Upsilon} - \Upsilon\right\| \geqslant \varepsilon\right) \leqslant 2s(d+1)e^{-\frac{m\varepsilon^2}{7sL_{01}^2}} \tag{23}$$

Proof We are using Lemma 11 with $Y_k = \mathrm{Re}\left(\gamma(\omega_k)\gamma(\omega_j)^H\right) - \overline{\Upsilon}$. We have:

$$\mathbb{E}Y = 0, \quad \|Y\| \leqslant sL_{01}^2 + \left\|\overline{\Upsilon}\right\| \leqslant 2sL_{01}^2$$

since $\left\|\overline{\Upsilon}\right\| \leqslant \|\gamma(\omega)\|^2 \leqslant sL_{01}^2$ (for simplicity). Denoting $A = \gamma(\omega)\gamma(\omega)^H$, we can write

$$\begin{aligned}\left\|\mathbb{E}\mathrm{Re}\,(A)\mathrm{Re}\,(A)^\top\right\| &\leqslant \left\|\mathbb{E}\mathrm{Re}\,(A)\mathrm{Re}\,(A)^\top + \mathbb{E}\mathrm{Im}\,(A)\mathrm{Im}\,(A)^\top\right\| = \left\|\mathbb{E}\mathrm{Re}\left(AA^H\right)\right\| \\ &\leqslant B_\gamma^2\left\|\mathbb{E}\mathrm{Re}\left(\gamma(\omega)\gamma(\omega)^H\right)\right\| = B_\gamma^2\left\|\overline{\Upsilon}\right\|\end{aligned}$$

where, for the first inequality, we have used the fact that for two positive definite matrices A, B we have $\|A\| = \sup_{x,\ \|x\|_2=1} x^\top A x \leqslant \sup_{x,\ \|x\|_2=1} x^\top (A+B) x = \|A+B\|$. Therefore

$$\begin{aligned} v(Y) &= \left\| \mathbb{E} \left(\mathrm{Re} \left(\gamma(\omega)\gamma(\omega)^H \right) - \overline{\Upsilon} \right)^2 \right\| \\ &= \left\| \mathbb{E}\mathrm{Re} \left(\gamma(\omega_j)\gamma(\omega_j)^H \right)^2 - \overline{\Upsilon}^2 \right\| \leqslant sL_{01}^2 \left\| \overline{\Upsilon} \right\| + \left\| \overline{\Upsilon} \right\|^2 \leqslant 3sL_{01}^2 \end{aligned}$$

Applying Lemma 11, for all $\varepsilon \leqslant 1/4$ (for simplicity), we obtain that

$$\mathbb{P} \left(\left\| \frac{1}{m} \sum_{j=1}^m Y_j \right\| \geqslant \varepsilon \right) \leqslant 2s(d+1) e^{-\frac{m\varepsilon^2}{(6+1/3)sL_{01}^2}} = 2s(d+1) e^{-\frac{m\varepsilon^2}{7sL_{01}^2}}$$

which is the desired result. ■

The next useful Lemma is used in the two corollaries that come after which we shall repeatedly use.

Lemma 14 *Let $g(\omega)$ be any complex function such that $|g(\omega)| \leqslant L$ almost surely, and $q \in \mathbb{R}^{s(d+1)}$ be any vector. Define $Y_k = \mathrm{Re}\left(\gamma(\omega_k) g(\omega_k)\right) q - \mathbb{E}\mathrm{Re}\left(\gamma(\omega) g(\omega)\right) q \in \mathbb{R}$. Then we have*

$$\mathbb{P} \left(\left| \frac{1}{m} \sum_{k=1}^m Y_k \right| \geqslant t \|q\| \right) \leqslant \mathbb{E} \left(- \frac{mt^2}{4L^2 \left\| \overline{\Upsilon} \right\| + 4LL_{01}\sqrt{s}t/3)} \right) \tag{24}$$

and as a corollary

$$\mathbb{P} \left(\left| \frac{1}{m} \sum_{k=1}^m Y_k \right| \geqslant t \|q\|_\infty \right) \leqslant \mathbb{E} \left(- \frac{mt^2}{4L^2 s(d+1) \left\| \overline{\Upsilon} \right\| + 4LL_{01}s\sqrt{(d+1)}t/3} \right) \tag{25}$$

Proof It is a simple use of Bernstein's inequality. We have

$$|Y_k| \leqslant 2LL_{01}\sqrt{s} \|q\|$$

and, defining the shorthand $v_k = \gamma(\omega_k) g(\omega_k)$ we get

$$\begin{aligned} \mathbb{E} |Y_k|^2 &\leqslant 2\mathbb{E} \left(q^\top \mathrm{Re}\,(v_k) \mathrm{Re}\,(v_k)^\top q \right) \\ &\leqslant 2\mathbb{E} \left(q^\top \left(\mathrm{Re}\,(v_k)\mathrm{Re}\,(v_k)^\top + \mathrm{Im}\,(v_k)\mathrm{Im}\,(v_k)^\top \right) q \right) \\ &= 2q^\top \mathbb{E} \left(|g(\omega_k)|^2 \, \mathrm{Re} \left(\gamma(\omega_k)\gamma(\omega_k)^H \right) \right) q \\ &\leqslant 2L^2 \left\| \overline{\Upsilon} \right\| \|q\|^2 \end{aligned}$$

Hence applying Bernstein's inequality (Lemma 10) we obtain (24). ■

Immediates corollary from the previous Lemma are the following

Corollary 15 *For any vector $q \in \mathbb{R}^{s(d+1)}$,*

$$\mathbb{P}\left(\left\|(\overline{\Upsilon}-\Upsilon)q\right\|_\infty \geqslant t\left\|q\right\|_\infty\right) \leqslant s(d+1)\mathbb{E}\left(-\frac{mt^2}{4L_{01}^2\left(\left\|\overline{\Upsilon}\right\|+s\sqrt{d+1}t/3\right)}\right) \tag{26}$$

Proof For each coordinate $\gamma_i(\omega)$ of γ, apply Lemma 14 with $g(\omega)$ defined as $\gamma_i(\omega)$, by noting that $|\gamma_i| \leqslant L_{01}$, to obtain a bound on $((\overline{\Upsilon}-\Upsilon)q)_i$, then apply a union bound. ■

Corollary 16 *For any vector q and $x \in \mathcal{X}$, we have*

$$\mathbb{P}\left(\left|(\overline{\mathbf{f}}(x)-\mathbf{f}(x))^\top q\right| \geqslant t\left\|q\right\|\right) \leqslant \mathbb{E}\left(-\frac{mt^2}{4L_0^2\left\|\overline{\Upsilon}\right\|+4L_0L_{01}\sqrt{s}t/3}\right) \tag{27}$$

and

$$\mathbb{P}\left(\left|(\overline{\mathbf{f}}(x)-\mathbf{f}(x))^\top q\right| \geqslant t\left\|q\right\|_\infty\right) \leqslant \mathbb{E}\left(-\frac{mt^2}{4L_0^2 s(d+1)\left\|\overline{\Upsilon}\right\|+4L_0L_{01}s\sqrt{(d+1)}t/3}\right) \tag{28}$$

Proof Just apply Lemma 14 with $g(\omega)=\varphi_\omega(x)$. ■

And finally, we have the same result for the Hessian:

Proposition 17 *For any vector q and $x \in \mathcal{X}$, we have*

$$\mathbb{P}\left(\left\|\nabla^2\left((\overline{\mathbf{f}}(x)-\mathbf{f}(x))^\top q\right)\right\| \geqslant t\left\|q\right\|\right) \leqslant 2d\mathbb{E}\left(-\frac{mt^2}{4L_2^2\left\|\overline{\Upsilon}\right\|+4L_2L_{01}\sqrt{s}t/3}\right) \tag{29}$$

and

$$\mathbb{P}\left(\left\|\nabla^2\left((\overline{\mathbf{f}}(x)-\mathbf{f}(x))^\top q\right)\right\| \geqslant t\left\|q\right\|_\infty\right) \leqslant 2d\mathbb{E}\left(-\frac{mt^2}{4L_2^2 s(d+1)\left\|\overline{\Upsilon}\right\|+4L_2L_{01}s\sqrt{(d+1)}t/3}\right) \tag{30}$$

Proof We define

$$Y_k = \mathrm{Re}\left((q^\top\gamma(\omega_k))\nabla^2\varphi_{\omega_k}(x)\right) - \sum_{l=1}^{Q} q_l \nabla^2 f_l(x)$$

which are indeed symmetric matrices.

We have $\mathbb{E}_\omega Y_k = 0$ and

$$\|Y_k\| \leqslant 2\left\|q\right\|_2\sqrt{s}L_{01}L_2$$

Furthermore, defining $A=(q^\top\gamma(\omega_k))\nabla^2\varphi_{\omega_k}(x)$ (which is symmetric). As in the proof of Lemma 13 we have

$$\begin{aligned}
\left\|\mathbb{E}\left(\mathrm{Re}\,(A)\mathrm{Re}\,(A)^\top\right)\right\| &\leqslant \left\|\mathbb{E}\left(\mathrm{Re}\,(A)\mathrm{Re}\,(A)^\top\right)+\mathbb{E}\left(\mathrm{Im}\,(A)\mathrm{Im}\,(A)^\top\right)\right\| \\
&= \left\|\mathbb{E}\left(AA^H\right)\right\| \leqslant L_2^2\mathbb{E}\left|q^\top\gamma(\omega)\right|^2 = L_2^2\mathbb{E}\mathrm{Re}\left(q^\top\gamma(\omega)\gamma(\omega)^H q\right) \\
&= L_2^2 q^\top\left(\mathbb{E}\mathrm{Re}\left(\gamma(\omega)\gamma(\omega)^H\right)\right)q \leqslant L_2^2\left\|q\right\|_2^2\left\|\overline{\Upsilon}\right\|
\end{aligned}$$

	Order 0	Order 1	Order 2	Order 3
$x = x'$	$\overline{K} = 1$	$\left\lvert \partial_i \overline{K} \right\rvert \leqslant a_1$	$\left\lvert \partial_{1,i}\partial_{2,j}\overline{K} \right\rvert \leqslant a_2$ $\partial_{1,i}\partial_{2,i}\overline{K} \geqslant v$	n/a
$\lVert x - x' \rVert_{\text{sep.}} \leqslant \varepsilon_{\text{near}}$	n/a	n/a	$-b_2 \leqslant \operatorname{eig}\left(\nabla_2^2 \overline{K}\right) \leqslant -\lambda_1$	$\left\lVert \nabla_1 \nabla_2^2 \overline{K} \right\rVert \leqslant b_3$
$\lVert x - x' \rVert_{\text{sep.}} \geqslant \varepsilon_{\text{near}}$	$\left\lvert \overline{K} \right\rvert \leqslant c_0$	$\left\lVert \nabla_1 \overline{K} \right\rVert \leqslant c_1$	n/a	n/a
$\lVert x - x' \rVert_{\text{sep.}} \geqslant \Delta/2$	$\left\lvert \overline{K} \right\rvert \leqslant \frac{e_0}{s_{\max}}$	$\left\lVert \nabla_1 \overline{K} \right\rVert \leqslant \frac{e_1}{s_{\max}}$	$\left\lVert \nabla_2^2 \overline{K} \right\rVert \leqslant \frac{e_2}{s_{\max}}$	$\left\lVert \nabla_1 \nabla_2^2 \overline{K} \right\rVert \leqslant \frac{e_3}{s_{\max}}$
$\lVert x - x' \rVert_{\text{sep.}} \geqslant \Delta$	$\left\lvert \overline{K} \right\rvert \leqslant \frac{h_0}{s_{\max}}$	$\left\lvert \partial_{1,i} \overline{K} \right\rvert \leqslant \frac{h_1}{s_{\max}}$	$\left\lVert \partial_{1,i} \nabla_2 \overline{K} \right\rVert_1 \leqslant \frac{h_2}{s_{\max}}$	n/a

Table 1: Assumptions on the covariance kernel for Thm 18. Each line $\overline{K}$ is a shorthand for $\overline{K}(x, x')$, and x, x' are any elements of $\mathcal{X}$ that are as described in the first column. The notation $\operatorname{eig}(\cdot)$ designates any eigenvalue of a symmetric matrix.

And therefore

$$\begin{aligned}\left\lVert \mathbb{E} Y_j Y_j^\top \right\rVert &= \left\lVert \mathbb{E}(\mathrm{Re}\,(A) - \mathbb{E}\mathrm{Re}\,(A))(\mathrm{Re}\,(A) - \mathbb{E}\mathrm{Re}\,(A))^\top \right\rVert \\ &= \left\lVert \mathbb{E}\left(\mathrm{Re}\,(A)^2\right) - \mathbb{E}\mathrm{Re}\,(A)^2 \right\rVert \leqslant 2 \left\lVert \mathbb{E}\left(\mathrm{Re}\,(A)\mathrm{Re}\,(A)^\top\right) \right\rVert \leqslant 2L_2^2 \lVert q \rVert_2^2 \left\lVert \overline{\Upsilon} \right\rVert\end{aligned}$$

We can therefore apply the matrix Bernstein's inequality to obtain the desired result. ■

Appendix E. Acceptable kernels

In this section, we precisely define what is an acceptable kernel (Def. 3). Although we do not explicitely require the kernel to be translation invariant (ie $\overline{K}(x, x') = \overline{K}(x - x')$), our analysis is a generalization of (Candès and Fernandez-Granda, 2014) and is taylored for translation-invariant kernels. Intuitively, we require the kernel to have a negative curvature for $\lVert x - x' \rVert_{\text{sep.}} \leqslant \varepsilon_{\text{near}}$, and to decrease sufficiently for $\lVert x - x' \rVert_{\text{sep.}} \geqslant \Delta$. If this is the case, we prove in the following Theorem that the function $\mathbf{u}_s^\top \overline{\Upsilon}^{-1} \overline{\mathbf{f}}(\cdot)$ is a non-degenerate certificate (for the limit problem $m \to \infty$).

