Support Localization and the Fisher Metric for off-the-grid Sparse Regularization

Clarice Poon[1] **Nicolas Keriven**[2] **Gabriel Peyré**[2]
[1]Centre for Mathematical Sciences
University of Cambridge
Wilberforce Rd, Cambridge, United Kingdom
[2]Département de Mathématiques et Applications
École Normale Supérieure
45 rue d'Ulm, Paris, France

Abstract

Sparse regularization is a central technique for both machine learning (to achieve supervised features selection or unsupervised mixture learning) and imaging sciences (to achieve super-resolution). Existing performance guaranties assume a separation of the spikes based on an ad-hoc (usually Euclidean) minimum distance condition, which ignores the geometry of the problem. In this article, we study the BLASSO (i.e. the off-the-grid version of ℓ^1 LASSO regularization) and show that the Fisher-Rao distance is the natural way to ensure and quantify support recovery, since it preserves the invariance of the problem under reparameterization. We prove that under mild regularity and curvature conditions, stable support identification is achieved even in the presence of randomized sub-sampled observations (which is the case in compressed sensing or learning scenario). On deconvolution problems, which are translation invariant, this generalizes to the multi-dimensional setting existing results of the literature. For more complex translation-varying problems, such as Laplace transform inversion, this gives the first geometry-aware guarantees for sparse recovery.

1 Introduction

1.1 Sparse Regularization

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 $y \in \mathbb{C}^m$, Let $\Phi : \mathcal{M}(\mathcal{X}) \mapsto \mathbb{C}^m$ be defined by

$$\Phi\mu \stackrel{\text{def.}}{=} \frac{1}{\sqrt{m}} \left(\int_{\mathcal{X}} \varphi_{\omega_k}(x) \mathrm{d}\mu(x) \right)_{k=1}^{m}, \qquad (1.1)$$

where $(\omega_1, \ldots, \omega_m)$ 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 $\varphi_\omega \in \mathscr{C}(\mathcal{X})$. We further assume that $\varphi_\omega(x)$ is normalized, that is

$$\mathbb{E}_\omega[|\varphi_\omega(x)|^2] = 1, \qquad \forall x \in \mathcal{X}. \qquad (1.2)$$

The observations are $y = \Phi\mu_0 + w$, where $w \in \mathbb{C}^m$ accounts for noise or modelling errors. 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), corresponds exactly to measurements of the form (1.1). This is a "continuous" analogous of the celebrated compressed sensing line of works (Candès et al., 2006; Donoho, 2006).
- *Regression using a continuous 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 template densities,

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 rewritten 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. 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 $\theta_\omega(t)$, that is, for $k = 1, \ldots, m$,

$$y_k \overset{\text{def.}}{=} \frac{1}{n} \sum_{j=1}^{n} \theta_{\omega_k}(t_j) \approx \int_{\mathcal{T}} \theta_{\omega_k}(t) \xi(t) \mathrm{d}t.$$

These sketches exactly have the form (1.1) 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.

BLASSO. In all these applications, and many 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 $(x_i, a_i) \in \mathcal{X} \times \mathbb{C}$. An increasingly popular method to estimate such a sparse measure corresponds to solving a infinite-dimensional analogous of the Lasso regression problem

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} \frac{1}{2} \|\Phi\mu - y\|_2^2 + \lambda |\mu|(\mathcal{X}). \qquad (\mathcal{P}_\lambda(y))$$

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

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

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 $|\mu|(\mathcal{X}) = \|a\|_1 \overset{\text{def.}}{=} \sum_i |a_i|$ 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 $|\mu|(\mathcal{X}) = \|a\|_1$.

Previous Works. The BLASSO problem $(\mathcal{P}_\lambda(y))$ was initially proposed by De Castro and Gamboa (2012), see also Bredies and Pikkarainen (2013). The first sharp analysis of the solution of this problem is provided by 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 $w = 0$ 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). A refined stability results is detailed by Duval and Peyré (2015) which shows that conditions based on minimum separation imply *support stability*, which means that when $\|w\|$ and $\|w\|/\lambda$ are small enough, then the solution of $(\mathcal{P}_\lambda(y))$ has the same number of Diracs as μ_0, and that both the amplitudes and positions of the spikes converges smoothly as $w \to 0$. These initial works have been extended by Tang et al. (2013) to the case of randomized compressive measurements of the form (1.1), when using Fourier sketching functions φ_ω. In all these results, the separation condition are given for the Euclidean cases, which is an ad-hoc choice which does not take into account the geometry of the problem, and gives vastly sub-optimal theories for spatially varying operators (such as data-dependent kernels in supervised learning, Gaussian mixture estimation and Laplace transform in imaging, see Section 1.2).

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). 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.

1.2 The Fisher information metric

The empirial covariance operator is defined as $\hat{K}(x, x') \overset{\text{def.}}{=} \frac{1}{m} \sum_i \overline{\varphi_{\omega_i}(x)} \varphi_{\omega_i}(x')$ and the deterministic limit as $m \to +\infty$ is denoted K with

$$K(x, x') \overset{\text{def.}}{=} \int_\Omega \overline{\varphi_\omega(x)} \varphi_\omega(x') \mathrm{d}\Lambda(\omega). \qquad (1.3)$$

Note that many covariance kernels can be written under the form (1.3). By Bochner's theorem, this includes all translation-invariant kernels, for which possible features

are $\varphi_\omega(x) = e^{\mathrm{i}\omega^\top x}$. The associated metric tensor is

$$\mathbf{H}_x \overset{\text{def.}}{=} \nabla_x \nabla_{x'} K(x,x) \in \mathbb{C}^{d\times d}. \tag{1.4}$$

Throughout, we assume that $\mathbf{H}_x$ is positive definite for all $x \in \mathcal{X}$. Then, $\mathbf{H}$ naturally induces a distance between points in our parameter space $\mathcal{X}$. Given a piecewise smooth curve $\gamma : [0,1] \to \mathcal{X}$, the length $\ell_{\mathbf{H}}[\gamma]$ of γ is defined by $\ell_{\mathbf{H}}[\gamma] \overset{\text{def.}}{=} \int_0^1 \sqrt{\langle \mathbf{H}_{\gamma(t)}\gamma'(t), \gamma'(t)\rangle}\mathrm{d}t$. Given two points $x, x' \in \mathcal{X}$, the distance from x to x', induced by $\mathbf{H}$ is $d_{\mathbf{H}}(x,x') \overset{\text{def.}}{=} \inf_{\gamma\in\mathcal{F}} \ell_{\mathbf{H}}[\gamma]$ where $\mathcal{F}$ is the set of all piecewise smooth paths $\gamma : [0,1] \to \mathcal{X}$ with $\gamma(0) = x$ and $\gamma(1) = x'$.

The metric $\mathbf{H}$ is closely linked to the Fisher information matrix (Fisher, 1925) associated with Φ: since (1.2) holds, $f(x,\omega) \overset{\text{def.}}{=} |\varphi_\omega(x)|^2$ can be interpreted as a probability density function for the random variable ω conditional on parameter x, and the metric $\mathbf{H}_x$ is equal (up to rescaling) to its Fisher information matrix, since

$$\begin{aligned}&\int \nabla\left(\log f(x,\omega)\right)\nabla\left(\log f(x,\omega)\right)^\top f(x,\omega)\mathrm{d}\Lambda(\omega)\\ &= 4\,\mathbb{E}_\omega[\mathrm{Re}\left(\overline{\nabla\varphi_\omega(x)}\nabla\varphi_\omega(x)^\top\right)] = 4\mathbf{H}_x.\end{aligned}$$

The distance $d_{\mathbf{H}}$ is called the "Fisher-Rao" geodesic distance (Rao, 1945) and is used extensively in information geometry for estimation and learning problems on parametric families of distributions (Amari and Nagaoka, 2007). The Fisher-Rao is the unique Riemannian metric on a statistical manifold (Cencov, 2000) and it is invariant to reparameterization, which matches the invariance of the BLASSO problem $(\mathcal{P}_\lambda(y))$ to reparameterization of the space $\mathcal{X}$. Although $d_{\mathbf{H}}$ has been used in conjunction with kernel methods (see for instance Burges (1999)), to the best of our knowledge, it is the first time this metric is put forward to analyze the performance of off-the-grid sparse recovery problems.

1.2.1 Examples

We detail some popular learning and imaging examples.

The Fejér kernel One of the first seminal result of super-resolution with sparse regularization was given by Candès and Fernandez-Granda (2014) for this kernel, which corresponds to discrete Fourier measurements on the torus. We give a multi-dimensional generalization of this result here. Let $f_c \in \mathbb{N}$, $\mathcal{X} \in \mathbb{T}^d$, $\Omega = \left\{\omega \in \mathbb{Z}^d \,;\, \|\omega\|_\infty \leqslant f_c\right\}$. Let $\varphi_\omega(x) \overset{\text{def.}}{=} e^{\mathrm{i}2\pi\omega^\top x}$ and $\Lambda(\omega) \propto \prod_{j=1}^d g(\omega_j)$ where $g(j) = \frac{1}{f_c}\sum_{k=\max(j-f_c,-f_c)}^{\min(j+f_c,f_c)}(1-|k/f_c|)(1-|(j-k)/f_c|)$. Note that this corresponds to sampling *discrete* Fourier frequencies. Then, the associated kernel is the Fejér kernel $K(x,x') = \prod_{i=1}^d \kappa(x_i - x'_i)$, where $\kappa(x) \overset{\text{def.}}{=} \mathrm{sinc}_{f_c/2+1}^4(x)$ where $\mathrm{sinc}_s(x) \overset{\text{def.}}{=} s^{-1}\sin(\pi s x)/\sin(\pi x)$, which has a constant metric tensor $\mathbf{H}_x = C_{f_c}\mathrm{Id}$ and $d_{\mathbf{H}}(x,x') = \sqrt{C_{f_c}}\,\|x-x'\|_2$ is a scaled Euclidean metric (quotiented by the action of translation modulo 1 on $\mathbb{T}^d$), where $C_{f_c} = -\kappa''(0) = \frac{\pi^2 f_c(f_c+4)}{3}$.

The Gaussian kernel Let $\Sigma \in \mathbb{R}^{d\times d}$ be a positive semidefinite matrix, $\mathcal{X} \subseteq \mathbb{R}^d$ and $\Omega = \mathbb{R}^d$. Let $\varphi_\omega(x) = e^{\mathrm{i}\omega^\top x}$ and $\Lambda(\omega) = \mathcal{N}(0,\Sigma^{-1})$, the centered Gaussian distribution with covariance Σ^{-1}. This can be interpreted as sampling *continuous* Fourier frequencies. Then, the associated kernel is $K(x,x') = e^{-\frac{1}{2}\|x-x'\|_{\Sigma^{-1}}^2}$ where $\|x\|_\Sigma = \sqrt{x^\top\Sigma x}$, with constant metric $\mathbf{H}_x = \Sigma^{-1}$, and $d_{\mathbf{H}}(x,x') = \|x-x'\|_{\Sigma^{-1}}$. In Section 3, we also detail how to exploit this kernel for Gaussian Mixture Model (GMM) estimation with the BLASSO.

The Laplace transform Let $\bar\alpha = (\alpha_j) \in \mathbb{R}_+^d$, $\mathcal{X} \subseteq (0,+\infty)^d$ and $\Omega = \mathbb{R}_+^d$. A (sampled) Laplace transform is defined by setting $\varphi_\omega(x) = \prod_{i=1}^d \sqrt{\frac{2(x_i+\alpha_i)}{\alpha_i}}e^{-\langle x,\omega\rangle}$ and $\Lambda(\omega) = \prod_{j=1}^d (2\alpha_j)e^{-\langle 2\bar\alpha,\omega\rangle}$. Then, $K(x,x') = \prod_{i=1}^d \kappa(x_i+\alpha_i, x'_i+\alpha_i)$ where $\kappa(a,b) = \frac{2\sqrt{ab}}{a+b}$, with metric $\mathbf{H}_x$ as the diagonal matrix with diagonal $\left((2(x_i+\alpha_i))^{-2}\right)_{i=1}^d$ and distance $d_{\mathbf{H}}(x,x') = \sqrt{\sum_i \left|\log\left(\frac{x_i+\alpha_i}{x'_i+\alpha_i}\right)\right|^2}$. We remark that this kernel, associated to the Laplace transform (which should not be confused with the translation-invariant Laplace kernel $\exp(-\|x-x'\|)$) appears in some microscopy imaging technique, see for instance Boulanger et al. (2014). Unlike the previous examples, it is not translation-invariant, and therefore the metric $\mathbf{H}_x$ is not constant. Our results show that the corresponding Fisher metric is the natural way to impose the separation condition in super-resolution.

1.3 Contributions.

Our main contribution is Theorem 1, which states that if the sought after spikes positions X_0 are sufficiently separated with respect to the Fisher distance $d_{\mathbf{H}}$, then the solution to $(\mathcal{P}_\lambda(y))$ is *support stable* (that is, the solution of the BLASSO is formed of exactly s Diracs) provided that the number of random noisy measurements m is, up to log factors and under the assumption of random signs of the amplitudes a_0, linear in s, and the noise level $\|w\|$ is less than $1/s$. In the case of translation invariant kernels, this generalizes existing results to a large class of multi-dimensional kernels, and also provides for the first time a quantitative bounds on the impact of the noise and sub-sampling on the spikes positions and amplitudes errors. For non-translation kernels, this provides for the first time a meaningful

support recovery guarantee, a typical example being the Laplace kernel (see Section 1.2).

2 Key concepts

Notation for derivatives. Given $f \in \mathscr{C}^\infty(\mathcal{X})$, by interpreting the r^{th} derivative as a multilinear map: $\nabla^r f : (\mathbb{C}^d)^r \to \mathbb{C}$, so given $Q \overset{\text{def.}}{=} \{q_\ell\}_{\ell=1}^r \in (\mathbb{C}^d)^r$,

$$\nabla^r f[Q] = \sum_{i_1,\cdots,i_r} \partial_{i_1} \cdots \partial_{i_r} f(x) q_{1,i_1} \cdots q_{r,i_r}.$$

and we define the r^{th} normalized derivative of f as

$$\mathrm{D}_r\left[f\right](x)[Q] \overset{\text{def.}}{=} \nabla^r f(x)[\{\mathbf{H}_x^{-\frac{1}{2}} q_i\}_{i=1}^r]$$

with norm $\left\|\mathrm{D}_r\left[f\right](x)\right\| \overset{\text{def.}}{=} \sup_{\forall \ell, \|q_\ell\| \leqslant 1} \left|\mathrm{D}_r\left[f\right](x)[Q]\right|$. For $i, j \in \{0,1,2\}$, let $K^{(ij)}(x,x')$ be a "bi"-multilinear map, defined for $Q \in (\mathbb{C}^d)^i$ and $V \in (\mathbb{C}^d)^j$ as

$$[Q]K^{(ij)}(x,x')[V] \overset{\text{def.}}{=} \mathbb{E}[\overline{\mathrm{D}_i\left[\varphi_\omega\right](x)[Q]}\mathrm{D}_j\left[\varphi_\omega\right](x')[V]]$$

and $\left\|K^{(ij)}(x,x')\right\| \overset{\text{def.}}{=} \sup_{Q,V} \left\|[Q]K^{(ij)}(x,x')[V]\right\|$ where the supremum is defined over all $Q \overset{\text{def.}}{=} \{q_\ell\}_{\ell=1}^i$, $V \overset{\text{def.}}{=} \{v_\ell\}_{\ell=1}^j$ with $\|q_\ell\| \leqslant 1$, $\|v_\ell\| \leqslant 1$. Note that $\mathrm{D}_2\left[f\right](x)$ and $K^{(02)}(x,x')$ can also be interpreted as a matrix in $\mathbb{C}^{d\times d}$, and we have the normalization $K^{(02)}(x,x) = -\mathrm{Id}$ for all x.

2.1 Admissible kernel and separation

In previous studies on the recovery properties of $(\mathcal{P}_\lambda(y))$ (Candès and Fernandez-Granda, 2014; Bhaskar et al., 2013; Bendory et al., 2016; Duval and Peyré, 2015; Fernandez-Granda, 2016), recovery bounds are attained in the context of K being *admissible* and a separation condition on the underlying positions $\{x_j\}_j$. Namely, given $X = \{x_j\}_j$, that $\min_{i\neq j} d_{\mathbf{H}}(x_i, x_j)$ is sufficiently large with respect to the decay properties of K. For example, in the case where Φ corresponds to Fourier sampling on a grid, up to frequency f_c, this separation condition is $\min_{j\neq\ell} \|x_j - x_\ell\|_2 \gtrsim 1/f_c$. In fact, if $\mathrm{sign}(a_j)$ can take arbitrary values in $\{+1,-1\}$, this separation condition is a necessary to ensure exact recovery for the BLASSO (Tang, 2015).

Following the aforementioned works, we introduce the notion of an admissible kernel.

Definition 1. *A kernel K will be said* admissible *with respect to $\mathcal{K} \overset{\text{def.}}{=} \{r_{\mathrm{near}}, \Delta, \varepsilon_i, B_{ij}, s_{\max}\}$, where $0 < r_{\mathrm{near}} < \Delta/4$ is a neighborhood size, $\varepsilon_0 \in (0,1)$, $\varepsilon_2 \in (0, r_{\mathrm{near}}^{-2})$ are respectively a distance to 1 and a curvature, $\Delta > 0$ is a minimal separation, $B_{ij} > 0$ for $i,j = 0,\ldots,2$ are some constants and $s_{\max} \in \mathbb{N}^*$ is a maximal sparsity level, if*

1. ***Uniform bounds**: For $(i,j) \in \{(0,0),(1,0)\}$, $\sup_{x,x'\in\mathcal{X}} \|K^{(ij)}(x,x')\| \leqslant B_{ij}$; for $(i,j) \in \{(0,2),(1,1),(1,2)\}$ and all x,x' such that $d_{\mathbf{H}}(x,x') \leqslant r_{\mathrm{near}}$ or $d_{\mathbf{H}}(x,x') > \Delta/4$, $\|K^{(ij)}(x,x')\| \leqslant B_{ij}$; and finally, $\sup_{x\in\mathcal{X}} \left\|K^{(22)}(x,x)\right\| \leqslant B_{22}$.*
2. ***Neighborhood of each point**: For all $x \in \mathcal{X}$, $K(x,x) = 1$ and for all $x,x' \in \mathcal{X}$ with $d_{\mathbf{H}}(x,x') \leqslant r_{\mathrm{near}}$, $\mathrm{Re}\left(K^{(02)}(x,x')\right) \preccurlyeq -\varepsilon_2 \mathrm{Id}$ and $\left\|\mathrm{Im}\left(K^{(02)}(x,x')\right)\right\| \leqslant c\varepsilon_2$, where $c \overset{\text{def.}}{=} \frac{1}{2}\sqrt{\frac{2-\varepsilon_2 r_{\mathrm{near}}^2}{\varepsilon_2 r_{\mathrm{near}}^2}}$ and for $d_{\mathbf{H}}(x,x') \geqslant r_{\mathrm{near}}$, $|K(x,x')| \leqslant 1-\varepsilon_0$.*
3. ***Separation**: For $d_{\mathbf{H}}(x,x') \geqslant \Delta/4$, for all $i,j \in \{0,\ldots,2\}$ with $i+j \leqslant 3$, $\|K^{(ij)}(x,x')\| \leqslant \frac{h}{s_{\max}}$, where $h \overset{\text{def.}}{=} \min_{i\in\{0,2\}} \left(\frac{\varepsilon_i}{32B_{1i}+32}, \frac{5\varepsilon_2}{16B_{12}+24}\right)$.*

Additionally, there exists $C_{\mathbf{H}} \geqslant 0$ such that for $d_{\mathbf{H}}(x,x_0) \leqslant r_{\mathrm{near}}$: $\left\|\mathrm{Id} - \mathbf{H}_{x_0}^{-\frac{1}{2}} \mathbf{H}_x^{\frac{1}{2}}\right\| \leqslant C_{\mathbf{H}} d_{\mathbf{H}}(x,x_0)$. We also denote $d_{\mathbf{H}}(X,X_0) = \sqrt{\sum_i d_{\mathbf{H}}(x_i,x_{0,i})^2}$ and $B \overset{\text{def.}}{=} \sum_{i+j\leqslant 3} B_{ij}$ and $\varepsilon \overset{\text{def.}}{=} \min\{\varepsilon_0,\varepsilon_2\}$.

Intuitively, these three conditions express the following facts: 1) the kernel and its derivatives are uniformly bounded, 2) near $x = x'$, the kernel has negative curvature, and otherwise it is strictly less than 1, and 3) for x and x' sufficiently separated, the kernel and all its derivatives have a small value.

2.2 Almost bounded random features

Ideally, we would like our features and its derivatives to be uniformly bounded for all ω. However this may not be the case: think of $e^{i\omega^\top x}$ where the support of the distribution Λ is not bounded. Hence our results will be dependent on the probability that the derivatives are greater than some value T decays sufficiently quickly as T increases. In the following, for $r \in \{0,1,2,3\}$, $L_r(\omega) \overset{\text{def.}}{=} \sup_{x\in\mathcal{X}} \left\|\mathrm{D}_r\left[\varphi_\omega\right](x)\right\|$, and let F_r be such that $\mathbb{P}_\omega\left(L_r(\omega) > t\right) \leqslant F_r(t)$.

2.3 Key assumptions

Our main result will be valid under the following assumptions.

I. On the domain and limit kernel Let $\mathcal{X}$ be a compact domain with radius $R_{\mathcal{X}} \overset{\text{def.}}{=} \sup_{x,x'\in\mathcal{X}} d_{\mathbf{H}}(x,x')$. Assume the kernel is admissible wrt $\mathcal{K} \overset{\text{def.}}{=} \{r_{\mathrm{near}}, \Delta, \varepsilon_i, B_{ij}, s_{\max}\}$.

II. Assumption on the underlying signal For $s \leqslant s_{\max}$, let $a_0 \in \mathbb{C}^s$ and let $X_0 \overset{\text{def.}}{=} (x_{0,j})_{j=0}^s$ be such that $d_{\mathbf{H}}(x_{0,i}, x_{0,j}) \geqslant \Delta$ for $i \neq j$. The underlying measure is assumed to be $\mu_0 = \sum_{j=1}^s a_{0,j}\delta_{x_{0,j}}$.

III. Assumption on the sampling complexity For $\rho > 0$, suppose that $m \in \mathbb{N}$ and $\{\bar{L}_i\}_{i=0}^3 \in \mathbb{R}_+^4$ are chosen such that

$$\sum_{j=0}^{3} F_j(\bar{L}_j) \leqslant \frac{\rho}{m}, \qquad \text{and} \qquad \max_{j=0}^{3}\{\bar{L}_j^2 \sum_{i=0}^{3} F_i(\bar{L}_i) + 2\int_{\bar{L}_j}^{\infty} t F_j(t)\mathrm{d}t\} \leqslant \frac{\varepsilon}{m}, \tag{2.1}$$

and either one of the following hold:

$$m \gtrsim C \cdot s \cdot \log\left(N^d/\rho\right) \log\left(s/\rho\right), \tag{2.2}$$

$$\text{or} \quad m \gtrsim C \cdot s^{3/2} \cdot \log\left(N^d/\rho\right), \tag{2.3}$$

where $C \overset{\text{def.}}{=} \varepsilon^{-2}(\bar{L}_2^2 B_{11} + \bar{L}_1^2 B_{22} + (B_0 + B_2)\bar{L}_{01}^2)$, $N \overset{\text{def.}}{=} \mathbb{L}_3 d R_{\mathcal{X}} (r_{\text{near}}\varepsilon)^{-1}$ and $\mathbb{L}_r = \max_{i=1}^{r} \bar{L}_i$.

Remark 1. *Our main theorem presents support stability guarantees under the sampling complexity rate* (2.2) *if* $\operatorname{sign}(a_0) = (a_{0,i}/\left|a_{0,i}\right|)_{i=1}^s$ *forms a Steinhaus sequence, that is, iid uniformly distributed on the complex unit circle. This assumption has been used before in compressed sensing (Candès and Romberg, 2007; Tang et al., 2013) to achieve this optimal complexity (see also Foucart and Rauhut (2013), Chap. 14). As noted in previous works, this random signs assumption is likely to be a proof artefact, however achieving optimal complexity without it may require more involved arguments (Candes and Plan, 2011). When the signs are arbitrary, we prove our results under* (2.3)*. Although this* $s^{3/2}$ *scaling is still sub-optimal in* s*, we remark it improves upon the previous theoretical rate of* s^2 *(up to log factors) (Li and Chi, 2017).*

Remark 2. *The assumption on the choice of* $\bar{L}_r$ *ensures that with high probability,* $\mathrm{D}_r\left[\varphi_\omega\right](x)$ *is uniformly bounded up to* $r = 3$*. Note also that, generally, the* $\{\bar{L}_r\}$ *depend on* m*, through* (2.1)*. However, in all our examples: either a)* $\sup_{x\in\mathcal{X}} \left\|\mathrm{D}_r\left[\varphi_\omega\right](x)\right\|$ *are already uniformly bounded, in which case* $\bar{L}_i$ *can be chosen independently of* ρ *and* m *(for instance this is the case of the Fejér kernel); or b) the* $F_r(t)$ *are exponentially decaying, in which case we can show that* $\bar{L}_r = \mathcal{O}(\log(m/\rho)^p)$ *for some* $p > 0$*, which only incurs additional logarithmic terms on the bounds* (2.2) *and* (2.3)*. This is the case of the Gaussian or Laplace transform kernel.*

3 Main result

Our main theorem below states quantitative exact support recovery bounds under a minimum separation condition according to $d_{\mathbf{H}}$.

Theorem 1. *Let* $\rho > 0$*, suppose that* K *is admissible, and that* a_0*,* X_0*,* m *and* $\bar{L}_i$ *satisfy the assumptions of Section 2.3. Let* $\mathcal{D}_{\lambda_0,c_0} \overset{\text{def.}}{=} \{(\lambda, w) \in \mathbb{R}_+ \times \mathbb{C}^m \,;\, \lambda \leqslant \lambda_0,\ \|w\| \leqslant c_0\lambda\}$ *where* $c_0 \sim \min\left(\frac{\varepsilon_0}{\bar{L}_0},\ \frac{\varepsilon_2}{\bar{L}_2}\right)$ *and* $\lambda_0 \sim D/s$ *with*

$$D \overset{\text{def.}}{=} \underline{a} \min\left(r_{\text{near}}\sqrt{s},\ \tfrac{\varepsilon\sqrt{s}}{\mathbb{L}_2^2\|a\|},\ \tfrac{\varepsilon}{C_{\mathbf{H}}(B+\mathbb{L}_2^2)}\right) \tag{3.1}$$

where $\underline{a} = \min\{\left|a_{0,i}\right|, \left|a_{0,i}\right|^{-1}\}$*. Suppose that either* $\operatorname{sign}(a_0)$ *is a Steinhaus sequence and* m *satisfies* (2.2) *or* $\operatorname{sign}(a_0)$ *is an arbitrary sign sequence and* m *satisfies* (2.3)*. Then, with probability at least* $1 - \rho$*,*

(i) for all $v \overset{\text{def.}}{=} (\lambda, w) \in \mathcal{D}_{\lambda_0,c_0}$*,* $(\mathcal{P}_\lambda(y))$ *has a unique solution which consists of exactly* s *spikes. Moreover, up to a permutation of indices, the solution can be written as* $\sum_{i=1}^s a_i^v \delta_{x_i^v}$*, and* $\operatorname{sign}(a_i^v) = \operatorname{sign}(a_{0,i})$ *for all* $i = 1, \ldots, s$

(ii) The mapping $v \in \mathcal{D}_{\lambda_0,c_0} \mapsto (a^v, X^v)$ *is* $\mathscr{C}^1$ *and we have the error bound*

$$\|a^v - a_0\| + d_{\mathbf{H}}(X^v, X_0) \leqslant \tfrac{\sqrt{s}(\lambda + \|w\|)}{\min_i |a_{0,i}|} \tag{3.2}$$

We detail below the values relating to the sampling complexity corresponding to each of the examples detailed in Section 1.2.1. The corresponding proofs can be found in Section F of the appendix.