Theorem 18 *Assume there exist $s_{\max} > 0$, $\varepsilon_{\text{near}} > 0$ and $\Delta > 0$ such that $\overline{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfy the bounds indicated in Table 1. Define $u = (a_1 + h_1)/\sqrt{v}$. Assume that there exist $\delta, \delta' < 1$ such that*

$$v^{-1}(da_2 + h_2) \leqslant \delta, \quad h_0 + \tfrac{du^2}{1-\delta} \leqslant \delta' \tag{31}$$

$$\varepsilon_\eta \stackrel{\text{def.}}{=} 1 - \left(\tfrac{c_0 + e_0}{1-\delta'} + \tfrac{u\sqrt{d}}{\sqrt{v}(1-\delta)(1-\delta')} \cdot (c_1 + e_1)\right) > 0 \tag{32}$$

$$\lambda_\eta \stackrel{\text{def.}}{=} \left(1 - \tfrac{\delta'}{1-\delta'}\right) \cdot \lambda_1 - \tfrac{e_2}{1-\delta'} - \tfrac{u\sqrt{d}}{\sqrt{v}(1-\delta)(1-\delta')} \cdot (b_3 + e_3) > 0 \tag{33}$$

Then, for $s \leqslant s_{\max}$, for all $a_1, \ldots, a_s \in \mathbb{R}$ and $x_1, ..., x_s \in \mathcal{X}$ such that $\|x - x'\|_{sep.} \geqslant \Delta$, the function $\overline{\eta} = \mathbf{u}_s^\top \overline{\Upsilon}^{-1} \overline{\mathbf{f}}(\cdot)$ is such that $\forall\, i = 1, \ldots, s$:

$$\begin{aligned}
&\eta(x_i) = \operatorname{sign}(a_i) \\
&\forall\, x \quad s.t. \quad \|x_i - x\|_{sep.} \geqslant \varepsilon_{\text{near}}, \quad |\eta(x)| < 1 - \varepsilon_\eta \\
&\forall\, x \quad s.t. \quad \|x_i - x\|_{sep.} \leqslant \varepsilon_{\text{near}}, \quad -\operatorname{sign}(a_i)\nabla^2\eta(x_i) \succ \lambda_\eta \mathrm{Id}
\end{aligned}$$

In the rest of the proofs, we will also assume that the following is satisfied:

$$\max\{\delta + u,\ \delta' + du\left(1 - \tfrac{u}{1-\delta}\right)\} \leqslant 1/2, \tag{34}$$

and make use of the bounds $B_0 \stackrel{\text{def.}}{=} \left(c_0^2 + c_1^2 + e_0^2 + e_1^2\right)^{\frac{1}{2}}$ and $B_2 \stackrel{\text{def.}}{=} \left(b_2^2 + e_2^2 + b_3^2 + e_3^2\right)^{\frac{1}{2}}$.

Proof [Proof of Theorem 18] Fix a_i and x_i such that $\|x_i - x_j\|_{\text{sep.}} \geqslant \Delta$. Define $\overline{\Upsilon}$, $\overline{\mathbf{f}}$ and $\overline{\alpha}$ as in the previous section. Note that they can be defined only in terms of $\overline{K}$, and therefore can be defined for any smooth symmetric function $C : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

Invertibility of $\overline{\Upsilon}$. We first prove that, under conditions (31), the matrix $\overline{\Upsilon}$ is invertible. For that we divide it as:

$$\overline{\Upsilon} = \begin{pmatrix} \overline{\Upsilon}_0 & \overline{\Upsilon}_1 \\ \overline{\Upsilon}_1^\top & \overline{\Upsilon}_2 \end{pmatrix} \tag{35}$$

where $\overline{\Upsilon}_0 \in \mathbb{R}^{s \times s}$ and $\overline{\Upsilon}_2 \in \mathbb{R}^{sd \times sd}$. Remember that $\overline{\Upsilon}$ has been normalized (through the diagonal matrix D) to have only 1's on its diagonal, and that is also true for $\overline{\Upsilon}_0$ and $\overline{\Upsilon}_2$.

To prove the invertibility of $\overline{\Upsilon}$ and derive useful bounds, we use the Schur complement of $\overline{\Upsilon}$, defined, if $\overline{\Upsilon}_2$ is invertible, as $\overline{\Upsilon}_S \stackrel{\text{def.}}{=} \overline{\Upsilon}_0 - \overline{\Upsilon}_1 \overline{\Upsilon}_2^{-1} \overline{\Upsilon}_1^\top$. If both $\overline{\Upsilon}_2$ and $\overline{\Upsilon}_S$ are invertible, then so is $\overline{\Upsilon}$ and its inverse can be expressed using $\overline{\Upsilon}_S$. Hence we must first prove that $\overline{\Upsilon}_2$ is invertible, and for that we use the following bound

$$\begin{aligned}
\left\|I - \overline{\Upsilon}_2\right\|_\infty &= \sup_{1 \leqslant i \leqslant d,\ 1 \leqslant l \leqslant s} \sum_{j \neq i} v_{il} v_{jl} \left|\partial_{1,j}\partial_{2,i}\overline{K}(x_l, x_l)\right| + \sum_{l' \neq l} \sum_{j=1}^{d} v_{il} v_{jl'} \left|\partial_{1,j}\partial_{2,i}\overline{K}(x_{l'}, x_l)\right| \\
&\leqslant v^{-1}\left((d-1)a_2 + (s-1)\frac{h_2}{s_{\max}}\right) \stackrel{(31)}{\leqslant} \delta
\end{aligned}$$

Since $\left\|I - \overline{\Upsilon}_2\right\|_\infty < 1$, $\overline{\Upsilon}_2$ is invertible, and we have $\left\|\overline{\Upsilon}_2^{-1}\right\|_\infty \leqslant \frac{1}{1 - \left\|I - \overline{\Upsilon}_2\right\|_\infty}$. Next, we can bound

$$\begin{aligned}
\left\|I - \overline{\Upsilon}_0\right\|_\infty &= \sup_{1 \leqslant l \leqslant s} \sum_{l' \neq l} \left|\overline{K}(x_{l'}, x_l)\right| \leqslant (s-1)\frac{h_0}{s_{\max}} \leqslant h_0 \\
\left\|\overline{\Upsilon}_1\right\|_\infty &= \sup_{1 \leqslant l \leqslant s} \sum_{i=1}^{d} v_{il} \left|\partial_{1,i}\overline{K}(x_l, x_l)\right| + \sum_{l' \neq l} \sum_{i=1}^{d} v_{il'} \left|\partial_{1,j}\overline{K}(x_{l'}, x_l)\right| \\
&\leqslant v^{-\frac{1}{2}}\left(da_1 + (s-1)d\frac{h_1}{s_{\max}}\right) \leqslant du \\
\left\|\overline{\Upsilon}_1^\top\right\|_\infty &= \sup_{1 \leqslant i \leqslant d,\ 1 \leqslant l \leqslant k} v_{il} \sum_{l'=1}^{s} \left|\partial_{2,i}\overline{K}(x_{l'}, x_l)\right| \leqslant v^{-\frac{1}{2}}\left(a_1 + (s-1)\frac{h_1}{s_{\max}}\right) \leqslant u
\end{aligned}$$

where $u = (a_1 + h_1)/\sqrt{v}$. Hence, we have

$$\left\|I - \overline{\Upsilon}_S\right\|_\infty \leqslant \left\|I - \overline{\Upsilon}_0\right\|_\infty + \left\|\overline{\Upsilon}_1\right\|_\infty \left\|\overline{\Upsilon}_2^{-1}\right\|_\infty \left\|\overline{\Upsilon}_1^\top\right\|_\infty \leqslant h_0 + \frac{du^2}{1-\delta} \overset{(31)}{\leqslant} \delta' \tag{36}$$

and therefore the Schur complement of $\overline{\Upsilon}$ is invertible and so is $\overline{\Upsilon}$.

Expression of $\overline{\eta}$. By definition, $\overline{\eta}$ satisfies $\overline{\eta}(x_i) = \operatorname{sign}(a_i)$ and $\nabla\overline{\eta}(x_i) = 0$.

Again we divide:

$$\overline{\alpha} = \begin{pmatrix}\overline{\alpha}_1 \\ \overline{\alpha}_2\end{pmatrix}, \qquad \overline{\mathbf{f}}(x) = \begin{pmatrix}\overline{\mathbf{f}}_1(x) \\ \overline{\mathbf{f}}_2(x)\end{pmatrix}$$

where $\overline{\alpha}_1$, $\overline{\mathbf{f}}_1(x)$ are vectors of size s and $\overline{\alpha}_2$, $\overline{\mathbf{f}}_2(x)$ are vectors of size sd. The Schur's complement of $\overline{\Upsilon}$ allows us to express $\overline{\alpha}_1$ and $\overline{\alpha}_2$ as

$$\overline{\alpha}_1 = \overline{\Upsilon}_s^{-1} \operatorname{sign}(a), \qquad \overline{\alpha}_2 = -\overline{\Upsilon}_2^{-1}\overline{\Upsilon}_1^\top\overline{\Upsilon}_S^{-1} \operatorname{sign}(a) \tag{37}$$

and therefore we can bound

$$\|\overline{\alpha}_1\|_\infty \leqslant \frac{1}{1-\delta'} \tag{38}$$

$$\|\overline{\alpha}_2\|_\infty \leqslant \frac{u}{(1-\delta)(1-\delta')} \tag{39}$$

Moreover, we have

$$\|\overline{\alpha}_1 - \operatorname{sign}(a)\|_\infty \leqslant \left\|I - \overline{\Upsilon}_S^{-1}\right\|_\infty \leqslant \left\|\overline{\Upsilon}_S^{-1}\right\|_\infty \left\|I - \overline{\Upsilon}_S\right\|_\infty \leqslant \frac{\delta'}{1-\delta'} \tag{40}$$

Non-degeneracy. We can now prove that $\overline{\eta}$ is non-degenerate. More precisely, we are going to prove that for all x such that $\|x - x_i\|_{\text{sep.}} \leqslant \varepsilon_{\text{near}}$, all eigenvalues of $-\operatorname{sign}(a_i)\nabla^2\overline{\eta}(x)$ are above λ_η (defined by (33)), and for all other x's, $|\overline{\eta}(x)| \leqslant 1 - \varepsilon_\eta$ where ε_η is defined by (32).

Let x be such that $\|x - x_i\|_{\text{sep.}} \leqslant \varepsilon_{\text{near}}$. Then, since $\varepsilon_{\text{near}} \leqslant \Delta/2$ and the x_l's are Δ-separated, for all $l' \neq l$ we have $\|x - x_{l'}\| \geqslant \Delta/2$. Then, we have

$$\begin{aligned}
-\operatorname{sign}(a_i)\nabla^2\overline{\eta}(x) = {} & -\operatorname{sign}(a_i)\Bigg[(\overline{\alpha}_1)_l\nabla_2^2\overline{K}(x_l,x) + \sum_{l'\neq l}(\overline{\alpha}_1)_{l'}\nabla_2^2\overline{K}(x_{l'},x) \\
& + \sum_{i=1}^d\Big((\overline{\alpha}_2)_{d(l-1)+i}v_{il}\nabla_2^2\partial_{1,i}\overline{K}(x_l,x) + \sum_{l'\neq l}(\overline{\alpha}_2)_{d(l'-1)+i}v_{il'}\nabla_2^2\partial_{1,i}\overline{K}(x_{l'},x)\Big)\Bigg] \\
\overset{(38),(39),(40)}{\succcurlyeq} {} & \left(\frac{1}{1-\delta'}\left(\lambda_{\overline{K}} - (s-1)\frac{e_2}{s_{\max}}\right) - \frac{\sqrt{d}u}{\sqrt{v}(1-\delta)(1-\delta')}(b_3 + (s-1)\frac{e_3}{s_{\max}})\right)\mathrm{Id} \\
\overset{(33)}{\succcurlyeq} {} & \lambda_\eta\mathrm{Id} \succ 0
\end{aligned}$$

where for the second term we have used the fact that for $t \in \mathbb{R}^d$

$$\sum_{i=1}^d t_i\nabla_2^2\partial_{1,i}\overline{K}(x,x') \preccurlyeq \|t\| \left\|\nabla_1\nabla_2^2\overline{K}(x,x')\right\| \mathrm{Id}$$

and $\|t\|_2 \leqslant \sqrt{d}\,\|t\|_\infty$. Thus we proved that $s_l \nabla^2 \eta$ is uniformly positive definite inside an $\varepsilon_{\text{near}}$ neighborhood of x_l, with all its eigenvalues greater than λ_η.

Next, for any x such that $\|x - x_i\|_{\text{sep.}} \geqslant \varepsilon_{\text{near}}$ for all x_i's, we can say that x is $\Delta/2$-far from all x_i's except one, for which it is only $\varepsilon_{\text{near}}$-far. Let us call this point x_l. We have

$$
\begin{aligned}
|\overline{\eta}(x)| = \Bigg| & (\overline{\alpha}_1)_l \overline{K}(x_l, x) + \sum_{l' \neq l} (\overline{\alpha}_1)_{l'} \overline{K}(x_{l'}, x) \\
& + \sum_{i=1}^{d} \Big((\overline{\alpha}_2)_{d(l-1)+i} v_{il} \partial_{1,i} \overline{K}(x_l, x) + \sum_{l' \neq l} (\overline{\alpha}_2)_{d(l'-1)+i} v_{il'} \partial_{1,i} \overline{K}(x_{l'}, x) \Bigg| \\
\overset{(38),(39)}{\leqslant} & \frac{1}{1-\delta'} \left(c_0 + (s-1) \frac{e_0}{s_{\max}} \right) + \frac{\sqrt{d} u}{\sqrt{v}(1-\delta)(1-\delta')} \left(c_1 + (s-1) \frac{e_1}{s_{\max}} \right) \overset{(32)}{\leqslant} 1 - \varepsilon_\eta < 1
\end{aligned}
$$

■

Additional bounds. We finish this section by outlining several bounds that are useful for the rest of the proofs.

Under (34), we have $\left\|\text{Id} - \overline{\Upsilon}\right\|_\infty \leqslant 1/2$, and therefore

$$
\begin{aligned}
\left\|\overline{\Upsilon}\right\|_\infty &\leqslant 3/2 \\
\left\|\overline{\Upsilon}^{-1}\right\|_\infty &\leqslant 2
\end{aligned} \tag{41}
$$

and similarly for the spectral norm, since $\overline{\Upsilon}$ and $\overline{\Upsilon}^{-1}$ are symmetric and therefore their spectral norm is lower than their ∞ norm.

Then, we note that for any vector $q \in \mathbb{R}^{s(d+1)}$ and any $x \in \mathcal{X}^{\text{far}}$, we have

$$
\left| q^\top \overline{\mathbf{f}}(x) \right| \leqslant \|q\| \left(\sum_{i=1}^{s} \left| \overline{K}(x_i, x) \right|^2 + \sum_{j=1}^{d} \left| \partial_{1,j} \overline{K}(x_i, x) \right|^2 \right)^{\frac{1}{2}} \leqslant \|q\|\, B_0 \tag{42}
$$

for which, similar to the proof above, we have used the fact that x is $\Delta/2$-separated from $s-1$ points x_i. Similarly,

$$
\left| q^\top \overline{\mathbf{f}}(x) \right| \leqslant \|q\|_\infty \left(c_0 + c_1 + \sqrt{d}(e_0 + e_1) \right) \leqslant 2\, \|q\|_\infty \sqrt{d} B_0 \tag{43}
$$

For the second derivative, for $x \in \mathcal{X}^{\text{near}}$ we have the bound:

$$
\begin{aligned}
\left\| \nabla^2 \left(q^\top \overline{\mathbf{f}}(x) \right) \right\| &= \left\| \sum_{i=1}^{s} q_i \nabla_2^2 \overline{K}(x_i, x) + \sum_{j=1}^{d} q_j^{(i)} \partial_{1,j} \nabla_2^2 \overline{K}(x_i, x) \right\| \quad \text{where} \quad q_j^{(i)} = q_{(i-1)d+j+s} \\
&\leqslant \sum_{i=1}^{s} |q_i| \left\| \nabla_2^2 \overline{K}(x_i, x) \right\| + \left\| q^{(i)} \right\| \left\| \nabla_1 \nabla_2^2 \overline{K}(x_i, x) \right\| \\
&\leqslant \begin{cases} \|q\|\, B_2 \\ 2\, \|q\|_\infty \sqrt{d} B_2 \end{cases}
\end{aligned} \tag{44}
$$

And finally, we will also use the bound

$$\sum_{l=1}^{s(d+1)} \left\|\nabla_2^2 f_l(x)\right\|^2 \leqslant b_2^2 + e_2^2 + \sum_{i=1}^{s}\sum_{j=1}^{d} \left\|\partial_{1,j}\nabla_2^2 \overline{K}(x_i,x)\right\|$$
$$\leqslant b_2^2 + e_2^2 + d\sum_{i=1}^{s} \left\|\nabla_1\nabla_2^2\overline{K}(x_i,x)\right\| \quad \text{since} \quad \left\|\partial_{1,j}\nabla_2^2\overline{K}\right\| \leqslant \left\|\nabla_1\nabla_2^2\overline{K}\right\|$$
$$\leqslant dB_2^2 \tag{45}$$

Appendix F. Proof of Theorem 8: vanishing derivative pre-certificate

In this section we rename the vanishing derivative pre-certificate $\eta_V = \alpha^\top \mathbf{f}(\cdot)$ simply η for the sake of shortness.

By Theorem 18, we know that the function $\overline{\eta} = \overline{\alpha}^\top \overline{\mathbf{f}}$ is non-degenerate. Our goal is to show that the vanishing derivative pre-certificate η is sufficiently close to $\overline{\eta}$ to keep this property.

Next we define appropriate neighborhoods of the x_i's and grids. For $0 \leqslant r \leqslant 2$, define the following constants:

$$M_r \overset{\text{def.}}{=} 4sL_{01}L_{r+1} \tag{46}$$

Define $\delta_{\eta,2} \overset{\text{def.}}{=} \lambda_\eta/(4M_2)$ and $\mathcal{X}_{\text{grid}}^{\text{near}}(j)$ a $\delta_{\eta,2}$-covering of $\mathcal{X}_j^{\text{near}}$ (for the Euclidean norm), of size $N_{\text{near}} \leqslant (1 + 4\varepsilon_{\text{near}}/\delta_{\eta,2})^d$ (a classical bound for covering number of balls, see (Gribonval et al., 2017)), and $\mathcal{X}_{\text{grid}}^{\text{near}} = \bigcup_j \mathcal{X}_{\text{grid}}^{\text{near}}(j)$. Define $\delta_{\eta,0} \overset{\text{def.}}{=} \varepsilon_\eta/(4M_0)$ and $\mathcal{X}_{\text{grid}}^{\text{far}}$ a $\delta_{\eta,0}$-covering of $\mathcal{X}^{\text{far}}$, of size $N_{\text{far}} \leqslant (1 + 4B_{\mathcal{X}}/\delta_{\eta,0})^d$.

F.1 Sufficient bounds

The following Lemma gathers all the sufficient conditions that we will then aim to prove.

Lemma 19 *Assume that the following hold:*

$$\left\|\Upsilon^{-1}\right\| \leqslant 4 \tag{47}$$
$$\forall x_{grid} \in \mathcal{X}_{grid}^{far}, \ \ |\overline{\eta}(x_{grid}) - \eta(x_{grid})| \leqslant \frac{\varepsilon_\eta}{4} \tag{48}$$
$$\forall x_{grid} \in \mathcal{X}_{grid}^{near}, \ \ \left\|\nabla^2\overline{\eta}(x_{grid}) - \nabla^2\eta(x_{grid})\right\| \leqslant \frac{\lambda_\eta}{4} \tag{49}$$

Then, the certificate η is non-degenerate. The constant 4 in (46) and (47) has been chosen for simplicity.

Proof Under (47), for $0 \leqslant r \leqslant 2$, it is immediate to see that $\nabla^r \eta$ is M_r-Lipschitz:

$$\begin{aligned}
\left\|\nabla^r \eta(x) - \nabla^r \eta(x')\right\| &= \left\| \frac{1}{m} \sum_{k=1}^m \nabla^r \mathrm{Re}\left((\Upsilon^{-1}\mathbf{u}_k)^\top \gamma(\omega_k)\varphi_{\omega_k}(x)\right) - \nabla^r \mathrm{Re}\left((\Upsilon^{-1}\mathbf{u}_s)^\top \gamma(\omega_k)\varphi_{\omega_k}(x')\right)\right\| \\
&= \left\| \frac{1}{m} \sum_{j=1}^m \mathrm{Re}\left(\left((\Upsilon^{-1}\mathbf{u}_s)^\top \gamma(\omega_k)\right) \cdot \left(\nabla^r \varphi_{\omega_k}(x) - \nabla^r \varphi_{\omega_k}(x')\right)\right)\right\| \\
&\leqslant \left\|\Upsilon^{-1}\right\| \left\|\mathbf{u}_s\right\| \sqrt{s} L_{01} L_{r+1} \left\|x - x'\right\|_2 \\
&\leqslant 4 s L_{01} L_{r+1} \left\|x - x'\right\|_2 \qquad (50)
\end{aligned}$$

Next we prove that, inside each $\mathcal{X}_i^{\text{near}}$, $-\operatorname{sign}(a_i)\nabla^2 \eta$ is positive definite: indeed, for any $x \in \mathcal{X}_i^{\text{near}}$, pick a $x_{\text{grid}} \in \mathcal{X}_{\text{grid}}^{\text{near}}(i)$ that is closest to x, and we have

$$\begin{aligned}
&\operatorname{sign}(a_i)\nabla^2\eta(x) \\
&= \operatorname{sign}(a_i)\left(\nabla^2\eta(x) - \nabla^2\eta(x_{\text{grid}})\right) + \operatorname{sign}(a_i)\left(\nabla^2\eta(x_{\text{grid}}) - \nabla^2\overline{\eta}(x_{\text{grid}})\right) + \operatorname{sign}(a_i)\nabla^2\overline{\eta}(x_{\text{grid}}) \\
&\overset{(49),\ (50)}{\preccurlyeq} \quad (M_3\delta_{\eta,2} + \lambda_\eta/4)\,\mathrm{Id} + \operatorname{sign}(a_i)\nabla^2\overline{\eta}(x_{\text{grid}}) = \frac{\lambda_\eta}{2}\mathrm{Id} + \operatorname{sign}(a_i)\nabla^2\overline{\eta}(x_{\text{grid}}) \prec 0
\end{aligned}$$

since $\frac{\lambda_\eta}{2}\mathrm{Id} \preccurlyeq \frac{\lambda_{\min}\left(-\operatorname{sign}(a_i)\nabla^2\overline{\eta}(x_{\text{grid}})\right)}{2}\mathrm{Id} \prec -\operatorname{sign}(a_i)\nabla^2\overline{\eta}(x_{\text{grid}})$. Therefore, for all i, $-\operatorname{sign}(a_i)\nabla^2\eta(x)$ is positive definite on $\mathcal{X}_i^{\text{near}}$, which proves that it is *a fortiori* positive definite in x_i, and that $|\eta(x)| < 1$ on $\mathcal{X}_i^{\text{near}} \backslash \{x_i\}$.

With a similar strategy, we now prove that $|\eta(x)| < 1$ on $\mathcal{X}^{\text{far}}$. For any $x \in \mathcal{X}^{\text{far}}$, there is an $x_{\text{grid}} \in \mathcal{X}_{\text{grid}}^{\text{far}}$ such that $|x - x_{\text{grid}}| \leqslant \delta_{\eta,0}$, and thus:

$$\begin{aligned}
|\eta(x)| &\leqslant |\overline{\eta}(x_{\text{grid}})| + |\overline{\eta}(x_{\text{grid}}) - \eta(x_{\text{grid}})| + |\eta(x_{\text{grid}}) - \eta(x)| \\
&\leqslant 1 - \varepsilon_\eta + \frac{\varepsilon_\eta}{4} + M_0\delta_{\eta,0} = 1 - \frac{\varepsilon_\eta}{2} < 1
\end{aligned}$$

Hence η is non-degenerate. ∎

We must therefore control the deviation between $\overline{\eta}$ and η on $\mathcal{X}_{\text{grid}}^{\text{far}}$ and that between $\nabla^2\overline{\eta}$ and $\nabla^2\eta$ on $\mathcal{X}_{\text{grid}}^{\text{near}}$. We decompose this deviation into two terms.

$$\overline{\eta}(x) - \eta(x) = \left(\alpha^\top\left(\overline{\mathbf{f}}(x) - \mathbf{f}(x)\right)\right) + \left((\overline{\alpha} - \alpha)^\top \overline{\mathbf{f}}(x)\right) = E_1(x) + E_2(x) \qquad (51)$$

We bound them individually in the next sections, starting with a bound between $\overline{\Upsilon}^{-1}$ and Υ^{-1}. The bound on E_2 will depend if we assume random signs or not.

F.2 Bound on $\overline{\Upsilon}$

We first prove that Υ^{-1} is close to $\overline{\Upsilon}^{-1}$ with high probability.

Using Lemma 13, with high probability Υ is close to $\overline{\Upsilon}$, and since $\overline{\Upsilon}$ is close to identity, their inverses are also close, as gathered in the following lemma.

Lemma 20 *Assume* Υ *has been drawn such that* $\left\|\overline{\Upsilon} - \Upsilon\right\| \leqslant \varepsilon_\Upsilon$ *for some* $\varepsilon_\Upsilon \leqslant \frac{1}{4}$. *Then,* Υ *is invertible, and we have the following:*

$$\left\|\Upsilon^{-1}\right\| \leqslant 4 \tag{52}$$

$$\left\|\Upsilon^{-1} - \overline{\Upsilon}^{-1}\right\| \leqslant 8\varepsilon_\Upsilon \tag{53}$$

Proof Since we have $\varepsilon_\Upsilon \leqslant 1/4$ and $\left\|I - \overline{\Upsilon}\right\| \leqslant 1/2$ by (34), we have

$$\|I - \Upsilon\| \leqslant \left\|I - \overline{\Upsilon}\right\| + \varepsilon_\Upsilon \leqslant 1/2 + 1/4 \leqslant 3/4 \tag{54}$$

Therefore Υ is invertible, and $\left\|\Upsilon^{-1}\right\| \leqslant \frac{1}{1-\|I-\Upsilon\|} \leqslant 4$. Then, it holds that

$$\left\|\Upsilon^{-1} - \overline{\Upsilon}^{-1}\right\| \leqslant \left\|\Upsilon^{-1}\right\| \left\|\overline{\Upsilon} - \Upsilon\right\| \left\|\overline{\Upsilon}^{-1}\right\| \leqslant 8\varepsilon_\Upsilon$$

■

F.3 Bound on E_1

Let us now bound $E_1(x)$ and $\nabla^2 E_1(x)$, conditionally on Υ being close to $\overline{\Upsilon}$.

Lemma 21 *Assume* Υ *is fixed such that* (47) *is satisfied. Then:*

1. *We have*

$$\forall x \in \mathcal{X}^{far},\ \mathbb{P}\left(|E_1(x)| \geqslant \frac{\varepsilon_\eta}{8}\right) \leqslant \rho \tag{55}$$

   if

$$m \gtrsim s \cdot \left(\left(\frac{L_0}{\varepsilon_\eta}\right)^2 + \frac{L_0}{\varepsilon_\eta} L_{01}\right) \log\left(\frac{1}{\rho}\right)$$

2. *We have*

$$\forall x \in \mathcal{X}^{near},\ \mathbb{P}\left(\left\|\nabla^2 E_1(x)\right\| \geqslant \frac{\lambda_\eta}{8}\right) \leqslant \rho \tag{56}$$

   if

$$m \gtrsim s \cdot \left(\left(\frac{L_2}{\lambda_\eta}\right)^2 + \frac{L_2}{\lambda_\eta} L_{01}\right) \log\left(\frac{d}{\rho}\right)$$

Proof The result is a consequence of Corollary 16 and Prop. 17, by taking q as $\alpha = \Upsilon^{-1}\mathbf{u}_s$, with $\|q\| \leqslant 4\sqrt{s}$. ■

F.4 Bound on E_2

Simiarly, we bound $E_2(x)$ and $\nabla^2 E_2(x)$ conditionally on Υ being close to $\overline{\Upsilon}$, depending if the random signs assumption holds or not.