Discrete Fourier sampling The Fejer kernel of order $f_c \geqslant 128$ is admissible with $\Delta = \mathcal{O}(\sqrt{d}\sqrt[4]{s_{\max}})$, $r_{\text{near}} = 1/(8\sqrt{2})$, $\varepsilon_0 = 0.00097$, $\varepsilon_2 = 0.941$, $B_{01} = \mathcal{O}(d)$, $B_{11} = B_{02} = B_{12} = \mathcal{O}(1)$ and $B_{22} = \mathcal{O}(d)$. Moreover, $\bar{L}_r = \mathcal{O}(d^{r/2})$. Hence, up to logarithmic terms, Thm. 1 is applicable with $m = \mathcal{O}(sd^3)$ when the random signs assumption holds, and $m = \mathcal{O}(s^{\frac{3}{2}}d^3)$ in the general case, with guaranteed support stability when $\lambda = \mathcal{O}(s^{-1}d^{-2})$, $\|w\| = \mathcal{O}(s^{-1}d^{-3})$. Note that our choice of Δ imposes that $\|x_i - x_j\|_2 \gtrsim \sqrt{d}s_{\max}^{1/4}/f_c$ whereas the previous result of Candès and Fernandez-Granda (2014) requires $\|x_i - x_j\|_\infty \gtrsim C_d/f_c$ with no dependency in $s_{\max}$, however, their proof would imply that the constant C_d grows exponentially in d. Since we are interested in having a general theory in arbitrary dimension, we have opted to present a polynomial dependency on $s_{\max}$.

Continuous Gaussian Fourier sampling In the appendix we prove that the kernel is admissible with $\Delta = \mathcal{O}\left(\sqrt{\log s_{\max}}\right)$, $r_{\text{near}} = 1/\sqrt{2}$, $\varepsilon_0 = 1 - e^{-\frac{1}{4}}$, $\varepsilon_2 = e^{-\frac{1}{4}}/2$, $B_{ij} = \mathcal{O}(1)$ for $i + j \leqslant 3$, $B_{22} = \mathcal{O}(d)$ and $\bar{L}_r = \left(d + \log\left(\frac{dm}{\rho}\right)^2\right)^{\frac{r}{2}}$ (as mentioned before, the dependence in m only incurs additional logarithmic factors in (2.2) and (2.3)). Hence, up to log factors, the

sample complexity and noise level for the application of Thm. 1 is the same as for the Fejér kernel.

Laplace sampling The associated kernel is admissible with $\Delta = \mathcal{O}(d + \log(ds_{\max}))$, $r_{\text{near}} = 0.2$, $\varepsilon_0 = 0.005$, $\varepsilon_2 = 1.52$, $B_{ij} = \mathcal{O}(1)$ for $i + j \leqslant 3$ and $B_{22} = \mathcal{O}(d)$. Define $\bar{R}_{\mathcal{X}} = \left(1 + \frac{R_{\mathcal{X}}}{\min_i \alpha_i}\right)^d$ (where we recall that $R_{\mathcal{X}}$ is the radius of $\mathcal{X}$). Assuming for simplicity that all α_j are distinct, we can set $\bar{L}_r = \bar{R}_{\mathcal{X}}(R_{\mathcal{X}} + \|\alpha\|_\infty)^r \left(\sqrt{d} + \max_i \frac{1}{\alpha_i} \log\left(\frac{d\beta_i m \bar{R}_{\mathcal{X}}}{\rho \alpha_i}\right)\right)^r$ Hence, choosing $\alpha_i \sim d$, we have that $\bar{R}_{\mathcal{X}} = (1)$ and up to log factors, (2.2) is $\mathcal{O}(sd^7)$ and (2.3) is $\mathcal{O}(s^{3/2}d^7)$, and support stability is guaranteed when $\lambda = \mathcal{O}(s^{-1}d^{-3})$ and $\|w\| = \mathcal{O}(s^{-1}d^{-5})$. Note that despite the stronger dependency on d, for practical applications (microscopy), one is typically only interested in the low dimensional setting of $d = 2, 3$.

Gaussian mixture learning Consider n datapoints $z_1, \ldots, z_n \in \mathbb{R}^d$ drawn *iid* from a mixture of Gaussians $\sum_i a_{0,i} \mathcal{N}(x_{0,i}, \Sigma)$ with means $x_{0,i} \in \mathcal{X} \subset \mathbb{R}^d$ and known covariance Σ, where $\mathcal{X}$ is bounded. Consider the following procedure:

- draw ω_j *iid* from $\mathcal{N}(0, \Sigma^{-1}/d)$ (the $1/d$ normalization is necessary to avoid an exponential dependency in d later on)
- compute the generalized moments $y = \frac{1}{\sqrt{m}} \sum_{i=1}^n (e^{i\langle \omega_j, x_i \rangle})_{j=1}^m$
- solve the BLASSO with features $\varphi_\omega(x) = e^{i\langle \omega, x \rangle} e^{-\frac{1}{2}\|\omega\|_\Sigma^2}$, to obtain a distribution $\tilde{\mu}$

Then, as described in the introduction, we can interpret y as noisy Fourier measurements of $\mu_0 = \sum_i a_{0,i} \delta_{x_{0,i}}$ *in the space of means* $\mathcal{X}$, where the "noise" w corresponds to using the empirical average over the z_i instead of a true integration. It is easily bounded with probability $1 - \rho$ by $\|w\| \leqslant \mathcal{O}\left(\sqrt{\frac{\log(1/\rho)}{n}}\right)$, by a simple application of Hoeffding's inequality (Gribonval et al., 2017).

The associated kernel is the Gaussian kernel with covariance $(2 + d)\Sigma$ and hence, our result states that, if $\|x_i - x_j\|_{\Sigma^{-1}} \geqslant \sqrt{d \log s}$, and the number of measurements and sample complexity satisfy, up to logarithmic terms, $m = \mathcal{O}\left(s^{\frac{3}{2}} d^3\right)$, $n = \mathcal{O}\left(s^2 d^6 / \min_i |a_{0,i}|^2\right)$ and $\lambda_0 = \mathcal{O}\left(\frac{\min_i |a_{0,i}|}{\sqrt{s} d^2 \|a_0\|_2}\right)$, then, with probability $1 - \rho$ on both samples z_j and frequencies ω_j, the distribution $\tilde{\mu}$ is formed of exactly s Diracs, and their positions and weights converge to the means and weights of the GMM. Let us give a few remarks on this result.

On model selection. Besides convexity (with respect to the distribution of means) of the BLASSO, which is not the case of classical likelihood- or moments-based methods for learning GMM, the most striking feature of our approach is probably the support stability: with a sample complexity that is polynomial in s and d, the BLASSO yields *exactly* the right number of components for the GMM. Despite the huge literature on model selection for GMM, to our knowledge, this is one of the only result which is *non-asymptotic* in sample complexity, as opposed to many approaches (Roeder and Wasserman, 1997; Huang et al., 2013) which guarantee that the selected number of components approaches the correct one when the number of samples grows to infinity.

On separation condition. Our separation condition of $\sqrt{d \log s}$ is, up to the logarithmic term, similar to the $\sqrt{d}$ found in the seminal work by Dasgupta (1999). This was later improved by different methods (Dasgupta and Schulman, 2000; Vempala and Wang, 2004), until the most recent results on the topic (Moitra and Valianty, 2010) show that it is possible to learn a GMM with *no* separation condition, provided the sample complexity is exponential in s, which is a necessary condition (Moitra and Valianty, 2010). As mentioned in the introduction, similar results exist for the BLASSO: Denoyelle et al. (2017) showed that in one dimension, one can identify s positive spikes with no separation, provided the noise level is exponentially small with s. Hence learning GMM with the BLASSO and no separation condition may be feasible, which we leave for future work, however we note that the multi-dimensional case is still largely an open problem (Poon and Peyré, 2017).

On known covariance. An important path for future work is to handle arbitrary covariance. When the components all share the same mean and have diagonal covariance, the Fisher metric is related, up to a change of variables, to the Laplace transform kernel case treated earlier. When both means and covariance vary, in one dimension, the Fisher metric is related to the Poincaré half-plane metric (Costa et al., 2015). In the general case, it does not have a closed-form expression. We leave the treatment of these cases for future work.

4 Sketch of proof

4.1 Background on dual certificates

Our approach to establishing that the solutions to $(\mathcal{P}_\lambda(y))$ are support stable is via the study of the associated dual solutions in accordance to the framework introduced in Duval and Peyré (2015). We first recall some of their key ideas. In order to study the support stability properties of $(\mathcal{P}_\lambda(y))$ in the small noise regime, we consider the limit problem as $\lambda \to 0$ and $\|w\| \to 0$,

that is

$$\min_{\mu\in\mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) \text{ subject to } \Phi\mu = y. \qquad (\mathcal{P}_0(y))$$

The dual of $(\mathcal{P}_\lambda(y))$ and $(\mathcal{P}_0(y))$ are

$$\min_p \left\{ \|y/\lambda - p\|_2^2 \ ; \ \|\Phi^* p\|_\infty \leqslant 1 \right\} \qquad (\mathcal{D}_\lambda(y))$$

$$\max_p \left\{ \langle y, p\rangle \ ; \ \|\Phi^* p\|_\infty \leqslant 1 \right\}. \qquad (\mathcal{D}_0(y))$$

Any solution μ_λ of $(\mathcal{P}_\lambda(y))$ to related to the (unique) solution p_λ of $(\mathcal{D}_\lambda(y))$ by $-p_\lambda = \frac{1}{\lambda}(\Phi\mu_\lambda - y)$ and writing $\eta_\lambda \overset{\text{def.}}{=} \Phi^* p_\lambda$, $\langle \eta_\lambda, \mu_\lambda\rangle = |\mu_\lambda|(\mathcal{X})$. Note that $\mathrm{Supp}(\mu_\lambda) \subseteq \{x \in \mathcal{X} \ ; \ |\Phi^* p_\lambda(x)| = 1\}$, so η_λ "certifies" the support of μ_λ and is often referred to as a *dual certificate*. Furthermore, by defining the minimal norm certificate η_0 as $\eta_0 \overset{\text{def.}}{=} \Phi^* p_0$ where

$$p_0 = \mathrm{argmin}\left\{ \|p\|_2 \ ; \ p \text{ is a solution to } (\mathcal{D}_0(y)) \right\} \qquad (4.1)$$

one can show that p_λ converges as $\lambda \to 0$ to p_0 and hence η_λ converges to $\eta_0 \overset{\text{def.}}{=} \Phi^* p_0$ in L^∞. When λ and $\|w\|$ are sufficiently small, solutions to $(\mathcal{P}_\lambda(y))$ are support stable provided that η_0 (called the minimal norm certificate) is *nondegenerate*, that is $\eta_0(x_i) = \mathrm{sign}(a_i)$ for $i = 1, \dots, s$ and $\nabla^2 |\eta_0|^2 (x_i)$ is negative definite. This is proven to be an almost sharp condition for support stability, since Duval and Peyré (2017) provided explicit examples where $|\eta_0(x)| = 1$ for some $x \notin \{x_i\}_i$ implies that $(\mathcal{P}_\lambda(y))$ recovers more than s spikes under arbitrarily small noise.

Pre-certificates In practice, the minimal norm certificate is hard to compute and analyse due to the nonlinear ℓ^∞ constraint in (4.1). So, one often introduces a proxy which can be computed in closed form by solving an linear system associated to the following least squares problem: $\eta_X \overset{\text{def.}}{=} \Phi^* p$ where

$$p_X \overset{\text{def.}}{=} \mathrm{argmin}\{\|p\|_2 \ ; \ (\Phi^* p)(x_i) = \mathrm{sign}(a_i), \ \nabla(\Phi^* p)(x_i) = 0\}. \qquad (4.2)$$

Note that if η_X satisfies $\|\eta_X\|_\infty \leqslant 1$, then $\eta_X = \eta_0$.

Computation of η_X For $x \in \mathcal{X}$, let $\varphi(x) \overset{\text{def.}}{=} \frac{1}{\sqrt{m}}(\varphi_{\omega_k}(x))_{k=1}^m$. For $X = \{x_i\}_{i=1}^s$ we define $\Gamma_X : \mathbb{C}^{s(d+1)} \to \mathbb{C}^m$ as $\Gamma_X([\alpha, \beta]) \overset{\text{def.}}{=} \sum_{i=1}^s \alpha_i \varphi(x_i) + \nabla\varphi(x_i)^\top \beta_i$ where $\nabla\varphi \in \mathbb{C}^{m\times d}$. Then, the minimizer of (4.2) is $p_X = \Gamma_X^{*,\dagger} \binom{\mathrm{sign}(a)}{\mathbf{0}_{sd}}$. Furthermore, when Γ_X is full rank, we can write $\hat{\eta}_X(x) \overset{\text{def.}}{=} \sum_i \hat{\alpha}_i \hat{K}(x_i, x) + \langle \hat{\beta}_i, \nabla_1 \hat{K}(x_i, x)\rangle$, where $\hat{\alpha}_i \in \mathbb{C}$, $\hat{\beta}_i \in \mathbb{C}^d$ are such that $\binom{\hat{\alpha}}{\hat{\beta}} = (\Gamma_X^* \Gamma_X)^{-1} \binom{\mathrm{sign}(a)}{0_{sd}}$, and the hat notation refers to the fact that we are using sub-sampled measurements. The limit precertificate is defined as $\eta_X(x) \overset{\text{def.}}{=} \sum_i \alpha_i K(x_i, x) + \langle \beta_i, \nabla_1 K(x_i, x)\rangle$, where $\binom{\alpha}{\beta} = (\mathbb{E}[\Gamma_X^* \Gamma_X])^{-1} \binom{\mathrm{sign}(a)}{0_{sd}}$.

The key to establishing our recovery results is to show that $\hat{\eta}_X$ is nondegenerate. In this paper, we will actually prove a stronger notion of nondegeneracy:

Definition 2. *Let $a \in \mathbb{C}^s$, $X = \{x_i\}_{i=1}^s \in \mathcal{X}^s$ for some $s \in \mathbb{N}$, and $\varepsilon_0, \varepsilon_2, r > 0$. We say that $\eta \in \mathscr{C}^1(\mathcal{X})$ is $(\varepsilon_0, \varepsilon_2)$-nondegenerate with respect to a, X and r if for all i, $\eta(x_i) = \mathrm{sign}(a_i)$, $\nabla\eta(x_i) = 0$ and*

$$\forall\, x \in \mathcal{X}^{\mathrm{far}}, \ |\eta(x)| \leqslant 1 - \varepsilon_0$$
$$\forall\, x \in \mathcal{X}_j^{\mathrm{near}}, |\eta(x)| \leqslant 1 - \varepsilon_2 d_{\mathbf{H}}(x, x_j)^2$$

where $\mathcal{X}_j^{\mathrm{near}} \overset{\text{def.}}{=} \{x \in \mathcal{X} \ ; \ d_{\mathbf{H}}(x_i, x) \leqslant r\}$ and $\mathcal{X}^{\mathrm{far}} \overset{\text{def.}}{=} \mathcal{X} \setminus \bigcup_{j=1}^s \mathcal{X}_j^{\mathrm{near}}$.

Our proof proceeds in three steps:

1. Show that under admissibility of the kernel and sufficient separation, the limit precertificate η_{X_0} is non-degenerate (see Theorem 2).
2. Show that this non-degeneracy transfers to $\hat{\eta}_X$ when m is large enough and X is close to X_0. This is the purpose of Section 4.3.
3. As discussed, nondegeneracy of $\hat{\eta}_{X_0}$ automatically guarantees support stability when $(\lambda, w) \in \mathcal{D}_{\lambda_0, c_0}$ for λ_0 and c_0 *sufficiently small*. To conclude we simply need to quantify λ_0 and c_0. This is the purpose of Section 4.4. In particular, given (λ, w), we construct a candidate solution by means of (a quantitative version of) the Implicit Function Theorem, and show that it is indeed a true solution using the previous results.

4.2 Non-degeneracy of the limit certificate

Our first result shows that the "limit precertificate" η_{X_0} is nondegenerate:

Theorem 2. *Assume the kernel is admissible wrt $\mathcal{K}$ (see Definition 1). Then, for $s \leqslant s_{\max}$, for all $a = (a_j)_{j=1}^s \in \mathbb{C}^s$ and $X = \{x_j\}_{j=1}^s \in \mathcal{X}^s$ such that $d_{\mathbf{H}}(x_i, x_j) \geqslant \Delta$, the function η_{X_0} is $(\frac{\varepsilon_0}{2}, \frac{\varepsilon_2}{2})$-nondegenerate with respect to a, X and r_{near}.*

The proof of this result can be found in Appendix B and is a generalization of the arguments of Candès and Fernandez-Granda (2014) (see also Bendory et al. (2016)). We remark that unlike previous works which focus on translation invariant kernels, the Fisher metric provides a natural way to understand the required separation between the points in X and thus open up the possibility of analysing more complex problems such as Laplace transform inversion.

4.3 The randomized setting

For the remainder of this paper, we consider solutions of $(\mathcal{P}_\lambda(y))$ given $y = \Phi\mu_{a_0,X_0} + w$ for some fixed $a_0 \in \mathbb{C}^s$ and $X_0 \in \mathcal{X}^s$. The following result shows that $\hat{\eta}_X$ is nondegenerate for all X close to X_0:

Theorem 3. *Let $\rho > 0$. Under the assumptions of Section 2.3, and assuming that either m satisfies* (2.2) *and* $\mathrm{sign}(a_0)$ *is a Steinhaus sequence, or m satisfies* (2.3) *and* $\mathrm{sign}(a_0)$ *is an arbitrary sign sequence, with probability at least $1-\rho$: for all $X \in \mathcal{X}^s$ such that*

$$d_{\mathbf{H}}(X, X_0) \lesssim \min\left(r_{\mathrm{near}}, \frac{\varepsilon_r}{C_{\mathbf{H}}\sqrt{s}\max\left(B, \bar{L}_{12}\bar{L}_r\right)}\right), \quad (4.3)$$

Γ_X is full rank and $\hat{\eta}_X$ is $(\varepsilon_0/8, \varepsilon_2/8)$-nondegenerate with respect to a_0, X and r_{near}.

The proof of this result is given in Appendix D. We simply make a remark on the proof here: We first prove that $\hat{\eta}_{X_0}$ is nondegenerate by bounding variations between η_{X_0} and $\hat{\eta}_{X_0}$. The proof of this fact is a generalization of the arguments in Tang et al. (2013) to the multidimensional and general operator case. We then exploit the fact the φ is smooth and hence, $\Gamma_X^*\Gamma_X$ satisfies certain Lipschitz properties with respect to X, to bound the local variation between $\hat{\eta}_X$ and $\hat{\eta}_{X_0}$.

4.4 Quantitative support recovery

This final section concludes the proof of Theorem 1 by quantifying the regions for λ and $\|w\|$ for which support stability is guaranteed.

Solution of the noisy BLASSO. Let $\Phi_X : \mathbb{C}^s \to \mathbb{C}^m$ be defined by $\Phi_X a = \sum_{i=1}^s a_i\varphi(x_i)$. Recall that $\mu_{a,X} = \sum_i a_i\delta_{x_i}$ is a solution to the BLASSO with $y = \Phi\mu_{a_0,X_0} + w$ if and only if $\hat{\eta}_\lambda = \Phi^* p_\lambda$, with $p_\lambda = \frac{1}{\lambda}(y - \Phi_X a)$, satisfies $\|\hat{\eta}_\lambda\|_\infty \leqslant 1$ and $\hat{\eta}(x_j) = \mathrm{sign}(a_j)$. In that case, p_λ is the *unique* solution to the dual of the BLASSO. Moreover, if $|\hat{\eta}_\lambda(x)| < 1$ for $x \neq x_i$ and Φ_X is full rank (which follows by Theorem D.2), then $\mu_{a,X}$ is also the unique solution of the primal.

Construction of a solution Following Denoyelle et al. (2017), we define the function $f : \mathbb{C}^s \times \mathcal{X}^s \times \mathbb{R}_+ \times \mathbb{C}^m$ by

$$f(u,v) \overset{\text{def.}}{=} \Gamma_X^*(\Phi_X a - \Phi_{X_0}a_0 - w) + \lambda\begin{pmatrix}\mathrm{sign}(a_0)\\ 0_{sd}\end{pmatrix}$$

where $u = (a, X)$ and $v = (\lambda, w)$. Observe that having $f(u,v) = 0$ ensures the existence of $\hat{\eta}_\lambda$ defined as above that satisfies $\hat{\eta}_\lambda(x_i) = \mathrm{sign}(a_{0,i})$ and $\nabla\hat{\eta}_\lambda(x_i) = 0$. We will use it to construct a non-degenerate solution to $\mathcal{D}_\lambda(y)$ for small λ and $\|w\|$. Now, f is continuously differentiable, with explicit forms of $\partial_v f(u,v)$ and $\partial_u f(u,v)$ given in (E.1) and (E.2) in the appendix, and in particular, letting $u_0 = (a_0, X_0)$, $\partial_u f(u_0, 0) = \Gamma_{X_0}^*\Gamma_{X_0}J_a$, where J_a is the diagonal matrix with $\binom{1}{a} \otimes 1_d \in \mathbb{C}^{s(d+1)}$ along its diagonal and Γ_{X_0} is full rank (with probability at least $1-\rho$) by Theorem D.2. So, $\partial_u f(u_0, 0)$ is invertible and $f(u_0, 0) = 0$. Hence, by the Implicit Function Theorem, there exists a neighbourhood V of 0 in $\mathbb{C} \times \mathbb{C}^m$, a neighbourhood U of u_0 in $\mathbb{C}^s \times \mathcal{X}^s$ and a Fréchet differentiable function $g : V \to U$ such that for all $(u,v) \in U \times V$, $f(u,v) = 0$ if and only if $u = g(v)$. So, to establish support stability for $(\mathcal{P}_\lambda(y))$, we simply need to estimate the size of the neighbourhood V on which g is well defined, and given $(\lambda, w) \in V$, for $(a, Z) = g((\lambda, w))$, to check that the associated certificate $\hat{\eta}_{\lambda,w} \overset{\text{def.}}{=} \Phi^* p_{\lambda,w}$ with $p_{\lambda,w} \overset{\text{def.}}{=} \frac{1}{\lambda}(\Phi_X a - \Phi_{X_0}a_0 - w)$ is nondegenerate.

Indeed, one can prove (see Theorem E.1) that with probability at least $1-\rho$, V contains the ball $B_r(0)$ with radius $r \sim \frac{1}{\sqrt{s}}\min\left(\frac{\min\{r_{\mathrm{near}}, (C_{\mathbf{H}}B)^{-1}\}}{\min_i|a_{0,i}|}, \frac{1}{\bar{L}_{01}\bar{L}_{12}(1+\|a_0\|)}\right)$ and given any $v \in B_r(0)$, $(a, X) = g(v)$ indeed satisfy the error bound (3.2).

Checking that the candidate solution is a true solution It remains to check that $g(\lambda, w)$ defines a valid certificate and is non-degenerate (and hence, $\sum_i a_i\delta_{x_i}$ is the unique solution to $(\mathcal{P}_\lambda(y))$) provided that λ, w satisfy (3.1). Given $(\lambda, w) \in V$, let $(a, X) = g((\lambda, w))$. Define $\hat{\eta}_{\lambda,w} \overset{\text{def.}}{=} \frac{1}{\lambda}\Phi^*(\Phi_X a - \Phi_{X_0}a_0 - w)$ and following Denoyelle et al. (2017), one can show that

$$\hat{\eta}_{\lambda,w} = \hat{\eta}_X + \varphi(\cdot)^\top\Pi_X\frac{w}{\lambda} + \frac{1}{\lambda}\varphi(\cdot)^\top\Pi_X\Phi_{X_0}a_0$$

where Π_X is the orthogonal projection onto $\mathrm{Im}(\Gamma_X)^\perp$.

Note that since we have the error bound (3.2), our choice of λ and $\|w\|$ ensures that (4.3) holds and hence, Theorem D.2 implies that $\hat{\eta}_X$ is nondegenerate with probablity at least $1-\rho$. To conclude, it is sufficient to show that the two remaining terms are sufficiently small, so that $\hat{\eta}_{\lambda,w}$ remains non-degenerate. Under $\bar{E}$, $\|\mathrm{D}_r[\varphi_\omega](\cdot)\| \leqslant \bar{L}_r$, and for any $z \in \mathbb{C}^m$, $\left\|\mathrm{D}_r\left[\varphi^\top z\right]\cdot\right\| \leqslant \bar{L}_r\|z\|$. Therefore, since Π_X is a projection, we have $\left\|\mathrm{D}_r\left[\varphi(\cdot)^\top\Pi_X\frac{w}{\lambda}\right]\right\| \lesssim \varepsilon_r$ when $\|w\|/\lambda \lesssim \varepsilon_r/\bar{L}_r$. Finally, since $\Phi_{X_0}a_0 = \sum_{j=1}^s\varphi(x_{0,j})$, by Taylor expansion of $\varphi(x_{0,j})$ around x_j and applying Π_X (see Lemma E.1 for this computation), we have

$$\left\|\frac{1}{\lambda}\Pi_X\Gamma_{X_0}\begin{pmatrix}a_0\\ 0_{sd}\end{pmatrix}\right\| \leqslant \frac{\bar{L}_2}{\lambda}\|a_0\|_\infty d_H(X, X_0)^2.$$

Since g satisfies (3.2) our choice of $\lambda_0 = \mathcal{O}(s^{-1})$ ensures that we can upper bound this

by $\bar{L}_2 \left\|a_0\right\|_\infty \frac{s\left(\lambda+\|w\|^2/\lambda\right)}{\min|a_{0,i}|^2} \lesssim \varepsilon$ and consequently, $\frac{1}{\lambda}\left\|\mathrm{D}_r\left[\varphi(\cdot)^\top \Pi_X \Phi_{X_0} a_0\right]\right\| \lesssim \varepsilon_r$.

Acknowledgements

We would like to thank Ben Adcock for a helpful conversation regarding the stochastic gradient bounds. This work was partly funded by the CFM-ENS chair "Modèles et Sciences des données" and the European Research Council, NORIA project.

References

M. Akkouchi. On the convolution of exponential distributions.

S.-i. Amari and H. Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

J.-M. Azais, Y. De Castro, and F. Gamboa. Spike detection from inaccurate samplings. *Applied and Computational Harmonic Analysis*, 38(2):177–195, 2015.

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

T. Bendory, S. Dekel, and A. Feuer. Robust recovery of stream of pulses using convex optimization. *Journal of mathematical analysis and applications*, 442(2):511–536, 2016.

B. N. Bhaskar, G. Tang, and B. Recht. Atomic norm denoising with applications to line spectral estimation. *IEEE Transactions on Signal Processing*, 61(23):5987–5999, 2013.

J. Boulanger, C. Gueudry, D. Münch, B. Cinquin, P. Paul-Gilloteaux, S. Bardin, C. Guérin, F. Senger, L. Blanchoin, and J. Salamero. Fast high-resolution 3D total internal reflection fluorescence microscopy by incidence angle scanning and azimuthal averaging. *Proceedings of the National Academy of Sciences*, 111(48):17164–17169, 2014.

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

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

C. J. Burges. Geometry and invariance in kernel based methods. 1999.

E. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, 2007.

E. J. Candès and C. Fernandez-Granda. Super-resolution from noisy data. *Journal of Fourier Analysis and Applications*, 19(6):1229–1254, 2013.

E. J. Candès and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956, 2014.

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

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

N. N. Cencov. *Statistical decision rules and optimal inference*. Number 53. American Mathematical Soc., 2000.

S. I. Costa, S. A. Santos, and J. E. Strapasson. Fisher information distance: A geometrical reading. In *Discrete Applied Mathematics*, volume 197, pages 59–69. Elsevier B.V., 2015.

S. Dasgupta. Learning mixtures of Gaussians. In *IEEE 51st Annual Symposium on Foundations of Computer Science*, number May, 1999.

S. Dasgupta and A. Gupta. An Elementary Proof of a Theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, 2003.

S. Dasgupta and L. J. Schulman. A Two-Round Variant of EM for Gaussian Mixtures. *Uncertainty in Artificial Intelligence*, pages 152–159, 2000.