Lemma 22 *Assume Υ is fixed, such that $\left\|\overline{\Upsilon} - \Upsilon\right\| \leqslant \varepsilon_\Upsilon$ for some $\varepsilon_\Upsilon \leqslant \frac{1}{4}$ is satisfied. Then:*

1. *If the random signs assumption does not hold, we have uniformly:*

$$\forall x \in \mathcal{X}^{far}, \ \ |E_2(x)| \leqslant 8\sqrt{k}\varepsilon_\Upsilon B_0 \tag{57}$$

and

$$\forall x \in \mathcal{X}^{near}, \ \ \left\|\nabla^2 E_2(x)\right\| \leqslant 8\sqrt{k}\varepsilon_\Upsilon B_2 \tag{58}$$

2. *If the random sign assumption holds, we have*

$$\forall x \in \mathcal{X}^{far}, \ \mathbb{P}\left(|E_2(x)| \geqslant \frac{\varepsilon_\eta}{8}\right) \leqslant \exp\left(-\frac{\varepsilon_\eta^2}{256\varepsilon_\Upsilon^2 B_0^2}\right) \tag{59}$$

and

$$\forall x \in \mathcal{X}^{near}, \ \mathbb{P}\left(\left\|\nabla^2 E_2(x)\right\| \geqslant \frac{\lambda_\eta}{8}\right) \leqslant 4d^2 \exp\left(-\frac{\lambda_\eta^2}{256 d\varepsilon_\Upsilon^2 B_2^2}\right) \tag{60}$$

Proof Let us start with the case where we do not assume random signs. In that case, for any $x \in \mathcal{X}_{\text{far}}$, we have

$$|E_2(x)| = \left|\mathbf{u}_k(\overline{\Upsilon} - \Upsilon)\overline{\mathbf{f}}(x)\right| \overset{(42)}{\leqslant} 8\sqrt{k}\varepsilon_\Upsilon B_0$$

and for any $x \in \mathcal{X}_{\text{near}}$

$$\left\|\nabla^2 E_2(x)\right\| = \left\|\sum_{l=1}^{Q} (\overline{\alpha}_l - \alpha_l)\, \nabla^2 f_l(x)\right\| \overset{(44)}{\leqslant} \|\overline{\alpha} - \alpha\|\, B_2 \leqslant 8\sqrt{k}\varepsilon_\Upsilon B_2$$

Let us now turn to the case with random signs. Using Lemma 12 (noting that it is valid even if some components of $\mathbf{u}_k$ are zeros), for any $x \in \mathcal{X}_{\text{far}}$ we have

$$\mathbb{P}\left(\left|\mathbf{u}_k^\top(\overline{\Upsilon}^{-1} - \Upsilon^{-1})\overline{\mathbf{f}}(x)\right| \geqslant \frac{\varepsilon_\eta}{8}\right) \leqslant \exp\left(-\frac{\varepsilon_\eta^2/64}{4\left\|(\overline{\Upsilon}^{-1} - \Upsilon^{-1})\overline{\mathbf{f}}(x)\right\|^2}\right) \leqslant \exp\left(-\frac{\varepsilon_\eta^2}{256\varepsilon_\Upsilon^2 B_0^2}\right)$$

which is the desired bound.

Let us turn to the bound on the Hessian matrix. Take any $x \in \mathcal{X}_{\text{near}}$. Using Lemma 12 with a union bound, we have that with probability at least $1 - \rho$,

$$\forall\ 1 \leqslant i,j \leqslant d, \ \left|\mathbf{u}_k^\top(\overline{\Upsilon}^{-1} - \Upsilon^{-1})\partial_{ij}\overline{\mathbf{f}}(x)\right| \leqslant 2\varepsilon_\Upsilon \left\|\partial_{ij}\overline{\mathbf{f}}(x)\right\| \sqrt{\log\left(\frac{4d^2}{\rho}\right)}$$

And therefore with the same probability we have

$$
\begin{aligned}
\left\|\nabla^2 E_2(x)\right\| \leqslant \left\|\nabla^2 E_2(x)\right\|_\infty &\leqslant \max_i \left(\sum_{j=1}^d 4\varepsilon_\Upsilon^2 \left\|\partial_{ij}\overline{\mathbf{f}}(x)\right\|^2 \log\left(\frac{4d^2}{\rho}\right)\right)^{\frac{1}{2}} \\
&= \max_i 2\varepsilon_\Upsilon \left(\sum_{j=1}^d \sum_{l=1}^Q |\partial_{ij} f_l(x)|^2\right)^{\frac{1}{2}} \sqrt{\log\left(\frac{4d^2}{\rho}\right)} \\
&= 2\varepsilon_\Upsilon \sqrt{\sum_{l=1}^Q \|\nabla^2 f_l(x)\|^2} \sqrt{\log\left(\frac{4d^2}{\rho}\right)} \leqslant 2\varepsilon_\Upsilon \sqrt{d} B_2 \sqrt{\log\left(\frac{4d^2}{\rho}\right)}
\end{aligned}
$$

whee we have used the fact that $\|\partial_i \nabla f_l\|^2 \leqslant \left\|\nabla^2 f_l\right\|^2$.

Putting this bound to $\lambda_\eta/8$ and computing ρ we obtain the desired result. ■

F.5 Summary

Let us now summarize the results to obtain the bound on m. First, using Lemma 13, we obtain that with probability at least $1-\rho_{\overline{\Upsilon}}$, we have

$$
\left\|\Upsilon - \overline{\Upsilon}\right\| \leqslant \varepsilon_\Upsilon := \sqrt{\frac{7sL_{01}^2}{m} \log\left(\frac{2s(d+1)}{\rho_1}\right)} \tag{61}
$$

if m is sufficiently big such that $\varepsilon_\Upsilon \leqslant 1/4$, i.e.

$$
m \gtrsim sL_{01}^2 \log \frac{sd}{\rho_1} \tag{62}
$$

Without random signs. When we do not assume random signs, using Lemma 21 and a union bound with (61) we see that $|E_1(x)| \leqslant \varepsilon_\eta/8$ for all $x \in \mathcal{X}_{\text{grid}}^{\text{far}}$ and $\left\|\nabla^2 E_1(x)\right\| \leqslant \lambda_\eta/8$ for all $x \in \mathcal{X}_{\text{grid}}^{\text{near}}$ with probability $1-\rho_1-\rho_2-\rho_3$ if

$$
m \gtrsim s \max\{\frac{L_0}{\varepsilon_\eta}\cdot\left(\frac{L_0}{\varepsilon_\eta}+L_{01}\right)\cdot\log\left(\frac{N_{\text{far}}}{\rho_2}\right), \frac{L_2}{\lambda_\eta}\cdot\left(\frac{L_2}{\lambda_\eta}+L_{01}\right)\cdot\log\left(\frac{dN_{\text{near}}}{\rho_3}\right)\} \tag{63}
$$

Then, using Lemma 22, we see that $|E_2(x)| \leqslant \varepsilon_\eta/8$ for all $x \in \mathcal{X}_{\text{far}}^{\text{grid}}$ and $\left\|\nabla^2 E_2(x)\right\| \leqslant \lambda_\eta/8$ for all $x \in \mathcal{X}_{\text{near}}^{\text{grid}}$ are immediately satisfied as soon as

$$
\varepsilon_\Upsilon \leqslant \frac{1}{64\sqrt{s}} \min\{\frac{\varepsilon_\eta}{B_0}, \frac{\lambda_\eta}{B_2}\}
$$

which happens when

$$
m \gtrsim s^2 L_{01}^2 \max\{\frac{B_0^2}{\varepsilon_\eta^2},\ \frac{B_2^2}{\lambda_\eta^2}\} \log\frac{sd}{\rho_1} \tag{64}
$$

Therefore, combining (63), (64) (the latter being strictly stronger than (62)) and using the expressions for the covering numbers N_{near} and N_{far} we obtain the desired bound.

With random signs. Under the random signs assumption, the bound (63) is still valid for the bound on E_1, and by applying Lemma 22 with ε_Υ given by (61)we have $|E_2(x)| \leqslant \varepsilon_\eta/8$ for all $x \in \mathcal{X}^{\text{far}}_{\text{grid}}$ and $\left\|\nabla^2 E_2(x)\right\| \leqslant \lambda_\eta/8$ for all $x \in \mathcal{X}^{\text{near}}_{\text{grid}}$ with probability $1-\rho_1-\rho_1'-\rho_1''$ if

$$m \gtrsim sL_{01}^2 \log\left(\frac{sd}{\rho_1}\right) \max\{\frac{B_0^2}{\varepsilon_\eta^2}\log\left(\frac{N_{\text{far}}}{\rho_1'}\right), \frac{dB_2^2}{\lambda_\eta^2}\log\left(\frac{N_{\text{near}}}{\rho_1''}\right)\} \tag{65}$$

Using this equation with (63) we obtain the final bound.

Appendix G. Proof of Theorem 7: an infinite dimensional golfing scheme

Define $C_0 \stackrel{\text{def.}}{=} 2\sqrt{d}B_0$, $C_2 \stackrel{\text{def.}}{=} 2\sqrt{d}B_2$, such that by (43) and (44) we have: in $\mathcal{X}^{\text{far}}$

$$\left|q^\top \overline{\mathbf{f}}(x)\right| \leqslant C_0 \left\|q\right\|_\infty$$

and in $\mathcal{X}^{\text{near}}$

$$\left\|\nabla^2\left(q^\top \overline{\mathbf{f}}(x)\right)\right\| \leqslant C_2 \left\|q\right\|_\infty$$

As before, we are going to construct a certificate that satisfies the right properties on a dense grid, and interpolate by bounding the Lipschitz constant on the constructed certificate and its second derivative. We thus define $\mathcal{X}^{\text{near}}_{\text{grid}}$ (resp. $\mathcal{X}^{\text{far}}_{\text{grid}}$) a δ_{near}-covering (resp. δ_{far}-covering) of $\mathcal{X}^{\text{near}}$ (resp. $\mathcal{X}^{\text{far}}$), with $\delta_{\text{near}} = \frac{\lambda_\eta}{16L_3L_{01}\sqrt{sd}}$ and $\delta_{\text{far}} = \frac{\varepsilon_\eta}{16L_1L_{01}\sqrt{sd}}$.

The construction of this certificate is done in two steps: first construct an approximate certificate by the golfing scheme, then correct this certificate by adding a small perturbation.

G.1 Step I: The golfing scheme

We follow the golfing scheme presented in Candes and Plan (2011).

Define the parameters

$$L = \left\lceil \max\{2,\ \log\left(\tfrac{32sL_0L_{01}\sqrt{d+1}}{\varepsilon_\eta C_0^2}\right),\ \log\left(\tfrac{32sL_2L_{01}\sqrt{d+1}}{\lambda_\eta C_0^2}\right)\}\right\rceil \quad \text{and} \quad L' = \left\lceil 3L + \frac{1}{2}\log\left(\frac{4}{\rho}\right)\right\rceil,$$

$$c_1 = c_2 = \frac{\delta}{C_0\sqrt{\log_2(s(d+1))}}, \quad \text{and} \quad c_i = \delta, \quad i \geqslant 3,$$

$$t_1 = 1 - \frac{\varepsilon_\eta}{2}, \quad t_2 = 4C_0, \quad \text{and} \quad t_i = 4C_0\log_2(s(d+1)), \quad i \geqslant 3,$$

$$b_1 = \lambda_\eta/2,\ b_2 = 4C_2\sqrt{\log(s(d+1))} \quad \text{and} \quad b_i = 4C_2\log(s(d+1)), \quad i \geqslant 3$$

where, $\delta = \min\{\frac{\varepsilon_\eta}{32}, \frac{B_0\lambda_\eta}{32B_2}, e^{-1}\}$. With our choice of L and δ, one can easily check that we have:

$$1 - \frac{\varepsilon_\eta}{2} + 4\left(\frac{\delta}{1-\delta} + \frac{\delta^L L_0L_{01}s\sqrt{d+1}}{C_0^2}\right) \leqslant 1 - \frac{\varepsilon_\eta}{4}$$

and

$$-\frac{\lambda_\eta}{2} + 4\left(\frac{\delta B_2}{(1-\delta)B_0} + \frac{\delta^L L_2L_{01}s\sqrt{d+1}}{C_0^2}\right) \leqslant -\frac{\lambda_\eta}{4}$$

Divide the indices $\{1,\ldots,m\}$ into L' blocks B_l of size m_l (whose exact size will be determined later) for $l = 1,\ldots,L'$. As detailed in Section D, we denote Υ_l and $\mathbf{f}_l$ the respective empirical version of $\overline{\Upsilon}$ and $\overline{\mathbf{f}}$ over the m_l frequencies included in B_l.