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

Q. Denoyelle, V. Duval, and G. Peyré. Support recovery for sparse super-resolution of positive measures. *to appear in Journal of Fourier Analysis and Applications*, 2017.

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

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

V. Duval and G. Peyré. Sparse spikes super-resolution on thin grids I: the LASSO. *Inverse Problems*, 33(5):055008, 2017.

C. Fernandez-Granda. Support detection in super-resolution. *Proc. Proceedings of the 10th International Conference on Sampling Theory and Applications*, pages 145–148, 2013.

C. Fernandez-Granda. Super-resolution of point sources via convex programming. *Information and Inference: A Journal of the IMA*, 5(3):251–303, 2016.

R. A. Fisher. Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge University Press, 1925.

S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Springer New York, NY, 2013.

R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin. Compressive statistical learning with random feature moments. *arXiv preprint arXiv:1706.07180*, 2017.

T. Huang, H. Peng, and K. Zhang. Model Selection for Gaussian Mixture Models. *Statistica Sinica*, pages 1–27, 2013.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Y. Li and Y. Chi. Stable separation and super-resolution of mixture models. *Applied and Computational Harmonic Analysis*, 2017.

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

S. Minsker. On some extensions of bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119, 2017.

A. Moitra and G. Valianty. Settling the polynomial learnability of mixtures of Gaussians. *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 93–102, 2010.

C. Poon and G. Peyré. Multi-dimensional Sparse Super-resolution. pages 1–42, 2017.

C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1945.

K. Roeder and L. Wasserman. Practical Bayesian density estimation using mixtures of normal. *JASA*, 92:894–902, 1997.

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

G. Schiebinger, E. Robeva, and B. Recht. Superresolution without separation. *arXiv preprint arXiv:1506.03144*, 2015.

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

K. Sridharan. A Gentle Introduction to Concentration Inequalities. Technical report, 2002.

G. Tang. Resolution limits for atomic decompositions via markov-bernstein type inequalities. In *Sampling Theory and Applications (SampTA), 2015 International Conference on*, pages 548–552. IEEE, 2015.

G. Tang, B. N. Bhaskar, P. Shah, and B. Recht. Compressed sensing off the grid. *IEEE transactions on information theory*, 59(11):7465–7490, 2013.

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

J. A. Tropp. An introduction to matrix concentration inequalities. (December), 2015.

S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

A Notations.

In this section, we recall and introduce some notation which will be used throughout the appendix.

Block norms. By default, $\|\cdot\|$ is the Euclidean norm for vector and spectral norm for matrices. For a vector $x = [x_1, \dots, x_s] \in \mathbb{C}^{sd}$ formed of s blocks $x_i \in \mathbb{C}^d$, $1 \leqslant i \leqslant s$, we define the block norm

$$\|x\|_{\text{block}} \stackrel{\text{def.}}{=} \sup_{1 \leqslant i \leqslant s} \|x_i\|_2$$

For a vector $q = [q_1, \dots, q_s, Q_1, \dots, Q_s] \in \mathbb{C}^{s(d+1)}$ decomposed such that $q_i \in \mathbb{C}$ and $Q_i \in \mathbb{C}^d$, we define

$$\|q\|_{*,\infty} \stackrel{\text{def.}}{=} \max_{i=1}^{s} \{|q_i|\,,\ \|Q_i\|\}.$$

Kernel The empirical kernel is defined as

$$\hat{K}(x, x') = \frac{1}{m} \sum_{k=1}^{m} \overline{\varphi_{\omega_k}(x)} \varphi_{\omega_k}(x')$$

and the limit kernel is $K(x, x) \stackrel{\text{def.}}{=} \mathbb{E}_\omega[\overline{\varphi_\omega(x)} \varphi_\omega(x')]$. The metric tensor associated to this kernel is

$$\mathbf{H}_x \stackrel{\text{def.}}{=} \mathbb{E}_\omega[\overline{\nabla \varphi_\omega(x)} \nabla \varphi_\omega(x)^\top]$$

Given an event E, we write $K_E(x, x') \stackrel{\text{def.}}{=} \mathbb{E}_\omega[\hat{K}(x, x')|E]$ to denote the conditional expectation on E.

Derivatives Given $f \in \mathscr{C}^\infty(\mathcal{X})$, by interpreting the r^{th} derivative as a multilinear map: $\nabla^r f : (\mathbb{C}^d)^r \to \mathbb{C}$, so given $Q \stackrel{\text{def.}}{=} \{q_\ell\}_{\ell=1}^r \in (\mathbb{C}^d)^r$,

$$\nabla^r f[Q] = \sum_{i_1, \cdots, i_r} \partial_{i_1} \cdots \partial_{i_r} f(x) q_{1,i_1} \cdots q_{r,i_r}.$$

and we define the r^{th} normalized derivative of f as

$$\mathrm{D}_r\left[f\right](x)[Q] \stackrel{\text{def.}}{=} \nabla^r f(x)[\{\mathbf{H}_x^{-\frac{1}{2}} q_i\}_{i=1}^r]$$

with norm $\|\mathrm{D}_r\left[f\right](x)\| \stackrel{\text{def.}}{=} \sup_{\forall \ell, \|q_\ell\| \leqslant 1} |\mathrm{D}_r\left[f\right](x)[Q]|$. We will sometimes make use the the multiarray interpretation: $\mathrm{D}_0\left[f\right] = f$, $\mathrm{D}_1\left[f\right](x) = \mathbf{H}_x^{-\frac{1}{2}} \nabla f(x) \in \mathbb{C}^d$, $\mathrm{D}_2\left[f\right](x) = \mathbf{H}_x^{-\frac{1}{2}} \nabla^2 f(x) \mathbf{H}_x^{-\frac{1}{2}} \in \mathbb{C}^{d \times d}$.

For a bivariate function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$, $\partial_{1,i}$ (resp. $\partial_{2,i}$) designates the derivative with respect to the i^{th} coordinate of the first variable (resp. second variable), and similarly ∇_i and ∇_i^2 denote the gradient and Hessian on the i^{th} coordinate respectively.

For $i, j \in \{0, 1, 2\}$, let $K^{(ij)}(x, x')$ be a "bi"-multilinear map, defined for $Q \in (\mathbb{C}^d)^i$ and $V \in (\mathbb{C}^d)^j$ as

$$[Q] K^{(ij)}(x, x')[V] \stackrel{\text{def.}}{=} \mathbb{E}[\overline{\mathrm{D}_i\left[\varphi_\omega\right](x)[Q]} \mathrm{D}_j\left[\varphi_\omega\right](x')[V]]$$

and $\left\|K^{(ij)}(x, x')\right\| \stackrel{\text{def.}}{=} \sup_{Q,V} \left\|[Q] K^{(ij)}(x, x')[V]\right\|$ where the supremum is defined over all $Q \stackrel{\text{def.}}{=} \{q_\ell\}_{\ell=1}^i$, $V \stackrel{\text{def.}}{=} \{v_\ell\}_{\ell=1}^j$ with $\|q_\ell\| \leqslant 1$, $\|v_\ell\| \leqslant 1$.

When $i + j \leqslant 2$, an equivalent definition is $K^{(ij)}(x, x') = \mathbb{E}[\overline{\mathrm{D}_i\left[\varphi_\omega\right](x)} \mathrm{D}_j\left[\varphi_\omega\right](x')^\top]$, and we note that $K^{(00)} = K$, and we have normalized so that $\mathrm{Re}\left(K^{(11)}(x, x)\right) = -\mathrm{Re}\left(K^{(02)}(x, x)\right)$. Finally, we will make use of the still equivalent definition: $[q] K^{(12)}(x, x') = \mathbb{E}[\overline{q^\top \mathrm{D}_1\left[\varphi_\omega\right](x)} \mathrm{D}_2\left[\varphi_\omega\right](x')^\top] \in \mathbb{C}^{d \times d}$.

Kernel constants For for $i, j \in \{(0, 0), (0, 1)\}$, define $B_{ij} \stackrel{\text{def.}}{=} \sup_{x, x' \in \mathcal{X}} \left|K^{(ij)}(x, x')\right|$, for $(i, j) \in \{(0, 2), (1, 2)\}$,

$$B_{ij} \stackrel{\text{def.}}{=} \sup \left\{ \left\|K^{(ij)}(x, x')\right\| \ ;\ d_{\mathbf{H}}(x, x') \leqslant r_{\text{near}} \text{ or } d_{\mathbf{H}}(x, x') > \Delta/2 \right\}.$$

and define for $i = 1, 2$

$$B_{ii} \stackrel{\text{def.}}{=} \sup_{x \in \mathcal{X}} \left\| K^{(ii)}(x, x) \right\|.$$

For convenience, we define

$$B_i \stackrel{\text{def.}}{=} B_{0i} + B_{1i} + 1, \quad B \stackrel{\text{def.}}{=} \sum_{\substack{i,j \in \{0,1,2\} \\ i+j \leqslant 3}} B_{ij} + 1. \tag{A.1}$$

Matrices and vectors We will make use of the following vectors and matrices throughout: Given $X \stackrel{\text{def.}}{=} \{x_j\}_{j=1}^s \in \mathcal{X}^s$ and $a \in \mathbb{C}^s$ which are always clear from context, define the vector $\gamma_X(\omega) \in \mathbb{C}^{s(d+1)}$ as

$$\gamma_X(\omega) \stackrel{\text{def.}}{=} \left(\left(\overline{\varphi_\omega(x_i)} \right)_{i=1}^s, \left(\overline{\mathrm{D}_1 \left[\varphi_\omega \right](x_i)}^\top \right)_{i=1}^s \right)^\top, \tag{A.2}$$

and

$$\begin{aligned} \Upsilon_X &\stackrel{\text{def.}}{=} \mathbb{E}_\omega[\gamma(\omega)\gamma(\omega)^*] \in \mathbb{C}^{s(d+1) \times s(d+1)} \\ \mathbf{f}_X(x) &\stackrel{\text{def.}}{=} \mathbb{E}_\omega[\gamma(\omega)\varphi_\omega(x)] \in \mathbb{C}^{s(d+1)} \\ \alpha &\stackrel{\text{def.}}{=} \Upsilon_X^{-1} \mathbf{u}_s, \qquad \mathbf{u}_s = \begin{pmatrix} \operatorname{sign}(a) \\ 0_{sd} \end{pmatrix}. \end{aligned}$$

Note that the diagonal of Υ has only 1's. For $\omega_1, \ldots, \omega_m$, we denote their empirical versions as:

$$\begin{aligned} \hat{\Upsilon}_X &\stackrel{\text{def.}}{=} \frac{1}{m} \sum_{k=1}^m \gamma(\omega_k)\gamma(\omega_k)^*, \\ \hat{\mathbf{f}}_X(x) &\stackrel{\text{def.}}{=} \frac{1}{m} \sum_{k=1}^m \gamma(\omega_k)\varphi_{\omega_k}(x), \quad \hat{\alpha} \stackrel{\text{def.}}{=} \hat{\Upsilon}_X^{-1} \mathbf{u}_s. \end{aligned}$$

which will serve us to construct our certificate, using the properties of their respective limit version.

We remark that $\mathbf{G}_X^{-1/2} \Gamma_X^* \Gamma_X \mathbf{G}_X^{-1/2} = \hat{\Upsilon}_X$, where Γ_X is defined in the main paper and

$$\mathbf{G}_X = \begin{pmatrix} \mathrm{Id}_s & & & 0 \\ & \mathbf{H}_{x_1} & & \\ & & \ddots & \\ 0 & & & \mathbf{H}_{x_s} \end{pmatrix}$$

The vanishing derivative pre-certificate $\hat{\eta}_X$ is $\hat{\alpha}^\top \hat{\mathbf{f}}_X(\cdot)$ and the limit pre-certificate is $\eta_X \stackrel{\text{def.}}{=} \alpha^\top \mathbf{f}_X(\cdot)$. When the set of points X is clear from context, we will drop the subscript X and write instead γ, Υ, $\mathbf{f}$, η, and so on.

Metric induced distances Given $X = (x_j)_{j=1}^s \in \mathcal{X}^s$ and $X' = (x'_j)_{j=1}^s \in \mathcal{X}^s$, denote $d_{\mathbf{H}}(X, X') \stackrel{\text{def.}}{=} \sqrt{\sum_j d_{\mathbf{H}}(x_j, x'_j)^2}$. Observe also that $\mathbf{G}_X$ is positive definite for all X and induces a metric on $\mathbb{R}^s \times \mathcal{X}^s$ so that given $a, a' \in \mathbb{R}^s$ and $X, X' \in \mathcal{X}^s$,

$$d_G((a, X), (a', X')) = \sqrt{\|a - a'\|_2^2 + d_{\mathbf{H}}(X, X')^2}.$$

Stochastic gradient bounds For $r \in \mathbb{N}$,

$$L_r(\omega) = \sup_{x \in \mathcal{X}} \left\| \mathrm{D}_r \left[\varphi_\omega \right](x) \right\|,$$

and $L_{ij}(\omega) \stackrel{\text{def.}}{=} \sqrt{L_i(\omega)^2 + L_j(\omega)^2}$. For $i = 0, 1, 2, 3$, let F_i be such that

$$\mathbb{P}_\omega \left(L_j(\omega) > t \right) \leqslant F_i(t),$$

Throughout, for $(\bar{L}_j)_{j=0}^3 \in \mathbb{R}_+^4$, the event $\bar{E}$ is defined as

$$\bar{E} \stackrel{\text{def.}}{=} \bigcap_{k=1}^{m} E_{\omega_k} \quad \text{where} \quad E_\omega \stackrel{\text{def.}}{=} \{L_j(\omega) \leqslant \bar{L}_j,\ \forall j = 0,1,2,3\}. \tag{A.3}$$

B Proof of Theorem 2

In this section, we consider the (limit) vanishing derivative pre-certificate

$$\eta(x) = \mathbf{u}^\top \Upsilon_X^{-1} \mathbf{f}_X(x).$$

Note that

$$\mathrm{D}_2\left[\eta\right](x) = \sum_{i=1}^{s} \alpha_{1,i} K^{(02)}(x_i, x) + [\alpha_{2,i}] K^{(12)}(x_i, x)$$

where we have decomposed $\alpha = [\alpha_{1,1}, \ldots, \alpha_{1,s}, \alpha_{2,1}, \ldots, \alpha_{2,s}] \in \mathbb{C}^{s(d+1)}$ where $\alpha_{2,i} \in \mathbb{C}^d$.