The golfing construction For some $L' > L$, define index sets S_j, vectors $q_j \in \mathbb{C}^{(d+1)s}$ and functions $\eta^j \in \mathscr{C}(X)$ by: $S_1 = \{1\}$, $S_2 = \{2\}$, $q_0 = \mathbf{u}_s$, $\eta^0 = 0$, and

$$2 \leqslant i \leqslant L', \qquad S_i = \begin{cases} S_{i-1} \cup \{i\} & \text{if event } E_i(q_{i-1}) \text{ occurs,} \\ S_{i-1} & \text{otherwise,} \end{cases}$$

$$1 \leqslant i \leqslant L', \qquad \eta^i = \begin{cases} \sum_{j\in S_i} \bar{q}_{j-1}^\top \mathbf{f}_j(\cdot) & \text{if } i \in S_i, \\ \eta^{i-1} & \text{otherwise,} \end{cases}$$

$$1 \leqslant i \leqslant L', \qquad q_i = \begin{cases} \mathbf{u}_s - \sum_{j\in S_i} \Upsilon_j \bar{q}_{j-1} & \text{if } i \in S_i, \\ q_{i-1} & \text{otherwise,} \end{cases}$$

where $\bar{q}_i = \overline{\Upsilon}^{-1} q_i$ and the event $E_i(q_{i-1})$ is said to occur if the following hold:

(I) $\left\|(\mathrm{Id} - \Upsilon_i \overline{\Upsilon}^{-1}) q_{i-1}\right\|_\infty \leqslant c_i \left\|q_{i-1}\right\|_\infty$

(II) $\forall x \in \mathcal{X}_{\mathrm{grid}}^{\mathrm{far}}$, $\left|\bar{q}_{i-1}^\top \mathbf{f}_i(x)\right| \leqslant t_i \left\|q_{i-1}\right\|_\infty$

(III) $\forall x \in \mathcal{X}_{\mathrm{grid}}^{\mathrm{near}}$, $\left\|\nabla^2 \left(\bar{q}_{i-1}^\top \mathbf{f}_i(x)\right)\right\| \leqslant b_i \left\|q_{i-1}\right\|_\infty$

where, if $i = 1$, then we replace (III) by

$$\mathrm{sign}(a_j) \nabla^2 \left((\overline{\Upsilon}^{-1} \mathbf{u}_s)^\top \mathbf{f}_1(x) \right) \preceq -b_1 \mathrm{Id} \qquad \forall x \in \mathcal{X}_{\mathrm{grid}}^{\mathrm{near}}(j),\ j = 1, \ldots, s. \tag{66}$$

Note that, by (22), we have the following relation between q_i and η^i:

$$q_i = \mathbf{u}_s - D\Psi(\eta^i) \tag{67}$$

where we recall that Ψ is the operator that computes the values of a function and its gradient at all x_i.

Therefore q_i represents the *error* between the value of η^i at each x_l and $\mathrm{sign}(a_l)$ and the deviation of its gradient from zero. At each golfing step, this error is reduced by assumption (I): indeed, for two consecutive elements $\tau(l-1), \tau(l)$ in S_i, it is easy to check that by definition we have

$$q_{\tau(l)} = \left(\mathrm{Id} - \Upsilon_{\tau(l)} \overline{\Upsilon}^{-1}\right) q_{\tau(l-1)} \tag{68}$$

This geometric progression of the error is the key to the golfing scheme.

We set the final inexact dual certificate to be $\eta^{\mathrm{app}} \stackrel{\text{def.}}{=} \eta^{L'}$.

Lemma 23 *Suppose that*

$$m \gtrsim s \cdot \log\left(\tfrac{1}{\rho}\right) \cdot \left(d^2 \left(\tfrac{L_{01} B_0}{\delta}\right)^2 \log(sd)^2 + \mathcal{L}_0 \log\left(\left|\mathcal{X}_{\mathrm{grid}}^{\mathrm{far}}\right|\right) + \mathcal{L}_2 \log\left(d \left|\mathcal{X}_{\mathrm{grid}}^{\mathrm{near}}\right|\right)\right) \tag{69}$$

where

$$\mathcal{L}_0 = \max\{\tfrac{dL_0^2}{\varepsilon_\eta^2}, \tfrac{\sqrt{d} L_0 L_{01}}{\varepsilon_\eta}\} + \log(sd) \max\left(\tfrac{L_0^2}{B_0^2}, \tfrac{L_0 L_{01}}{B_0}\right)$$
$$\mathcal{L}_2 = \max\{\tfrac{dL_2^2}{\lambda_\eta^2}, \tfrac{\sqrt{d} L_2 L_{01}}{\lambda_\eta}\} + \log(sd) \max\left(\tfrac{L_2^2}{B_2^2}, \tfrac{L_2 L_{01}}{B_2}\right)$$

Then with probablity at least $1 - \rho$, the approximate certificate η^{app} satisfies $\eta \in \mathrm{Im}(\Phi^*)$ such that

- $\|D\Psi\eta^{\text{app}} - \mathbf{u}_s\|_2 \leqslant \frac{\delta^L\sqrt{s(d+1)}}{C_0^2}$,
- *for all* $x \in \mathcal{X}_{\text{grid}}^{\text{far}}$, $|\eta(x)| \leqslant 1 - \frac{\varepsilon_\eta}{2} + 4\frac{\delta}{(1-\delta)}$,
- *for all* $x \in \mathcal{X}_{\text{grid}}^{\text{near}}(j)$, $\text{sign}(a_j)\nabla^2\eta(x) \preceq \left(-\frac{\lambda_\eta}{2} + \frac{4\delta C_2}{(1-\delta)C_0}\right)\text{Id}$.

Proof Enumerating the elements of $S_{L'}$ by $\tau(1), \tau(2), \cdots$, we have seen that by definition (see (68)) we have

$$q_{\tau(i)} = \left(\text{Id} - \Upsilon_i \overline{\Upsilon}^{-1}\right) q_{\tau(i-1)}.$$

and by (67)

$$e \overset{\text{def.}}{=} q_{\tau(|S_{L'}|)} = \mathbf{u}_s - D\Psi\eta^{\text{app}}$$

since $\eta^{\text{app}} = \eta^{L'}$ in the golfing scheme. In the event that $|S_{L'}| \geqslant L$ and events $E_1(q_0)$ and $E_2(q_1)$ hold, we have the bounds:

$$\|e\| \leqslant \sqrt{(d+1)s}\,\|q_\tau(|S_{L'}|)\|_\infty \leqslant \sqrt{s(d+1)}\prod_{j=1}^{L} c_j \overset{\text{(I)}}{\leqslant} \frac{\delta^L\sqrt{s(d+1)}}{C_0^2}, \tag{70}$$

and for all $x \in \mathcal{X}_{\text{grid}}^{\text{far}}$,

$$\begin{aligned}|\eta^{\text{app}}(x)| &\leqslant \sum_{i=1}^{L}\left|(\overline{\Upsilon}^{-1}q_{i-1})^\top \mathbf{f}_i(x)\right| \overset{\text{(II)}}{\leqslant} \sum_{i=1}^{L} t_i\,\|q_{i-1}\|_\infty \leqslant \sum_{i=1}^{L} t_i \prod_{j=1}^{i-1} c_i \\ &= 1 - \frac{\varepsilon_\eta}{2} + 4\frac{\delta}{\sqrt{\log_2(s(d+1))}} + 4\frac{\delta^2}{C_0} + 4\frac{\delta^3}{C_0} + \cdots < 1 - \frac{\varepsilon_\eta}{2} + 4\frac{\delta}{(1-\delta)}. \end{aligned} \tag{71}$$

and finally for all $x \in \mathcal{X}_{\text{grid}}^{\text{near}}(j)$,

$$\begin{aligned}&\text{sign}(a_j)(\nabla^2\eta^{\text{app}})(x) \preceq \text{sign}(a_j)\nabla^2\left((\overline{\Upsilon}^{-1}\mathbf{u}_s)^\top \mathbf{f}_1(x)\right) + \sum_{i=2}^{L}\left\|\nabla^2\left((\overline{\Upsilon}^{-1}q_{i-1})^\top \mathbf{f}_i(x)\right)\right\|\text{Id} \\ &\overset{\text{(III)}}{\preceq} -b_1\text{Id} + \sum_{i=2}^{L} b_i \prod_{j=1}^{i-1} c_i \text{Id} \preceq \left(-\frac{\lambda_\eta}{2} + 4C_2\frac{\delta}{C_0} + 4C_2\frac{\delta^2}{C_0^2} + 4C_2\frac{\delta^3}{C_0^2} + \cdots\right)\text{Id} \\ &\preceq \left(-\frac{\lambda_\eta}{2} + \frac{4\delta C_2}{(1-\delta)C_0}\right)\text{Id}.\end{aligned} \tag{72}$$

It remains to lower bound $\mathbb{P}\left[|S_{L'}| \geqslant L \text{ and } E_1(q_0) \text{ and } E_2(q_1)\right]$: By the union bound

$$\mathbb{P}\left[|S_{L'}| \geqslant L \text{ and } E_1(q_0) \text{ and } E_2(q_1)\right] \geqslant 1 - \mathbb{P}[|S_{L'}| > L] - \mathbb{P}[\neg E_1(q_0)] - \mathbb{P}[\neg E_2(q_1)],$$

therefore, it remains to show that $\mathbb{P}(|S_{L'}| > L) \leqslant \rho/3$ and $\mathbb{P}(\neg E_i(q_{i-1})) \leqslant \rho/3$ for $i = 1, 2$. From Adcock et al. (2017), by defining the random variables $X_j = \begin{cases} 0 & q_{j+2} \neq q_{j+1} \\ 1 & \text{otherwise,} \end{cases}$ we have that

$$\mathbb{P}(|S_{L'}| < L) \leqslant \mathbb{P}(X_1 + \cdots + X_{L'-2} > L' - L) \leqslant \rho/3$$

provided that $L' \geqslant 8\lceil 3L + \frac{1}{2}\log(\frac{3}{\rho})\rceil$ and

$$\frac{1}{4} \geqslant \mathbb{P}(X_{\pi(j)} = 1 | X_{\pi(l)} = 1,\ l < j), \qquad \forall\{\pi(1) < \pi(2) < \cdots < \pi(l)\} \subset \{1, \ldots, L'-2\} \tag{73}$$

Now, for $i \geqslant 2$, since we have $t_i \geqslant 4C_0$, then for all $x \in \mathcal{X}^{\text{far}}$:

$$\left|\bar{q}_{i-1}^\top \bar{\mathbf{f}}(x)\right| \leqslant (t_i/4) \left\|\bar{q}_{i-1}\right\|_\infty \tag{74}$$

and similarly for $x \in \mathcal{X}^{\text{near}}$ and $i \geqslant 2$:

$$\left\|\nabla^2 \left(\bar{q}_{i-1}^\top \bar{\mathbf{f}}(x)\right)\right\| \leqslant (b_i/4) \left\|\overline{\Upsilon}^{-1} q_{i-1}\right\|_\infty$$

Since the assumptions of Theorem 18 hold and therefore $\left\|\overline{\Upsilon}^{-1}\right\|_\infty \leqslant 2$ (see (41)), (I), (II) and (III) hold if

(I') $\left\|(\overline{\Upsilon} - \Upsilon_i)\bar{q}_{i-1}\right\|_\infty \leqslant \frac{c_i}{2} \left\|\bar{q}_{i-1}\right\|_\infty$

(II') $\forall\, x \in \mathcal{X}^{\text{far}}_{\text{grid}}, \quad \left|\bar{q}_{i-1}^\top(\mathbf{f}_i(x) - \bar{\mathbf{f}}(x))\right| \leqslant r_i \left\|\bar{q}_{i-1}\right\|_\infty$ where $r_1 = \frac{\varepsilon_\eta}{4}$ and $r_i = t_i/4$ for $i \geqslant 2$

(III') $\forall\, x \in \mathcal{X}^{\text{near}}_{\text{grid}}, \quad \left\|\nabla^2 \left(\bar{q}_{i-1}^\top(\mathbf{f}_i(x) - \bar{\mathbf{f}}(x))\right)\right\| \leqslant \frac{b_i}{4} \left\|\bar{q}_{i-1}\right\|_\infty$

.

where the implication "(II') implies (II)" is valid since:

$$\begin{aligned}\left|(\overline{\Upsilon}^{-1} q_{i-1})^\top \mathbf{f}_i(x)\right| &\leqslant \left|(\overline{\Upsilon}^{-1} q_{i-1})^\top \bar{\mathbf{f}}(x)\right| + \left|(\overline{\Upsilon}^{-1} q_{i-1})^\top (\mathbf{f}_i(x) - \bar{\mathbf{f}}(x))\right| \\ &\overset{(74),\ (\text{II'}),\text{Thm. } 18}{\leqslant} \begin{cases} 1 - \varepsilon_\eta + r_1 \left\|\overline{\Upsilon}^{-1}\right\|_\infty \leqslant t_1 \text{ for the case } i = 1 \\ 2 \cdot \frac{t_i}{4} \left\|\overline{\Upsilon}^{-1} q_{i-1}\right\|_\infty \leqslant t_i \left\|q_{i-1}\right\|_\infty, \text{ for the case } i \geqslant 2 \end{cases}\end{aligned}$$

and the implication "(III') implies (III)" is valid for the same reasons, by noting that $-\lambda_\eta + b_1/4 \leqslant -b_1$ for the case $i = 1$.