We aim to prove that η is nondegenerate if K is an admissible kernel. Our first lemma shows that nondegeneracy of η within each small neighbourhood of x_i can be established by controlling the real and imaginary parts of $\mathrm{D}_2\left[\eta\right]$ in each small region:

Lemma B.1. *Let $\varepsilon > 0$. Let $a_0 \neq 0$, $x_0 \in \mathcal{X}$ and let $\sigma \in \mathbb{C}$ be such that $|\sigma| = 1$. Suppose that $\eta \in \mathscr{C}^2(\mathcal{X}; \mathbb{C})$ is such that $\eta(x_0) = \sigma$, $\nabla\eta(x_0) = 0$ and $\operatorname{Re}\left(\overline{\sigma}\mathrm{D}_2\left[\eta\right](x_0)\right) \prec -\varepsilon\mathrm{Id}$. Then, $\nabla^2 \left|\eta\right|^2(x_0) \prec -2\varepsilon\mathrm{Id}$. If in addition, we have $c, r > 0$ with $\varepsilon r < 1$ and $c^2 \leqslant (1-\varepsilon r^2)/(\varepsilon r^2)$ such that for all x such that $d_{\mathbf{H}}(x, x_0) \leqslant r$,*

$$\operatorname{Re}\left(\overline{\sigma}\mathrm{D}_2\left[\eta\right](x)\right) \prec -\varepsilon\mathrm{Id} \quad \text{and} \quad \left\|\operatorname{Im}\left(\overline{\sigma}\mathrm{D}_2\left[\eta\right](x)\right)\right\| \leqslant c\varepsilon,$$

then, $|\eta(x)|^2 \leqslant 1 - \varepsilon^2 d_{\mathbf{H}}(x, x_0)^2$ for all x such that $d_{\mathbf{H}}(x, x_0) \leqslant r$.

Proof. The first claim follows immediately from the computation: by writing $\eta = \eta_r(x) + \mathrm{i}\eta_i(x)$ where η_i and η_r are real valued functions,

$$\frac{1}{2}\mathrm{D}_2\left[|\eta|^2\right] = \operatorname{Re}\left(\overline{\mathrm{D}_1\left[\eta\right]}\mathrm{D}_1\left[\eta\right]^\top + \mathrm{D}_2\left[\eta\right]\overline{\eta}\right),$$

and evaluation at x_0 gives the required result.

Let $\gamma : [0,1] \to \mathcal{X}$ be a piecewise smooth path such that $\gamma(0) = x_0$, $\gamma(1) = x$.

$$\begin{aligned}
\eta(x) &= \eta(x_0) + \int_0^1 (1-t)\langle \nabla^2\eta(\gamma(t))\gamma'(t),\, \gamma'(t)\rangle \mathrm{d}t \\
&= \eta(x_0) + \int_0^1 (1-t)\langle \mathrm{D}_2\left[\eta\right](\gamma(t))\mathbf{H}_{\gamma(t)}^{\frac{1}{2}}\gamma'(t),\, \mathbf{H}_{\gamma(t)}^{\frac{1}{2}}\gamma'(t)\rangle \mathrm{d}t.
\end{aligned}$$

So,

$$\operatorname{Re}\left(\overline{\operatorname{sign}(a_0)}\eta(x)\right) = 1 + \inf_\gamma \operatorname{Re}\left(\overline{\operatorname{sign}(a_0)}\int_0^1 (1-t)\langle \mathrm{D}_2\left[\eta\right](\gamma(t))\mathbf{H}_{\gamma(t)}^{\frac{1}{2}}\gamma'(t),\, \mathbf{H}_{\gamma(t)}^{\frac{1}{2}}\gamma'(t)\rangle \mathrm{d}t\right) \leqslant 1 - \varepsilon d_{\mathbf{H}}(x, x')^2$$

if we minimise over all paths from x to x_0. Similarly,

$$\left\|\operatorname{Im}\left(\overline{\operatorname{sign}(a_0)}\eta(x)\right)\right\| \leqslant c\varepsilon d_{\mathbf{H}}(x, x_0)^2$$

Therefore,

$$\begin{aligned}
|\eta(x)|^2 &\leqslant \left|1 - \varepsilon d_{\mathbf{H}}(x, x_0)^2\right|^2 + \left|c\varepsilon d_{\mathbf{H}}(x, x_0)^2\right|^2 \\
&\leqslant 1 - 2\varepsilon d_{\mathbf{H}}(x, x_0)^2 + \varepsilon^2 d_{\mathbf{H}}(x, x_0)^4 + c^2\varepsilon^2 d_{\mathbf{H}}(x, x_0)^4 \\
&= 1 - \varepsilon d_{\mathbf{H}}(x, x_0)^2 - \varepsilon d_{\mathbf{H}}(x, x_0)^2\left(1 - \varepsilon d_{\mathbf{H}}(x, x_0)^2\left(1 + c^2\right)\right) \leqslant 1 - \varepsilon d_{\mathbf{H}}(x, x_0)^2.
\end{aligned}$$

□

Proof of Theorem 2. In order to show that η is $(\varepsilon_0/2, \varepsilon_2/2)$-nondegenerate, it is enough to show that

$$\forall x \in \mathcal{X}^{\text{far}}, \quad |\eta(x)| \leqslant 1 - \varepsilon_0/2 \tag{B.1}$$

$$\forall x \in \mathcal{X}^{\text{near}}, \quad \text{Re}\left(\overline{\text{sign}(a_j)}\text{D}_2\left[\eta\right](x)\right) \prec -\frac{\varepsilon_2}{2}\text{Id} \quad \text{and} \quad \left\|\text{Im}\left(\overline{\text{sign}(a_j)}\text{D}_2\left[\eta\right](x)\right)\right\| \leqslant \frac{p}{4}\varepsilon_2 \tag{B.2}$$

where $p = \sqrt{\frac{1-\varepsilon_2 r_{\text{near}}^2/2}{\varepsilon_2 r_{\text{near}}^2/2}}$.

We first prove that the matrix Υ is invertible. To this end, we write

$$\Upsilon = \begin{pmatrix} \Upsilon_0 & \Upsilon_1^\top \\ \Upsilon_1 & \Upsilon_2 \end{pmatrix} \tag{B.3}$$

where $\Upsilon_0 \overset{\text{def.}}{=} (K(x_i, x_j))_{i,j=1}^s \in \mathbb{C}^{s\times s}$, $\Upsilon_1 \overset{\text{def.}}{=} (K^{(10)}(x_i, x_j))_{i,j=1}^s \in \mathbb{C}^{sd\times s}$, and $\Upsilon_2 \overset{\text{def.}}{=} (K^{(11)}(x_i, x_j))_{i,j=1}^s \in \mathbb{C}^{sd\times sd}$. By definition of $K^{(ij)}$, Υ (and also Υ_0 and Υ_2) has only 1's on its diagonal.

To prove the invertibility of Υ, we use the Schur complement of Υ, and in particular it suffices to prove that Υ_2 and the Schur complement $\Upsilon_S \overset{\text{def.}}{=} \Upsilon_0 - \Upsilon_1\Upsilon_2^{-1}\Upsilon_1^\top$ are both invertible. To show that Υ_2 is invertible, we define $A_{ij} = K^{(11)}(x_i, x_j)$. So Υ_2 has the form:

$$\Upsilon_2 = \begin{pmatrix} \text{Id} & A_{12} & \dots & A_{1s} \\ A_{21} & \text{Id} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ A_{s1} & \dots & \dots & \text{Id} \end{pmatrix}$$

and by Lemma G.6, we have

$$\|\text{Id} - \Upsilon_2\|_{\text{block}} \leqslant \max_i \sum_j \|A_{ij}\| \leqslant 1/4.$$

Since $\|\text{Id} - \Upsilon_2\|_{\text{block}} < 1$, Υ_2 is invertible, and we have $\left\|\Upsilon_2^{-1}\right\|_{\text{block}} \leqslant \frac{1}{1-\|I-\Upsilon_2\|_{\text{block}}} \leqslant \frac{4}{3}$. Next, again with Lemma G.6, we can bound

$$\begin{aligned}
\|I - \Upsilon_0\|_\infty &= \max_i \sum_{j\neq i} |K(x_i, x_j)| \leqslant \frac{\varepsilon_0}{16} \\
\|\Upsilon_1\|_{\infty\to\text{block}} &\leqslant \max_i \sum_j \left\|K^{(10)}(x_i, x_j)\right\| \leqslant h \quad \text{since } K^{(10)}(x, x) = 0 \\
\left\|\Upsilon_1^\top\right\|_{\text{block}\to\infty} &\leqslant \max_i \sum_j \left\|K^{(10)}(x_j, x_i)\right\| \leqslant h
\end{aligned}$$

Hence, we have

$$\|I - \Upsilon_S\|_\infty \leqslant \|I - \Upsilon_0\|_\infty + \left\|\Upsilon_1^\top\right\|_{\text{block}\to\infty} \left\|\Upsilon_2^{-1}\right\|_{\text{block}} \|\Upsilon_1\|_{\infty\to\text{block}} \leqslant \frac{\varepsilon_0}{16} + \frac{4}{3}h^2 \leqslant \frac{\varepsilon_0}{8} \tag{B.4}$$

since $h \leqslant \frac{\varepsilon_0}{32}$. Therefore the Schur complement of Υ is invertible and so is Υ.

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

We divide:

$$\alpha = \Upsilon^{-1}\mathbf{u}_s = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

where $\alpha_1 \in \mathbb{C}^s$ and $\alpha_2 \in \mathbb{C}^{sd}$, and we denote $\alpha_{2,i} \in \mathbb{C}^d$ blocks such that $\alpha_2 = [\alpha_{2,1}, \dots, \alpha_{2,s}]$.

The Schur's complement of Υ allows us to express α_1 and α_2 as

$$\alpha_1 = \Upsilon_S^{-1}\operatorname{sign}(a), \qquad \alpha_2 = -\Upsilon_2^{-1}\Upsilon_1\Upsilon_S^{-1}\operatorname{sign}(a) \tag{B.5}$$

and therefore we can bound

$$\|\alpha_1\|_\infty \leqslant \frac{1}{1-\varepsilon_0/8} \tag{B.6}$$

$$\|\alpha_2\|_{\text{block}} \leqslant \frac{8}{3}h \leqslant 4h \tag{B.7}$$

Moreover, we have

$$\|\alpha_1 - \operatorname{sign}(a)\|_\infty \leqslant \left\|I - \Upsilon_S^{-1}\right\|_\infty \leqslant \left\|\Upsilon_S^{-1}\right\|_\infty \|I - \Upsilon_S\|_\infty \leqslant \frac{1}{4} \tag{B.8}$$

Non-degeneracy. We can now prove that η is non-degenerate.

Let x be such that $d_{\mathbf{H}}(x_i, x) \leqslant r_{\text{near}}$. We need to prove that for all x such that $d_{\mathbf{H}}(x, x_i) \leqslant r$,

$$\operatorname{Re}\left(\overline{\operatorname{sign}(a_i)}\mathrm{D}_2\left[\eta\right](x)\right) \prec -\frac{\varepsilon_2}{2}\mathrm{Id} \quad \text{and} \quad \left\|\operatorname{Im}\left(\overline{\operatorname{sign}(a_i)}\mathrm{D}_2\left[\eta\right](x)\right)\right\| \leqslant \frac{\varepsilon_2}{2}\sqrt{\frac{2-\varepsilon r_{\text{near}}^2}{\varepsilon_2 r_{\text{near}}^2}}.$$

Then, since $r_{\text{near}} \leqslant \Delta/2$ and the x_i's are Δ-separated, for all $j \neq i$ we have $d_{\mathbf{H}}(x, x_j) \geqslant \Delta/2$. Then, we have

$$\begin{aligned}\overline{\operatorname{sign}(a_i)}\mathrm{D}_2\left[\eta\right](x) = \overline{\operatorname{sign}(a_i)}\Bigg[&\alpha_{1,i}K^{(02)}(x_i,x) + \sum_{j\neq i}\alpha_{1,j}K^{(02)}(x_j,x) \\ &+ [\alpha_{2,i}]K^{(12)}(x_i,x) + \sum_{j\neq i}[\alpha_{2,j}]K^{(12)}(x_j,x)\Bigg]\end{aligned}$$

$$\begin{aligned}\operatorname{Re}\left(\overline{\operatorname{sign}(a_i)}\mathrm{D}_2\left[\eta\right](x)\right) &\preccurlyeq (1 - \|\alpha_1 - \operatorname{sign}(a)\|_\infty)\operatorname{Re}\left(K^{(02)}(x_i,x)\right) + \|\alpha_1\|_\infty \sum_{j\neq i}\left\|K^{(02)}(x_j,x)\right\|\mathrm{Id} \\ &\quad + \left(\left\|K^{(12)}(x_i,x)\right\| + \sum_{j\neq i}\left\|K^{(12)}(x_j,x)\right\|\right)\|\alpha_2\|_{\text{block}}\,\mathrm{Id} \\ &\preccurlyeq \left(-\frac{3}{4}\varepsilon_2 + \frac{1}{1-\varepsilon_0/8}\frac{\varepsilon_2}{16} + 4h(B_{12}+1)\right)\mathrm{Id} \preccurlyeq \varepsilon_2\left(-\frac{3}{4}+\frac{1}{4}\right)\mathrm{Id} \preccurlyeq -\frac{\varepsilon_2}{2}\mathrm{Id}\,.\end{aligned}$$

Taking the imaginary part, we have

$$\begin{aligned}\left\|\operatorname{Im}\left(\overline{\operatorname{sign}(a_i)}\mathrm{D}_2\left[\eta\right](x)\right)\right\| \leqslant\ & (1 + \|\alpha_1 - \operatorname{sign}(a)\|)\left\|\operatorname{Im}\left(K^{(02)}(x_i,x)\right)\right\| + \|\alpha_1\|_\infty \sum_{j\neq i}\left\|K^{(02)}(x_j,x)\right\| \\ &+ \left(\left\|K^{(12)}(x_i,x)\right\| + \sum_{j\neq i}\left\|K^{(12)}(x_j,x)\right\|\right)\|\alpha_2\|_{\text{block}} \\ \leqslant \left(\frac{5c\varepsilon_2}{4} + \frac{1}{(1-\varepsilon_0/8)}h + 4h(B_{12}+1)\right) &\leqslant \frac{5c\varepsilon_2}{4} + h\left(4B_{12}+6\right) \leqslant \frac{\varepsilon_2}{2}\sqrt{\frac{2-\varepsilon r_{\text{near}}^2}{\varepsilon_2 r_{\text{near}}^2}}.\end{aligned}$$

So, by Lemma B.1, for each $i = 1, \ldots, s$, $|\eta(x)| \leqslant 1 - \varepsilon_2/2 d_{\mathbf{H}}(x, x_i)$ for all $x \in \mathcal{X}$ such that $d_{\mathbf{H}}(x, x_i) \leqslant r_{\text{near}}$.

Next, for any x such that $d_{\mathbf{H}}(x, x_i) \geqslant r_{\text{near}}$ for all x_i's, we can say that there exists (at most) one index i such

that $d_{\mathbf{H}}(x, x_i) \geqslant r_{\mathrm{near}}$ and for all $j \neq i$ we have $d_{\mathbf{H}}(x, x_j) \geqslant \Delta/2$. We have

$$\begin{aligned}
|\eta(x)| &= \Big| \alpha_{1,i} K(x_i, x) + \sum_{j \neq i} \alpha_{1,j} K(x_j, x) \\
&\qquad + K^{(10)}(x_i, x)^\top \alpha_{2,i} + \sum_{j \neq i} K^{(10)}(x_j, x)^\top \alpha_{2,j} \Big| \\
&\leqslant \|\alpha_1\|_\infty \left(|K(x_i, x)| + \sum_{j \neq i} |K(x_j, x)| \right) \\
&\qquad + \|\alpha_2\|_{\mathrm{block}} \left(\left\| K^{(10)}(x_i, x) \right\| + \sum_{j \neq i} \left\| K^{(10)}(x_j, x) \right\| \right) \\
&\leqslant \frac{1 - \varepsilon_0 + \varepsilon_0/16}{1 - \varepsilon_0/8} + 4h(B_{10} + 1) \leqslant 1 - \frac{\varepsilon_0}{2} \,.
\end{aligned}$$

□

Remark B.1. *Assuming that the derivatives of the kernel decay like a function $f(\|x - x'\|)$ when, there is always a separation $\Delta \propto f^{-1}(1/(Cs_{\max}))$ such that the kernel is admissible. Ex: when $f = x^{-p}$, we have $\Delta \propto s_{\max}^{1/p}$ (eg Cauchy). When $f = e^{-x^p}$, we have $\Delta \propto \log^{1/p}(s_{\max})$ (eg Gaussian).*

C Preliminaries

In this section, we present some preliminary results which will be used for proving our main results. We assume that K is admissible, and given a set of points $X \in \mathcal{X}^s$, let $\mathcal{X}_j^{\mathrm{near}} \stackrel{\text{def.}}{=} \{x \in \mathcal{X} \,;\, d_{\mathbf{H}}(x, x_j) \leqslant r_{\mathrm{near}}\}$, $\mathcal{X}^{\mathrm{near}} \stackrel{\text{def.}}{=} \bigcup_{j=1}^s \mathcal{X}_j^{\mathrm{near}}$ and $\mathcal{X}^{\mathrm{far}} \stackrel{\text{def.}}{=} \mathcal{X} \setminus \mathcal{X}^{\mathrm{near}}$.

C.1 On the determistic kernel

For an admissible kernel, we have the following additional bounds that will be handy.

Lemma C.1. *Assume K is an admissible kernel, let $X \in \mathcal{X}^s$ be Δ-separated points. Then we have the following:*

(i) We have seen that Υ is invertible. Additionally it satisfies

$$\|\mathrm{Id} - \Upsilon\| \leqslant \frac{1}{2} \quad \textit{and} \quad \|\mathrm{Id} - \Upsilon\|_{*,\infty} \leqslant \frac{1}{2}. \tag{C.1}$$

(ii) For any vector $q \in \mathbb{C}^{s(d+1)}$ and any $x \in \mathcal{X}^{\mathrm{far}}$, we have

$$\|\mathbf{f}(x)\| \leqslant B_0 \quad \textit{and} \quad \left|q^\top \mathbf{f}(x)\right| \leqslant B_0 \|q\|_{*,\infty} \tag{C.2}$$

(iii) For any vector $q \in \mathbb{C}^{s(d+1)}$ and any $x \in \mathcal{X}^{\mathrm{near}}$ we have the bound:

$$\left\|\mathrm{D}_2 \left[q^\top \mathbf{f}(.)\right](x)\right\| \leqslant \|q\| B_2 \quad \textit{and} \quad \left\|\mathrm{D}_2 \left[q^\top \mathbf{f}(.)\right](x)\right\| \leqslant \|q\|_{*,\infty} B_2 \tag{C.3}$$

Proof. We bound the spectral norm of $\mathrm{Id} - \Upsilon$. Define $y \in \mathbb{C}^{s(d+1)}$ decomposed as $y = [y_1, \ldots, y_s, Y_1, \ldots, Y_s]$

where $Y_i \in \mathbb{R}^d$, such that $\|y\| \leqslant 1$. We have

$$\begin{aligned}
\|(\mathrm{Id} - \Upsilon)y\|^2 = & \sum_{i=1}^{s} \left| \sum_{j \neq i} K(x_i, x_j) y_j + \sum_{j=1}^{s} K^{(10)}(x_i, x_j)^\top Y_j \right|^2 \\
& + \left\| \sum_j y_j K^{(10)}(x_i, x_j) + \sum_{j \neq i} K^{(11)}(x_i, x_j) Y_j \right\|^2 \\
\leqslant & \sum_{i=1}^{s} \left(\sum_{j \neq i} |K(x_i, x_j)| \, |y_j| + \sum_{j=1}^{s} \left\| K^{(10)}(x_i, x_j) \right\| \|Y_j\| \right)^2 \\
& + \left(\sum_j |y_j| \left\| K^{(10)}(x_i, x_j) \right\| + \sum_{j \neq i} \left\| K^{(11)}(x_i, x_j) \right\| \|Y_j\| \right)^2 \\
\leqslant & \max_{d_{\mathbf{H}}(x,x') \geqslant \Delta} \left(|K(x, x')| \, , \left\| K^{(10)}(x, x') \right\| , \left\| K^{(11)}(x, x') \right\| \right)^2 \sum_i 2 \left(\sum_j |y_j| + \|Y_j\| \right)^2 \\
\leqslant & \, 4s^2 \max_{d_{\mathbf{H}}(x,x') \geqslant \Delta} \left(|K(x, x')| \, , \left\| K^{(10)}(x, x') \right\| , \left\| K^{(11)}(x, x') \right\| \right)^2
\end{aligned}$$

by Cauchy-Schwartz inequality and since $K^{(10)}(x, x) = 0$ for all $x \in \mathcal{X}$. Since by hypothesis we have

$$\max_{d_{\mathbf{H}}(x,x') \geqslant \Delta} \left(|K(x, x')| \, , \left\| K^{(10)}(x, x') \right\| , \left\| K^{(11)}(x, x') \right\| \right) \leqslant \frac{1}{4 s_{\max}} \, ,$$

we obtain

$$\|\mathrm{Id} - \Upsilon\| \leqslant \frac{1}{2} \tag{C.4}$$

and we deduce (i). A near identical argument also yields $\|\Upsilon - \mathrm{Id}\|_{*,\infty} \leqslant \frac{1}{4}$.

For (ii), let $x \in \mathcal{X}^{\mathrm{far}}$, then we have

$$\begin{aligned}
\|\mathbf{f}(x)\| & \leqslant \left(\sum_{i=1}^{s} |K(x_i, x)|^2 + \left\| K^{(10)}(x_i, x) \right\|^2 \right)^{\frac{1}{2}} \\
& \leqslant \left(B_{00}^2 + \frac{(s-1)\varepsilon_0^2}{(16 s_{\max})^2} + B_{10}^2 + \frac{(s-1)}{s_{\max}^2} \right)^{\frac{1}{2}} \leqslant B_0
\end{aligned}$$

for which, similar to the proof above, we have used the fact that x is $\Delta/2$-separated from at least $s - 1$ points x_i. Similarly, for any vector $q = [q_1, \ldots, q_s, Q_1, \ldots, Q_s] \in \mathbb{C}^{s(d+1)}$ and any $x \in \mathcal{X}^{\mathrm{far}}$, we have

$$\begin{aligned}
\left\| q^\top \mathbf{f}(x) \right\| & \leqslant \sum_{i=1}^{s} |q_i| \, |K(x_i, x)| + \|Q_i\| \left\| K^{(10)}(x_i, x) \right\| \\
& \leqslant \|q\|_{*,\infty} \left(B_{00} + \frac{(s-1)\varepsilon_0}{32 s_{\max})} + B_{10} + \frac{(s-1)\varepsilon_0}{32 s_{\max}} \right) \leqslant B_0 \, \|q\|_{*,\infty} \, .
\end{aligned}$$

For any $x \in \mathcal{X}^{\text{near}}$ we have the bound:

$$\begin{aligned}\left\|\mathrm{D}_2\left[q^\top \mathbf{f}\right](x)\right\| &= \left\|\sum_{i=1}^{s} q_i K^{(02)}(x_i, x) + [Q_i] K^{(12)}(x_i, x)\right\| \\ &\leqslant \|q\| \left(\sum_{i=1}^{s} \left\|K^{(02)}(x_i, x)\right\|^2 + \left\|K^{(12)}(x_i, x)\right\|^2\right)^{\frac{1}{2}} \\ &\leqslant \|q\| B_2\end{aligned}$$

and

$$\begin{aligned}\left\|\mathrm{D}_2\left[q^\top \mathbf{f}\right](x)\right\| &= \left\|\sum_{i=1}^{s} q_i K^{(02)}(x_i, x) + [Q_i] K^{(12)}(x_i, x)\right\| \\ &\leqslant \|q\|_{*,\infty} \left(\sum_{i=1}^{s} \left\|K^{(02)}(x_i, x)\right\| + \left\|K^{(12)}(x_i, x)\right\|\right) \\ &\leqslant \|q\|_{*,\infty} B_2\end{aligned}$$

□

C.2 Lipschitz bounds

Lemma C.2 (Local Lipschitz constant of φ_ω and higher order derivatives)**.** *Suppose that $\|\mathrm{D}_j[\varphi_\omega](x)\| \leqslant \bar{L}_j$ for all $x \in \mathcal{X}$. For all x, x' with $d_{\mathbf{H}}(x, x') \leqslant r_{\text{near}}$, we have*

(i) $|\varphi_\omega(x) - \varphi_\omega(x')| \leqslant \mathcal{L}_0 d_{\mathbf{H}}(x, x')$,

(ii) $\|\mathrm{D}_1[\varphi_\omega](x) - \mathrm{D}_1[\varphi_\omega](x')\| \leqslant \mathcal{L}_1 d_{\mathbf{H}}(x, x')$,

(iii) $\|\mathrm{D}_2[\varphi_\omega](x) - \mathrm{D}_2[\varphi_\omega](x')\| \leqslant \mathcal{L}_2 d_{\mathbf{H}}(x, x')$,

where $\mathcal{L}_0 \overset{\text{def.}}{=} \bar{L}_1$, $\mathcal{L}_1 \overset{\text{def.}}{=} \bar{L}_1 C_{\mathbf{H}} + \bar{L}_2(1 + C_{\mathbf{H}} r_{\text{near}})$ and $\mathcal{L}_2 \overset{\text{def.}}{=} \bar{L}_2\left(C_{\mathbf{H}} + C_{\mathbf{H}}^2 r_{\text{near}} + 1\right) + \bar{L}_3(1 + C_{\mathbf{H}} r_{\text{near}})^2$. As a consequence, for all $X = (x_j)$ and $X' = (x'_j)$ such that $d_{\mathbf{H}}(x_j, x'_j) \leqslant r_{\text{near}}$, we have

$$\sup_{\|q\|=1} \left\|\mathrm{D}_r\left[q^\top(\hat{\mathbf{f}}_X - \hat{\mathbf{f}}_{X'})\right](y)\right\| \leqslant \bar{L}_r \sqrt{\mathcal{L}_0^2 + \mathcal{L}_1^2}\, d_{\mathbf{H}}(X, X').$$

Proof. Let $x, x' \in \mathcal{X}$ with $d_{\mathbf{H}}(x, x') \leqslant r_{\text{near}}$. Recall that $\left\|\mathbf{H}_{x'}^{\frac{1}{2}} \mathbf{H}_x^{-\frac{1}{2}} - \mathrm{Id}\right\| \leqslant C_{\mathbf{H}} d_{\mathbf{H}}(x, x')$, and so, $\left\|\mathbf{H}_{x'}^{\frac{1}{2}} \mathbf{H}_x^{-\frac{1}{2}}\right\| \leqslant 1 + C_{\mathbf{H}} r_{\text{near}}$.

Let $p : [0, 1] \to \mathcal{X}$ be a piecewise smooth path such that $p(0) = x'$, $p(1) = x$. Then, by Taylor's theorem,

$$\varphi_\omega(x) - \varphi_\omega(x') = \int_{t=0}^{1} \langle \mathbf{H}_{p(t)}^{-\frac{1}{2}} \nabla \varphi_\omega(p(t)),\, \mathbf{H}_{p(t)}^{\frac{1}{2}} p'(t) \rangle \mathrm{d}t \leqslant \bar{L}_1 \int_0^1 \left\|\mathbf{H}_{p(t)}^{\frac{1}{2}} p'(t)\right\| \mathrm{d}t \tag{C.5}$$

so taking the minimum over all paths p yields $|\varphi_\omega(x) - \varphi_\omega(x')| \leqslant \bar{L}_1 d_{\mathbf{H}}(x, x')$.