We have:

1. By corollary 15, (I') holds with probability at least ρ if

$$m_i \gtrsim sd \cdot c_i^{-2} \cdot L_{01}^2 \log\left(\frac{sd}{\rho}\right)$$

2. by corollary 16, (II') holds with probability at least ρ if

$$m_i \gtrsim s \cdot \frac{\sqrt{d}L_0}{r_i}\left(\frac{\sqrt{d}L_0}{r_i} + L_{01}\right) \cdot \log\left(\frac{\left|\mathcal{X}^{\text{far}}_{\text{grid}}\right|}{\rho}\right)$$

3. finally, by Proposition 17, (III') holds with probability at least ρ if

$$m_i \gtrsim s \cdot \frac{\sqrt{d}L_2}{b_i}\left(\frac{\sqrt{d}L_2}{b_i} + L_{01}\right) \cdot \log\left(\frac{d\left|\mathcal{X}^{\text{near}}_{\text{grid}}\right|}{\rho}\right)$$

So, (73) holds if for $i > 2$:

$$m_i \gtrsim s \cdot \left(d \left(\tfrac{L_{01}}{\delta}\right)^2 \log(sd) + \sum_{i \in \{0,2\}} \mathcal{L}_i \tfrac{\log(N_i)}{\log(sd)} \right) \tag{75}$$

where $\mathcal{L}_i = \max\{\frac{L_i^2}{B_i^2}, \frac{L_i L_{01}}{B_i}\}$, $N_0 = \left|\mathcal{X}_{\text{grid}}^{\text{far}}\right|$, $N_2 = d\left|\mathcal{X}_{\text{grid}}^{\text{near}}\right|$. Moreover, $\mathbb{P}(\neg E_i(q_{i-1})) \leqslant \rho/3$ for $i = 1, 2$ hold if

$$m_1, m_2 \gtrsim s \cdot \left(d^2 \left(\tfrac{L_{01} B_0}{\delta}\right)^2 \log(sd) \log\left(\tfrac{sd}{\rho}\right) + \sum_{i \in \{0,2\}} \mathcal{L}'_i \log\left(\tfrac{N_i}{\rho}\right) \right) \tag{76}$$

where $\mathcal{L}'_0 = \max\{\frac{dL_0^2}{\varepsilon_\eta^2}, \frac{\sqrt{d} L_0 L_{01}}{\varepsilon_\eta}\}$, $\mathcal{L}'_2 = \max\{\frac{dL_2^2}{\lambda_\eta^2}, \frac{\sqrt{d} L_2 L_{01}}{\lambda_\eta}\}$. Therefore, the result follows provided that

$$\begin{aligned} m &= m_1 + m_2 + \cdots + m_{L'} \\ &\gtrsim s \cdot \left(d^2 \left(\tfrac{L_{01} B_0}{\delta}\right)^2 \log(sd) \log\left(\tfrac{sd}{\rho}\right) + \sum_{i \in \{0,2\}} \bar{\mathcal{L}}_i \left(\log\left(\tfrac{N_i}{\rho}\right) + \log\left(\tfrac{1}{\rho}\right) \tfrac{\log N_i}{\log(sd)} \right) \right) \end{aligned}$$

where $\bar{\mathcal{L}}_i = (\mathcal{L}_i + \mathcal{L}'_i) \log\left(\frac{L_0}{\varepsilon_\eta} + \frac{L_2}{\lambda_\eta}\right)$. ■

G.2 Step II: Correcting the approximate certificate.

With the lower bound on m that we consider in this section and a union bound, we know using Lemma 20 from the previous section that with probability $1 - \rho/2$ we have $\left\|\Upsilon^{-1}\right\| \leqslant 4$. We can now fix our approximate certificate.

Lemma 23 constructed an approximate certificate $\eta^{\text{app}} = \eta^{L'}$ (where the η^i are the successive "golfing" iterations).

We define $\eta^e = (\Upsilon^{-1} e)^\top \mathbf{f}(\cdot)$ where $e = D\Psi\eta^{\text{app}} - \mathbf{u}_s$, and "fix" the approximate certificate by putting the final certificate as $\eta = \eta^{\text{app}} - \eta_e$. One can verify that

$$\Psi\eta = \Psi(\eta^{\text{app}} - \eta^e) \overset{(21)}{=} \Psi\eta^{app} - D^{-1} e = D^{-1}\mathbf{u}_s = \mathbf{u}_s$$

and thus we have indeed $\eta(x_i) = \text{sign}(a_i)$ and $\nabla\eta(x_i) = 0$. We now check that with our choice of parameters this pre-certificate satisfy the right bounds on the grid.

For all $x \in \mathcal{X}_{\text{grid}}^{\text{far}}$,

$$\begin{aligned} |\eta(x)| &\leqslant |\eta^{\text{app}}(x)| + \left\|\Upsilon^{-1}\right\| \left\|\mathbf{f}(x)\right\| \left\|e\right\| \overset{(71)}{\leqslant} 1 - \frac{\varepsilon_\eta}{2} + \frac{\delta}{1-\delta} + 4\sqrt{s} L_0 L_{01} \left\|e\right\| \\ &\overset{(70)}{\leqslant} 1 - \frac{\varepsilon_\eta}{2} + 4\left(\frac{\delta}{1-\delta} + \frac{\delta^L L_0 L_{01} s\sqrt{d+1}}{C_0^2}\right) \leqslant 1 - \frac{\varepsilon_\eta}{4} \end{aligned}$$

by our choice of δ, and similarly for all $x \in \mathcal{X}_{\text{grid}}^{\text{near}}$,

$$
\begin{aligned}
\nabla^2\eta(x) &\preceq \nabla^2\eta^{\text{app}}(x) + 4\sqrt{s}L_{01}L_2\left\|e\right\|\mathrm{Id} \\
&\overset{(72)}{\preceq} \left(-\frac{\lambda_\eta}{2} + 4\left(\frac{\delta B_2}{(1-\delta)B_0} + \frac{\delta^L L_2 L_{01} s\sqrt{d+1}}{C_0^2}\right)\right)\mathrm{Id} \preceq -\frac{\lambda_\eta}{4}\mathrm{Id}
\end{aligned}
$$

To finish the proof, it suffices to show that the constructed η and its Hessian have a controlled Lipschitz constant. Given that $\eta = \sum_{j=1}^{L'}(\overline{\Upsilon}^{-1}q_{j-1})^\top \mathbf{f}_j(\cdot) - (\Upsilon^{-1}e)^\top\mathbf{f}(\cdot)$, by using the same computations as in the proof of Lemma 19 we obtain for $r = 0, 2$:

$$
\left\|\nabla^r\eta(x) - \nabla^r\eta(x')\right\| \lesssim L_{r+1}L_{01}(\sum_j \left\|q_{j-1}\right\| + \left\|e\right\|)\left\|x - x'\right\|
$$

and

$$
\begin{aligned}
L_{r+1}L_{01}(\sum_j \left\|q_{j-1}\right\| + \left\|e\right\|) &\leqslant L_{r+1}L_{01}\left(\sqrt{s(d+1)}\sum_j\prod_l c_l + \frac{\delta^L\sqrt{s(d+1)}}{C_0^2}\right) \\
&\leqslant L_{r+1}L_{01}\left(\sqrt{s(d+1)}\frac{1}{1-\delta} + 1\right) \\
&\leqslant 4L_{r+1}L_{01}\sqrt{s(d+1)}
\end{aligned}
$$

By our choice of δ_{far} and δ_{near} we can conclude that η satisfies the desired properties.

Appendix H. Fejér kernel: Proof of Prop 4

The kernel we study has the form $K(x, x') = K(t) = \prod_{i=1}^d \kappa(t_i)$ for some univariate function κ. In the following, for a t that is always clear from the context, we shall write $\kappa_i = \kappa(t_i)$ and its derivatives κ_i', κ_i'' and so on, and $K_i = \prod_{j=1,\ j\neq i}^d \kappa_j$, K_{ij} and K_{ijl} in the same way. With this, we have:

$$
\begin{aligned}
\partial_{1,i}K(x,x') &= \kappa_i' K_i \\
\partial_{1,i}\partial_{2,i}K(x,x') &= -\kappa_i'' K_i
\end{aligned}
$$

and using Gershgorin theorem:

$$
\begin{aligned}
\left\|\nabla_2^2 K(x,x')\right\| &\leqslant \max_{1\leqslant i\leqslant d}\{\left|\kappa_i'' K_i\right| + \left|\kappa_i'\right|\sum_{j\neq i}\left|\kappa_j'\right|\left|K_{ij}\right|\} \\
\lambda_{\min}\left(\nabla_2^2 K(x,x')\right) &\in \bigcup_{1\leqslant i\leqslant d}\left(\kappa_i'' K_i \pm \left|\kappa_i'\right|\sum_{j\neq i}\left|\kappa_j'\right|\left|K_{ij}\right|\right) \\
\left\|\partial_{1,i}\nabla_2^2 K(x,x')\right\| &\leqslant \max\left\{\left|\kappa_i''' K_i\right| + \left|\kappa_i''\right|\sum_{j\neq i}\left|\kappa_j'\right|\left|K_{ij}\right|,\right. \\
&\qquad \left.\max_{j\neq i}\{\left|\kappa_j''\kappa_i' K_{ij}\right| + \left|\kappa_j'\kappa_i'' K_{ij}\right| + \left|\kappa_i'\right|\left|\kappa_j'\right|\sum_{l\neq i,j}\left|\kappa_l'\right|\left|K_{ijl}\right|\}\right\}
\end{aligned}
$$

Now, in the particular case of the univariate Fejér kernel we can write the following bounds Candès and Fernandez-Granda (2014) whose value we shall detail later.

- for $t = a/f_c \in [0, a_{\lim}/f_c]$ and $\ell = 0, 1, 2, 3$:

$$\begin{aligned}
\left|\kappa^{(\ell)}(t)\right| / f_c^{\ell} &\leqslant \kappa_\ell^{<}(a) && \text{which is decreasing for even } \ell \text{ and increasing for odd } \ell, \\
\kappa(t) &\geqslant \kappa_0^{<,l}(a) && \text{which is positive and decreasing,} \\
\kappa''(t)/f_c^2 &\leqslant \kappa_2^{<,l}(a) && \text{which is negative and increasing,}
\end{aligned}$$

- for $t = a/f_c \in [a_{\lim}/f_c, 1/2]$:

$$\left|\kappa^{(\ell)}(t)\right| / f_c^{\ell} \leqslant \kappa_\ell^{>}(a)$$

  which is decreasing for all $\ell = 0, 1, 2, 3$.

In the particular case of the Fejér kernel, if $f_c \geqslant 128$, we have:

$$\begin{aligned}
\kappa_0^{<}(a) &= 1 - \frac{\pi^2}{6} a^2 + \frac{\pi^4 \left(1 + \frac{1}{64}\right)^4}{72} \cdot a^4 \\
\kappa_1^{<}(a) &= \frac{\pi^2 \left(1 + \frac{1}{32}\right)}{3} \cdot a \\
\kappa_2^{<}(a) &= \frac{\pi^2 \left(1 + \frac{1}{32}\right)}{3} \\
\kappa_3^{<}(a) &= \frac{\pi^4 \left(1 + \frac{1}{64}\right)^4}{3} \cdot a \\
\kappa_0^{<,l}(a) &= 1 - \frac{\pi^2}{6} a^2 \\
\kappa_2^{<,l}(a) &= -\frac{\pi^2}{3} + \frac{\pi^4 \left(1 + \frac{1}{64}\right)^4}{6} \cdot a^2
\end{aligned}$$

and $\kappa_\ell^{>}(a) = \frac{\pi^\ell H_\ell(a)}{\left(2 + \frac{1}{128}\right)^{4-\ell} a^4}$ for $a \in [a_{\lim}; \sqrt{2} f_c/\pi]$ and decreasing after, with

$$\begin{aligned}
H_0(a) &= \alpha^4(a) \\
H_1(a) &= \alpha^4(a)(2 + 2\beta(a)) \\
H_2(a) &= \alpha^4(a)(4 + 7\beta(a) + 6\beta(a)^2) \\
H_3(a) &= \alpha^4(a)(8 + 24\beta(a) + 30\beta(a)^2 + 15\beta(a)^3)
\end{aligned}$$

where

$$\alpha(a) = \frac{2}{\pi \left(1 - \frac{\pi^2 a^2}{6 f_c^2}\right)}, \quad \beta(a) = \frac{\alpha(a)}{a}$$

We note that, for $f_c \geqslant 128\sqrt{d}\sqrt[4]{s_{\max}}$, we have

$$\alpha(a\sqrt{d}\sqrt[4]{s_{\max}}) \leqslant \frac{2}{\pi \left(1 - \frac{\pi^2 a^2}{6 \cdot 128^2}\right)}$$

and that is therefore the expression that we shall use for the bounds involving $\Delta \propto f_c^{-1}\sqrt{d}\sqrt[4]{s_{\max}}$.

Now, setting $\varepsilon_{\text{near}} = a \cdot d^{-\frac{1}{2}} \cdot f_c^{-1}$ with $a \leqslant a_{\lim}$, for $\|t\|_\infty \leqslant \varepsilon_{\text{near}}$ we get, using the fact that $(1-c/d)^{d-1} \geqslant e^{-\frac{c+1}{1-c}}$ for all d,

$$\begin{aligned}
\left|\kappa_i'' K_i\right| / f_c^2 &\leqslant \gamma_0(a) \stackrel{\text{def.}}{=} \kappa_2^<\left(\tfrac{a}{\sqrt{d}}\right) = \mathcal{O}(1) \\
-\kappa_i'' K_i / f_c^2 &\geqslant -\kappa_2^{<,l}\left(\tfrac{a}{\sqrt{d}}\right)\left(\kappa_0^{<,l}\left(\tfrac{a}{\sqrt{d}}\right)\right)^{d-1} \\
&\geqslant \gamma_1(a) \stackrel{\text{def.}}{=} \left(\frac{\pi^2}{3} - \frac{\pi^4\left(1+\frac{1}{64}\right)^4}{6}\cdot a^2\right)\exp\left(-\frac{\frac{\pi^2}{6}a^2+1}{1-\frac{\pi^2}{6}a^2}\right) \\
\left|\kappa_i'\kappa_j' K_{ij}\right| / f_c^2 &\leqslant \gamma_2(a) \stackrel{\text{def.}}{=} \kappa_1^<\left(\tfrac{a}{\sqrt{d}}\right)^2 = \mathcal{O}(1/d) \\
\left|\kappa_i''' K_i\right| / f_c^3 &\leqslant \gamma_3(a) \stackrel{\text{def.}}{=} \kappa_3^<\left(\tfrac{a}{\sqrt{d}}\right) = \mathcal{O}\left(1/\sqrt{d}\right) \\
\left|\kappa_i''\kappa_j' K_{ij}\right| f_c^3 &\leqslant \gamma_4(a) \stackrel{\text{def.}}{=} \kappa_2^<\left(\tfrac{a}{\sqrt{d}}\right)\kappa_1^<\left(\tfrac{a}{\sqrt{d}}\right) = \mathcal{O}\left(1/\sqrt{d}\right) \\
\left|\kappa_i'\kappa_j'\kappa_l' K_{ijl}\right| / f_c^3 &\leqslant \gamma_5(a) \stackrel{\text{def.}}{=} \kappa_1^<\left(\tfrac{a}{\sqrt{d}}\right)^3 = \mathcal{O}\left(1/d^{3/2}\right)
\end{aligned}$$

Next, for all $\|t\|_\infty \geqslant a/(\sqrt{d}f_c)$ with $a \leqslant a_{\lim}$, we will need

$$\begin{aligned}
|K(t)| &\leqslant \kappa_0^<\left(\tfrac{a}{\sqrt{d}}\right) = \max\{1-\gamma_7(a)/d, \sup_{t \geqslant a_{\lim}/f_c} \kappa_0^>(t)\} \\
&= 1-\gamma_7(a)/d \text{ where } \gamma_7(a) = \frac{\pi^2}{6}a^2 - \frac{\pi^4\left(1+\frac{1}{64}\right)^4}{72}a^4 \\
\left|\kappa_i' K_i\right| / f_c &\leqslant \kappa_1^{\max} = \mathcal{O}(1)
\end{aligned}$$

And finally for $\|t\|_\infty \geqslant \bar{A}\sqrt{d}s_{\max}^{1/4}/f_c$ for $\bar{A} \geqslant a_{\lim}$, denoting by $A = \bar{A}\sqrt{d}s_{\max}^{1/4}$, we have

$$\begin{aligned}
|K| &\leqslant \gamma_8(\bar{A}) \stackrel{\text{def.}}{=} \kappa_0^>(A) \\
\left|\kappa_i' K_i\right| / f_c &\leqslant \gamma_9(\bar{A}) \stackrel{\text{def.}}{=} \max\{\kappa_1^>(A),\ \kappa_1^{\max}\kappa_0^>(A)\} \\
\left|\kappa_i'' K_i\right| / f_c^2 &\leqslant \gamma_{10}(\bar{A}) \stackrel{\text{def.}}{=} \max\{\kappa_2^>(A),\ \kappa_2^{\max}\kappa_0^>(A)\} \\
\left|\kappa_i'\kappa_j' K_{ij}\right| / f_c^2 &\leqslant \gamma_{11}(\bar{A}) \stackrel{\text{def.}}{=} \kappa_1^{\max}\max\{\kappa_1^>(A),\ \kappa_0^>(A)\} \\
\left|\kappa_i''' K_i\right| / f_c^3 &\leqslant \gamma_{12}(\bar{A}) \stackrel{\text{def.}}{=} \max\{\kappa_3^>(A),\ \kappa_3^{\max}\kappa_0^>(A)\} \\
\left|\kappa_i''\kappa_j' K_{ij}\right| f_c^3 &\leqslant \gamma_{13}(\bar{A}) \stackrel{\text{def.}}{=} \max\{\kappa_2^>(A)\kappa_1^{\max},\ \kappa_1^>(A)\kappa_2^{\max}, \kappa_2^{\max}\kappa_1^{\max}\kappa_0^>(A)\} \\
\left|\kappa_i'\kappa_j'\kappa_l' K_{ijl}\right| / f_c^3 &\leqslant \gamma_{14}(\bar{A}) \stackrel{\text{def.}}{=} \max\{\kappa_1^>(A)(\kappa_1^{\max})^2,\ (\kappa_1^{\max})^3\kappa_0^>(A)\}
\end{aligned}$$

and all these bounds are as $\mathcal{O}\left(\frac{1}{d^2 s_{\max}}\right)$.

1. **At distance 0.** For all $x \in \mathbb{R}^d$,

$$a_1 = 0,$$
$$a_2 = 0,$$
$$v = \frac{\pi^2}{3}$$

2. **At small distance.** For $x, x' \in \mathbb{R}^d$ such that $\|x - x'\| \leqslant \varepsilon_{\text{near}} = a/(\sqrt{d} f_c)$,

$$b_2/f_c^2 = \gamma_0(a) + d\gamma_2(a) = \mathcal{O}\left(1\right)$$
$$\lambda_1/f_c^2 = \gamma_1(a)$$
$$b_3 = \sqrt{d}\left(\gamma_3(a) + (d+1)\gamma_4(a) + d\gamma_5(a)\right) = \mathcal{O}\left(d\right)$$

3. **At distance larger than $\varepsilon_{\text{near}}$.** For $x, x' \in \mathbb{R}^d$ such that $\|x - x'\| \geqslant \varepsilon_{\text{near}}$,

$$c_0 = 1 - \gamma_7(a)/d$$
$$c_1/f_c = \sqrt{d}\kappa_1^{\max} = \mathcal{O}\left(\sqrt{d}\right)$$

4. **At distance larger than $\Delta/2$.** For $x, x' \in \mathbb{R}^d$ such that $\|x - x'\| \geqslant \Delta/2$,

$$e_0 = s_{\max}\gamma_8(\bar{A}/2) = \mathcal{O}\left(d^{-2}\right)$$
$$e_1/f_c = s_{\max}\sqrt{d}\gamma_9(\bar{A}/2) = \mathcal{O}\left(d^{-3/2}\right)$$
$$e_2/f_c^2 = s_{\max}\left(\gamma_{10}(\bar{A}/2) + d\gamma_{11}(\bar{A})\right) = \mathcal{O}\left(d^{-1}\right)$$
$$e_3/f_c^3 = s_{\max}\left(\sqrt{d}\left(\gamma_{12}(\bar{A}/2) + (d+1)\gamma_{13}(\bar{A}/2) + d\gamma_{14}(\bar{A}/2)\right)\right) = \mathcal{O}\left(d^{-1/2}\right)$$

5. **At distance larger than Δ.** Finally, for $x, x' \in \mathbb{R}^d$ such that $\|x - x'\| \geqslant \Delta$,

$$h_0 = s_{\max}\gamma_8(\bar{A}) = \mathcal{O}\left(d^{-2}\right)$$
$$h_1/f_c = s_{\max}\gamma_9(\bar{A}) = \mathcal{O}\left(d^{-2}\right)$$
$$h_2/f_c^2 = s_{\max}\left(\gamma_{10}(\bar{A}) + d\gamma_{11}(\bar{A})\right) = \mathcal{O}\left(d^{-1}\right)$$

Then proceeding to write conditions (31) to (34), we obtain $u = \mathcal{O}\left(1/d^2\right)$, $\delta = \mathcal{O}\left(1/d\right)$, $\delta' = \mathcal{O}\left(1/d^2\right)$, $\varepsilon_\eta = \mathcal{O}\left(1/d\right)$ and $\lambda_\eta/f_c^2 = \mathcal{O}\left(1\right)$. We then adjust the constants to obtain the desired result.

Additional bounds. Finally, one can trivially check that (34) is satisfied, and that we have $B_0 = \mathcal{O}\left(\sqrt{d}\right)$ and $B_2 = \mathcal{O}\left(d f_c^2\right)$.

Appendix I. Gaussian kernel: proof of proposition 5

In this section we prove Proposition 17. Despite their apparent complexity, the computations are actually quite natural in the case of the Gaussian kernel.

We consider the Gaussian kernel $\overline{K}(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$ in $\mathbb{R}^d$. For simplicity define $t = x - x'$ and $K(t) = \exp\left(-\frac{\|t\|^2}{2\sigma^2}\right)$. Denote by $\{e_i\}$ the canonical basis of $\mathbb{R}^d$. We have the following:

$$
\begin{aligned}
\partial_{1,i} K(x, x') &= -\frac{1}{\sigma^2} t_i K(t) \\
\partial_{1,i}\partial_{2,i} K(x, x') &= \left(\frac{1}{\sigma^2} - \frac{1}{\sigma^4} t_i^2\right) K(t) \\
\nabla_2^2 C(x, x') &= \left(-\frac{1}{\sigma^2}\mathrm{Id} + \frac{1}{\sigma^4} tt^\top\right) K(t) \\
\partial_{1,i}\nabla_2^2 C(x, x') &= \left(\frac{1}{\sigma^4} t_i \mathrm{Id} + \frac{1}{\sigma^4}\left(te_i^\top + e_i t^\top\right) - \frac{1}{\sigma^6} t_i tt^\top\right) K(t)
\end{aligned}
$$

We are going to repeatedly use the fact that for any $s \geqslant 1$ the function $f(r) = r^s e^{-\frac{r^2}{2\sigma^2}}$ defined on $\mathbb{R}_+$ is increasing on $[0, \sqrt{s}\sigma]$ and decreasing after, and its maximum value is $f(\sqrt{s}\sigma) = (s\sigma^2)^{s/2} e^{-s/2}$. Similarly, for $q > 0$, the function $g(r) = (\log r)/r^q$ defined on $[1, \infty)$ is bounded by $A/(qe)$.

Therefore we obtain the following bounds.

1. **At distance 0.**

$$
a_1 \overset{\text{def.}}{=} 0,\; a_2 \overset{\text{def.}}{=} 0,\; v \overset{\text{def.}}{=} \frac{1}{\sigma^2}
$$

2. **At small distance.** Taking $\varepsilon_{\text{near}} = \sigma/\sqrt{2}$, for all $t = x - x'$ such that $\|x - x'\| \leqslant \varepsilon_{\text{near}}$:

$$
\begin{aligned}
\left\|\nabla_2^2 K(x, x')\right\|_{sp.} &\leqslant \left(\frac{1}{\sigma^2} + \frac{1}{\sigma^4}\|t\|^2\right) K(t) \leqslant \frac{1}{\sigma^2} + \frac{e^{-1/4}}{2\sigma^2} \leqslant b_2 \overset{\text{def.}}{=} \frac{1.3895}{\sigma^2} \\
\lambda_{\min}(\nabla_2^2 K(x, x')) &\leqslant -\left(\frac{1}{\sigma^2} + \frac{\|t\|^2}{\sigma^4}\right) K(t) \leqslant -\frac{e^{-1/4}}{\sigma^2} + \frac{e^{-1/4}}{2\sigma^2} \leqslant -\lambda_1 \overset{\text{def.}}{=} -\frac{0.3893}{\sigma^2}
\end{aligned}
$$

$$
\begin{aligned}
\forall u \in \mathbb{R}^d,\quad \left\|\sum_{i=1}^d u_i \partial_{1,i}\nabla_2^2 K(x, x')\right\| &= \left\|\frac{1}{\sigma^4} t^\top u \mathrm{Id} + \frac{1}{\sigma^4}(tu^\top + ut^\top) - \frac{1}{\sigma^6}(t^\top u) tt^\top\right\| K(t) \\
&\leqslant \|u\| \left(\frac{3\|t\|}{\sigma^4} + \frac{\|t\|^3}{\sigma^6}\right) K(t) \leqslant \|u\| \frac{3/\sqrt{2} + (1/2)^{3/2}}{\sigma^3} e^{-1/4} \\
&\leqslant \|u\|\, b_3 \text{ with } b_3 \overset{\text{def.}}{=} 2.4750/\sigma^3
\end{aligned}
$$

3. **At distance larger than $\varepsilon_{\text{near}}$.** For $x, x' \in \mathbb{R}^d$ such that $\|x - x'\| \geqslant \varepsilon_{\text{near}}$,

$$
\begin{aligned}
\left|K(x, x')\right| &\leqslant e^{-1/4} \leqslant c_0 \overset{\text{def.}}{=} 0.7789 \\
\left\|\nabla_1 K(x, x')\right\|_2 &\leqslant \frac{1}{\sigma^2}\|t\| K(t) \leqslant \frac{e^{-1/2}}{\sigma} \leqslant c_1 \overset{\text{def.}}{=} 0.6066/\sigma
\end{aligned}
$$

4. **At distance larger than $\Delta/2$.** We define $\Delta = \sqrt{2}\sigma\sqrt{A\log s_{\max} + B\log d + C}$ for $A, B, C > 0$ that we will adjust later. Denote $E_{kd} = A\log s_{\max} + B\log d + C$ for shortness. For $x, x' \in \mathbb{R}^d$ such that $\|x - x'\| \geqslant \Delta/2$, we have

$$
\begin{aligned}
\left|K(x,x')\right| &\leqslant \frac{e_0}{s_{\max}} \quad \text{where} \quad e_0 \overset{\text{def.}}{=} e^{-\frac{C}{4}} \cdot s_{\max}^{-\frac{A}{4}+1} \cdot d^{-\frac{B}{4}} \\
\left\|\nabla_1 K(x,x')\right\|_2 &= \frac{1}{\sigma^2}\|t\| K(t) \leqslant \frac{\Delta}{2\sigma^2} K(\Delta/2) \\
&= \frac{e_1}{s_{\max}} \quad \text{where} \quad e_1 \overset{\text{def.}}{=} \frac{\sqrt{E_{kd}}}{\sqrt{2}\sigma e^{\frac{C}{4}} \cdot s_{\max}^{\frac{A}{4}-1} \cdot d^{\frac{B}{4}}} \\
\left\|\nabla_2^2 K(x,x')\right\| &\leqslant \left(\frac{1}{\sigma^2} + \frac{1}{\sigma^4}\|t\|^2\right) K(t) \leqslant \left(\frac{1}{\sigma^2} + \frac{\Delta^2}{4\sigma^4}\right) K(\Delta/2) \\
&= \frac{e_2}{s_{\max}} \quad \text{where} \quad e_2 \overset{\text{def.}}{=} \frac{4 + E_{kd}}{4\sigma^2 e^{\frac{C}{4}} \cdot s_{\max}^{\frac{A}{4}-1} \cdot d^{\frac{B}{4}}} \\
\forall u \in \mathbb{R}^d, \quad \left\|\sum_{i=1}^d u_i \partial_{1,i} \nabla_2^2 K(x,x')\right\|_{sp.} &\leqslant \|u\| \left(\frac{3\|t\|}{\sigma^4} + \frac{\|t\|^3}{\sigma^6}\right) K(t) \leqslant \|u\| \left(\frac{3\Delta}{2\sigma^4} + \frac{\Delta^3}{8\sigma^6}\right) K(\Delta/2) \\
&\leqslant \|u\| \frac{e_3}{s_{\max}} \quad \text{where} \quad e_3 \overset{\text{def.}}{=} \frac{\sqrt{E_{kd}}(6 + E_{kd})}{2\sqrt{2}\sigma^3 e^{\frac{C}{4}} \cdot s_{\max}^{\frac{A}{4}} \cdot d^{\frac{B}{4}}}
\end{aligned}
$$

5. **At distance larger than Δ.** Finally, for $x, x' \in \mathbb{R}^d$ such that $\|x - x'\| \geqslant \Delta$,

$$
\begin{aligned}
\left|K(x,x')\right| &\leqslant \frac{h_0}{s_{\max}} \quad \text{where} \quad h_0 \overset{\text{def.}}{=} e^{-C} \cdot s_{\max}^{-A+1} \cdot d^{-B} \\
\left|\partial_{1,i} K(x,x')\right| &\leqslant \frac{1}{\sigma^2}\|t\| K(t) \leqslant \frac{\Delta}{\sigma^2} K(\Delta) = \frac{h_1}{s_{\max}} \quad \text{where} \quad h_1 \overset{\text{def.}}{=} \frac{\sqrt{2E_{kd}}}{\sigma e^C \cdot s_{\max}^{A-1} \cdot d^B} \\
\left\|\partial_{1,i} \nabla_2 K(x,x')\right\|_1 &= \left|\left(\frac{1}{\sigma^2} - \frac{1}{\sigma^4} t_i^2\right)\right| K(t) + \sum_{j \neq i} \frac{1}{\sigma^4} |t_i t_j| K(t) \\
&\leqslant \left(\frac{1}{\sigma^2} + \frac{\sqrt{d}\|t\|^2}{\sigma^4}\right) K(t) \leqslant \left(\frac{1}{\sigma^2} + \frac{2\sqrt{d}\Delta^2}{\sigma^4}\right) K(\Delta) \\
&= \frac{h_2}{s_{\max}} \quad \text{where} \quad h_2 \overset{\text{def.}}{=} \frac{1 + 4\sqrt{d}E_{kd}}{\sigma^2 e^C \cdot s_{\max}^{A-1} \cdot d^B}
\end{aligned}
$$

Final bounds. We have $u = h_1/\sqrt{v} = \frac{\sqrt{2E_{kd}}}{e^C \cdot s_{\max}^{A-1} \cdot d^B}$. Using the fact that for all $s_{\max}, d$ we have $E_{kd} s_{\max}^{-a} d^{-b} \leqslant A/(ae) + B/(be) + C$, the equations (31) - (33) becomes

$$\begin{aligned}
v^{-1}(da_2 + h_2) &= \frac{1 + 4\sqrt{d}E_{kd}}{e^C \cdot s_{\max}^{A-1} \cdot d^B} \leqslant \delta \overset{\text{def.}}{=} e^{-C}\left(1 + \frac{4A}{e(A-1)} + \frac{4B}{e(B-1/2)} + 4C\right) \\
h_0 + du^2(1-\delta)^{-1} &= e^{-C} \cdot s_{\max}^{-A+1} \cdot d^{-B} + \frac{2E_{kd}}{(1-\delta)e^{2C} \cdot s_{\max}^{2A-2} \cdot d^{2B-1}} \\
&\leqslant \delta' \overset{\text{def.}}{=} e^{-C} + \frac{2e^{-2C}}{1-\delta}\left(\frac{A}{2e(A-1)} + \frac{B}{e(2B-1)} + C\right)
\end{aligned}$$

and

$$\begin{aligned}
&1 - \left[(1-\delta')^{-1}(c_0 + e_0) + \tfrac{u\sqrt{d}}{\sqrt{v}(1-\delta)(1-\delta')} \cdot (c_1 + e_1)\right] \\
&\quad = 1 - \Bigg[\frac{0.7789 + e^{-\frac{C}{4}} \cdot s_{\max}^{-\frac{A}{4}+1} \cdot d^{-\frac{B}{4}}}{1-\delta'} \\
&\qquad + \frac{\sqrt{2E_{kd}}e^{-C}s_{\max}^{-A+1}d^{-B+1/2}}{(1-\delta)(1-\delta')}\left(0.6066 + \frac{\sqrt{E_{kd}}}{\sqrt{2}e^{\frac{C}{4}} \cdot s_{\max}^{\frac{A}{4}-1} \cdot d^{\frac{B}{4}}}\right)\Bigg] \\
&\quad \geqslant \varepsilon_\eta \overset{\text{def.}}{=} 1 - \Bigg[\frac{0.7789 + e^{-\frac{C}{4}}}{1-\delta'} \\
&\qquad + \frac{\sqrt{2}e^{-C}\left(\frac{A}{2e(A-1)} + \frac{B}{e(2B-1)} + C\right)^{\frac{1}{2}}}{(1-\delta)(1-\delta')}\left(0.6066 + \frac{e^{-\frac{C}{4}}}{\sqrt{2}} \cdot \left(\frac{A}{e(\frac{A}{2}-2)} + \frac{2}{e} + C\right)^{\frac{1}{2}}\right)\Bigg]
\end{aligned}$$

and finally

$$\begin{aligned}
&\left(1 - \tfrac{\delta'}{1-\delta'}\right) \cdot \lambda_1 - (1-\delta')^{-1}e_2 - \tfrac{u\sqrt{d}}{\sqrt{v}(1-\delta)(1-\delta')} \cdot (b_3 + e_3) \\
&\quad = \frac{1}{\sigma^2}\Bigg[\left(1 - \tfrac{\delta'}{1-\delta'}\right)0.3893 - \frac{(4 + E_{kd})e^{-\frac{C}{4}} \cdot s_{\max}^{-\frac{A}{4}+1} \cdot d^{-\frac{B}{4}}}{4(1-\delta')} \\
&\qquad - \frac{\sqrt{2E_{kd}}e^{-C}s_{\max}^{-A+1}d^{-B+1/2}}{(1-\delta)(1-\delta')}\left(2.4750 + \frac{\sqrt{E_{kd}}(6 + E_{kd})}{2\sqrt{2}e^{\frac{C}{4}} \cdot s_{\max}^{\frac{A}{4}-1} \cdot d^{\frac{B}{4}}}\right)\Bigg] \\
&\quad \geqslant \lambda_\eta \overset{\text{def.}}{=} \frac{1}{\sigma^2}\Bigg[\left(1 - \tfrac{\delta'}{1-\delta'}\right)0.3893 - \frac{\left(1 + \frac{A}{e(A-4)} + \frac{1}{e} + \frac{C}{4}\right)e^{-\frac{C}{4}}}{1-\delta'} \\
&\qquad - \frac{\sqrt{2}e^{-C}\left(\frac{A}{2e(A-1)} + \frac{B}{e(2B-1)} + C\right)^{\frac{1}{2}}}{(1-\delta)(1-\delta')}\Bigg(2.4750 \\
&\qquad + \frac{e^{-\frac{C}{4}}}{2\sqrt{2}}\left(6\left(\frac{A}{e(A/2-2)} + \frac{2}{e} + C\right)^{\frac{1}{2}} + \left(\frac{3A}{e(A/2-2)} + \frac{6}{e} + C\right)^{\frac{3}{2}}\right)\Bigg)\Bigg]
\end{aligned}$$

By taking $A = 5, B = 2$ and $C = 12$ we obtain $\delta \leqslant 3.2443 \cdot 10^{-4}$, $\delta' \leqslant 6.1453 \cdot 10^{-6}$, $\varepsilon_\eta \geqslant 0.1712$ and $\lambda_\eta \geqslant 0.0800/\sigma^2$.

Additional bounds. Finally, one can trivially check that (34) is satisfied, and that we have $B_0 = \mathcal{O}(1)$ and $B_2 = \mathcal{O}\left(1/\sigma^2\right)$.

Appendix J. Proof of Proposition 9, summary of the examples

By the expression of the characteristic function of a Gaussian, we have

$$y_k = \frac{M_K}{n} \sum_{i=1}^{n} e^{i\omega_k^\top t_i} \approx M_K \sum_{i=1}^{s} a_i e^{i\omega_k^\top x_i} e^{-\frac{\|\omega_k\|^2}{2}} = \langle \varphi_{\omega_k}, \mu_0 \rangle$$

where $\mu_0 = \sum_i a_i \delta_{x_i}$ and $\varphi_\omega(x) = M_K e^{i\omega_k^\top x_i} e^{-\frac{\|\omega_k\|^2}{2}}$. In that case ((Gribonval et al., 2017), Lemma G.1), we can verify that (7) recovers the Gaussian kernel with $\sigma^2 = 2 + \sigma_K^{-2}$, when the distribution of the ω is a Gaussian with variance σ_K^2. The noise can be bounded using a generalized Hoeffding's inequality (Rahimi and Recht (2008), Lemma 4).

Hence we can apply the results for the Gaussian kernel (Prop. 5), with $\varepsilon_\eta = \mathcal{O}(1)$, $\lambda_\eta = \mathcal{O}\left(1/\sigma^2\right)$, $B_0 = \mathcal{O}(1)$, $B_2 = \mathcal{O}\left(1/\sigma^2\right)$, $\varepsilon_{\text{near}} = \mathcal{O}(\sigma)$, and for the features $L_i = \mathcal{O}(M_K)$.

We finish by writing in Table 2 a summary of all the objects involved in our examples.

	Fejér (discrete Fourier)	Gaussian (continuous fourier)	GMM
Domain	$\mathcal{X} = \mathbb{T}^d$	Compact $\mathcal{X} \subset \mathbb{R}^d$	Param.: $\mathcal{X} \subset \mathbb{R}^d$ Data: $T = \mathbb{R}^d$
Kernel $\overline{K}$	$\prod_{i=1}^{d} \left(\frac{\sin\left(\left(\frac{f_c}{2}+1\right)\pi(x-x')\right)}{\left(\frac{f_c}{2}+1\right)\sin(\pi(x-x'))} \right)^4$	$e^{-\frac{\Vert x-x'\Vert^2}{2\sigma^2}}$	$e^{-\frac{\Vert x-x'\Vert^2}{2\left(1+\sigma_K^{-2}\right)}}$
Feat. dom.	$\Omega = [\![-f_c\ ;\ f_c]\!]^d$	$\Omega = \mathbb{R}^d$	$\Omega = \mathbb{R}^d$
Features	$\varphi_\omega(x) = e^{\mathrm{i}2\pi\omega^\top x}$	$\varphi_\omega(x) = e^{\mathrm{i}\omega^\top x}/f(\omega)$ where $f(\omega) = \frac{1}{2}\left(\sum_{i=0}^{3} \frac{\Vert\omega\Vert^{2i}}{\gamma_{2i}}\right)^{\frac{1}{2}}$ $\gamma_i = \mathcal{O}\left(\frac{d^{i/2}}{\sigma^i}\right)$	$\varphi_\omega(x) = M_K e^{i\omega^\top x - \frac{\Vert\omega\Vert^2}{2}}$ where $M_K = (1+2\sigma_K^2)^{d/4}$
Feat. Distrib.	$\Lambda(\omega) = \prod_{i=1}^d g(\omega_i)$	$\Lambda(\omega) = f(\omega)^2 \mathcal{N}(0, \sigma^{-2}\mathrm{Id})$	$\Lambda(\omega) = \mathcal{N}(0, \sigma_K^2 \mathrm{Id})$
Meas.	$y_k = \langle \varphi_{\omega_k}, \mu_0 \rangle + \varepsilon_k$	$y_k = \langle \varphi_{\omega_k}, \mu_0 \rangle + \varepsilon_k$	$y_k = \frac{M_K}{n} \sum_{i=1}^n e^{i\omega_k^\top t_i}$
Norm sep.	$\Vert\cdot\Vert_{\text{sep.}} = \Vert\cdot\Vert_\infty$	$\Vert\cdot\Vert_{\text{sep.}} = \Vert\cdot\Vert_2$	$\Vert\cdot\Vert_{\text{sep.}} = \Vert\cdot\Vert_2$
Sep.	$\Delta = \mathcal{O}\left(f_c^{-1}\sqrt{d}\sqrt[4]{s_{\max}}\right)$	$\Delta = \mathcal{O}\left(\sigma\sqrt{\log(d s_{\max})}\right)$	$\Delta = \mathcal{O}\left(\sqrt{(1+\sigma_K^{-2})\log(d s_{\max})}\right)$
Neighb. Size	$\varepsilon_{\text{near}} = \mathcal{O}\left(1/(\sqrt{d} f_c)\right)$	$\varepsilon_{\text{near}} = \mathcal{O}\left(\sigma\right)$	$\varepsilon_{\text{near}} = \mathcal{O}\left(\sqrt{1+\sigma_K^{-2}}\right)$
Curv.	$\lambda_\eta = \mathcal{O}\left(f_c^2\right)$	$\lambda_\eta = \mathcal{O}\left(1/\sigma^2\right)$	$\lambda_\eta = \mathcal{O}\left(\frac{1}{1+\sigma_K^{-2}}\right)$
Dist. to 1	$\varepsilon_\eta = \mathcal{O}\left(1/d\right)$	$\varepsilon_\eta = \mathcal{O}\left(1\right)$	$\varepsilon_\eta = \mathcal{O}\left(1\right)$
Lip. const.	$L_i = \mathcal{O}\left(d^{i/2} f_c^i\right)$	$L_i = \mathcal{O}\left(d^{i/2}/\sigma^i\right)$	$L_i = \mathcal{O}\left(M_K\right)$
Bounds	$B_0 = \mathcal{O}\left(\sqrt{d}\right)$ $B_2 = \mathcal{O}\left(d f_c^2\right)$	$B_i = \mathcal{O}\left(1/\sigma^i\right)$	$B_i = \mathcal{O}\left((1+\sigma_K^2)^{-i/2}\right)$

Table 2: Summary of all quantities involved in the three examples of Section 3.4. The Fejér kernel is defined with $g(j) = \frac{1}{f_c}\sum_{k=\max(j-f_c,-f_c)}^{\min(j+f_c,f_c)} \left(1 - \left|\frac{k}{f_c}\right|\right)\left(1 - \left|\frac{j-k}{f_c}\right|\right)$.