Given $q \in \mathbb{R}^d$, by Taylor's theorem,

$$\begin{aligned}&\mathrm{D}_1[\varphi_\omega](x)[q] = \nabla \varphi(x)[\mathbf{H}_x^{-\frac{1}{2}} q] = \nabla \varphi(x')[\mathbf{H}_x^{-\frac{1}{2}} q] + \int \nabla^2 \varphi_\omega(p(t))[\mathbf{H}_x^{-\frac{1}{2}} q, p'(t)] \mathrm{d}t \\ &= \mathrm{D}_1[\varphi_\omega](x')[q] + \mathrm{D}_1[\varphi_\omega](x')[(\mathbf{H}_{x'}^{\frac{1}{2}} \mathbf{H}_x^{-\frac{1}{2}} - \mathrm{Id}) q] + \int \mathrm{D}_2[\varphi_\omega](p(t))[\mathbf{H}_{p(t)}^{\frac{1}{2}} \mathbf{H}_x^{-\frac{1}{2}} q, \mathbf{H}_{p(t)}^{\frac{1}{2}} p'(t)] \mathrm{d}t\end{aligned} \tag{C.6}$$

Therefore,

$$\|\mathrm{D}_1[\varphi_\omega](x) - \mathrm{D}_1[\varphi_\omega](x')\| \leqslant \bar{L}_1 C_{\mathbf{H}} d_{\mathbf{H}}(x, x') + \bar{L}_2(1 + C_{\mathbf{H}} r_{\text{near}}) d_{\mathbf{H}}(x, x').$$

Finally, for all $q_1, q_2 \in \mathbb{R}^d$, by Taylor's theorem

$$\begin{aligned}
&\mathrm{D}_2\left[\varphi_\omega\right](x)[q_1,q_2] - \mathrm{D}_2\left[\varphi_\omega\right](x')[q_1,q_2] \\
&= \nabla^2\varphi_\omega(x)[\mathbf{H}_x^{-\frac{1}{2}}q_1, \mathbf{H}_x^{-\frac{1}{2}}q_2] - \nabla^2\varphi_\omega(x')[\mathbf{H}_{x'}^{-\frac{1}{2}}q_1, \mathbf{H}_{x'}^{-\frac{1}{2}}q_2] \\
&= \mathrm{D}_2\left[\varphi_\omega\right](x')[\mathbf{H}_{x'}^{\frac{1}{2}}\mathbf{H}_x^{-\frac{1}{2}}q_1, (\mathbf{H}_{x'}^{\frac{1}{2}}\mathbf{H}_x^{-\frac{1}{2}} - \mathrm{Id})q_2] + \mathrm{D}_2\left[\varphi_\omega\right](x')[(\mathbf{H}_{x'}^{\frac{1}{2}}\mathbf{H}_x^{-\frac{1}{2}} - \mathrm{Id})q_1, q_2] \\
&\quad + \int \mathrm{D}_3\left[\varphi_\omega\right](p(t))[\mathbf{H}_{p(t)}^{\frac{1}{2}}\mathbf{H}_x^{-\frac{1}{2}}q_1, \mathbf{H}_{p(t)}^{\frac{1}{2}}\mathbf{H}_x^{-\frac{1}{2}}q_2, \mathbf{H}_{p(t)}^{\frac{1}{2}}p'(t)]\mathrm{d}t.
\end{aligned} \tag{C.7}$$

Therefore,

$$\left\|\mathrm{D}_2\left[\varphi_\omega\right](x) - \mathrm{D}_2\left[\varphi_\omega\right](x')\right\| \leqslant \left(\bar{L}_2\left((1 + C_{\mathbf{H}} r_{\mathrm{near}})C_{\mathbf{H}} + 1\right) + \bar{L}_3(1 + C_{\mathbf{H}} r_{\mathrm{near}})^2\right) d_{\mathbf{H}}(x, x').$$

By applying these Lipschitz bounds, we obtain

$$\begin{aligned}
&\sup_{\|q\|=1} \left\|\mathrm{D}_r\left[q^\top(\hat{\mathbf{f}}_X - \hat{\mathbf{f}}_{X'})\right](y)\right\|^2 \\
&\quad \leqslant \sum_{j=1}^s \left\|\hat{K}^{(0r)}(x_j, y) - \hat{K}^{(0r)}(x'_j, y)\right\|^2 + \sum_{j=1}^s \left\|\hat{K}^{(1r)}(x_j, y) - \hat{K}^{(1r)}(x'_j, y)\right\|^2 \\
&\quad \leqslant \sum_{j=1}^s \mathcal{L}_0^2 \bar{L}_r^2 d_{\mathbf{H}}(x_j, x'_j)^2 + \sum_{j=1}^s \mathcal{L}_1^2 \bar{L}_r^2 d_{\mathbf{H}}(x_j, x'_j)^2 \\
&\quad = \left(\mathcal{L}_0^2 + \mathcal{L}_1^2\right) \bar{L}_r^2 d_{\mathbf{H}}(X, X')^2
\end{aligned}$$

□

Lemma C.3 (Local Lipschitz constant of $\hat{K}^{(ij)}$)**.** *Let $x_1, x_0 \in \mathcal{X}$. Let $i, j \in \{0,1,2\}$ with $i + j \leqslant 3$. Define*

$$A_{ij} = \sup_x \left\|\hat{K}^{(ij)}(x, x_0)\right\|$$

where x ranges over $d_{\mathbf{H}}(x, x_1) \leqslant r_{\mathrm{near}}$. Then, for all x such that $d_{\mathbf{H}}(x, x_1) \leqslant r_{\mathrm{near}}$,

$$\begin{aligned}
\left\|\hat{K}^{(0j)}(x, x_0) - \hat{K}^{(0j)}(x_1, x_0)\right\| &\leqslant A_{1j} d_{\mathbf{H}}(x, x_1) \\
\left\|\hat{K}^{(1j)}(x, x_0) - \hat{K}^{(1j)}(x_1, x_0)\right\| &\leqslant \left(C_{\mathbf{H}} A_{1j} + (1 + C_{\mathbf{H}} r_{\mathrm{near}}) A_{2j}\right) d_{\mathbf{H}}(x, x_1)
\end{aligned}$$

The same results hold if we replace $\hat{K}$ by K.

Proof. The Lipschitz bounds on $\hat{K}^{ij}$ follow by combining

$$\begin{aligned}
&[q_1, \ldots, q_i](\hat{K}^{(ij)}(x, x_0) - \hat{K}^{(ij)}(x_1, x_0))[v_1, \ldots, v_j] \\
&= \hat{\mathbb{E}}\mathrm{Re}\left(\overline{(\mathrm{D}_i\left[\varphi_\omega\right](x) - \mathrm{D}_i\left[\varphi_\omega\right](x_1))[q_1, \ldots, q_i]}\mathrm{D}_j\left[\varphi_j\right](x_0)[v_1, \ldots, v_j]\right)
\end{aligned}$$

where $\hat{\mathbb{E}}$ indicates either empirical expectation or true expectation with (C.5), (C.6) and (C.7).

□

C.3 Probability bounds

In the proof of our main results, we will often assume that event $\bar{E}$ (see (A.3)) holds since our assumptions in Section 2.3 imply that $\mathbb{P}(\bar{E}^c) \leqslant \rho/m$. The following lemma shows that our assumptions also imply that $\mathbb{E}_\omega[L_i(\omega)^2 1_{E_\omega^c}] \leqslant \frac{\varepsilon}{m}$. and this is a condition which our proofs will often rely upon.

Lemma C.4. *The following holds.* $\mathbb{P}(E_\omega^c) \leqslant \sum_i F_i(\bar{L}_i)$ *and*

$$\mathbb{E}_\omega[L_j(\omega)^2 1_{E_\omega^c}] \leqslant 2\int_{\bar{L}_j}^\infty tF_j(t)\mathrm{d}t + \bar{L}_j^2 \sum_i F_i(\bar{L}_i)$$

Proof. Let $E_{\omega,j}$ be the event that $L_r(\omega) \leqslant \bar{L}_r$, so $E_\omega = \cap_{j=0}^3 E_{\omega,j}$. By the union bound, $\mathbb{P}(E_\omega^c) \leqslant \sum_j \mathbb{P}(E_{\omega,j}^c) \leqslant \sum_i F_i(\bar{L}_i)$.

For the second claim, observe that $E_\omega^c = \cup_i E_{\omega,i}^c$ so that $\mathbb{E}[L_j(\omega)^2 1_{E_\omega^c}] \leqslant \sum_i \mathbb{E}[L_j(\omega)^2 1_{E_{\omega,i}^c}]$ and we have

$$\begin{aligned}
\mathbb{E}[L_j(\omega)^2 1_{E_{\omega,i}^c}] &= \int_0^\infty \mathbb{P}(L_j(\omega)^2 1_{E_{\omega,i}^c} \geqslant t)\mathrm{d}t \\
&= \int_0^\infty \mathbb{P}\left((L_j(\omega)^2 \geqslant t) \cap (L_i(\omega) \geqslant \bar{L}_i)\right)\mathrm{d}t \\
&\leqslant \bar{L}_j^2 F_i(\bar{L}_i) + \int_{\bar{L}_j^2}^\infty F_j(\sqrt{t})\mathrm{d}t = \bar{L}_j^2 F_i(\bar{L}_i) + 2\int_{\bar{L}_j}^\infty tF_j(t)\mathrm{d}t
\end{aligned}$$

where we have bounded $\mathbb{P}\left((L_j(\omega)^2 \geqslant t) \cap (L_i(\omega) \geqslant \bar{L}_i)\right)$ by respectively $\mathbb{P}(L_i(\omega) \geqslant \bar{L}_i) \leqslant F_i(\bar{L}_i)$ in the first term and by $\mathbb{P}(L_j(\omega)^2 \geqslant t) \leqslant F_j(\sqrt{t})$ in the second term. □

C.3.1 Concentration inequalities

The following result is an adaption of the Matrix Bernstein inequality for dealing with conditional probabilities.

Lemma C.5 (Adapted unbounded Matrix Bernstein)**.** *Let $A_j \in \mathbb{R}^{d_1 \times d_2}$ be a family of iid matrices for $j = 1, \dots, m$. Let $Z = \frac{1}{m}\sum_{j=1}^m A_j$ and let $\bar{Z} = \mathbb{E}[Z]$. Let $t \in (0, 4\|\mathbb{E}[A_1]\|]$. Let events E_j be independent events such that $E_j \subseteq \{\|A_j\| \leqslant L\}$ and let $E = \cap_j E_j$. Suppose that we have*

$$\mathbb{P}(E_j^c) \leqslant \frac{t}{t + 4\|\mathbb{E}[A_1]\|} \quad \text{and} \quad \mathbb{E}[\|A_j\| 1_{E_j^c}] \leqslant \frac{t}{4}$$

Then a first consequence is that we have $\mathbb{E}_E[Z] = \mathbb{E}_{E_j}[A_j]$ for all j and $\|\mathbb{E}[Z] - \mathbb{E}_E[Z]\| \leqslant \frac{t}{2}$.

Finally, assuming that

$$\sigma^2 \overset{\text{def.}}{=} \max_j\{\left\|\mathbb{E}_{E_j}[A_jA_j^*]\right\|, \left\|\mathbb{E}_{E_j}[A_j^*A_j]\right\|\} < \infty$$

we have

$$\mathbb{P}_E\left(\|Z - \mathbb{E}[Z]\| \geqslant t\right) \leqslant (d_1 + d_2)\exp\left(-\frac{mt^2/4}{\sigma^2 + Lt/3}\right).$$

Proof. We first bound $\|\mathbb{E}[Z] - \mathbb{E}_E[Z]\|$. First observe that $\mathbb{E}[Z] = \mathbb{E}_{E_1}[A_1]$ and $\mathbb{E}_E Z = \mathbb{E}_{E_1}[A_1]$ since A_j are iid. Moreover,

$$\mathbb{E}[A_1] = \mathbb{E}[A_1 1_{E_1}] + \mathbb{E}[A_1 1_{E_1^c}] = \mathbb{E}[A_1|E_1]\mathbb{P}(E_1) + \mathbb{E}[A_1 1_{E_1^c}].$$

Hence,

$$\begin{aligned}
\|\mathbb{E}[A_1] - \mathbb{E}_{E_1}[A_1]\| &= \left\|(P(E_1) - 1)\mathbb{E}_{E_1}[A_1] + \mathbb{E}[A_1 1_{E_1^c}]\right\| \\
&\leqslant \mathbb{P}(E_1^c)\|\mathbb{E}[A_1]\| + P(E_1^c)\|\mathbb{E}[A_1] - \mathbb{E}_{E_1}[A_1]\| + \mathbb{E}[\|A_1\| 1_{E_1^c}].
\end{aligned}$$

Therefore,

$$\|\mathbb{E}[A_1] - \mathbb{E}_{E_1}[A_1]\| \leqslant \frac{P(E_1^c)\|\mathbb{E}[A_1]\| + \mathbb{E}[\|A_1\| 1_{E_1^c}]}{1 - \mathbb{P}(E_1^c)} \leqslant \frac{t}{2}$$

For the second statement,

$$\begin{aligned}
\mathbb{P}_E(\|Z - \mathbb{E}[Z]\| \geqslant t) &\leqslant \mathbb{P}_E(\|Z - \mathbb{E}_E[Z]\| \geqslant t - \|\mathbb{E}[Z] - \mathbb{E}_E[Z]\|) \\
&\leqslant \mathbb{P}_E(\|Z - \mathbb{E}_E[Z]\| \geqslant t/2).
\end{aligned}$$

To conclude, we apply Bernstein's inequality (Lemma G.2) to $Y_j = A_j - \mathbb{E}[A_j|E] = Y_j = A_j - \mathbb{E}[A_j|E_j]$ conditional to E. Observe that

$$0 \preceq \mathbb{E}_E[Y_j Y_j^\top] \preceq \mathbb{E}_E[A_j A_j^\top] - \mathbb{E}_E[A_j]\mathbb{E}_E[A_j]^\top] \preceq \mathbb{E}_E[A_j A_j^\top],$$

which yields $\left\|\mathbb{E}_E[Y_j Y_j^\top]\right\| \leqslant \left\|\mathbb{E}[A_j A_j^\top]\right\|$ and similarly, $\left\|\mathbb{E}_E[Y_j^\top Y_j]\right\| \leqslant \left\|\mathbb{E}_E[A_j^\top A_j]\right\|$. So by Bernstein's inequality

$$\mathbb{P}_E(\|Z - \mathbb{E}_E[Z]\| \geqslant t/2) \leqslant 2(d_1+d_2)\exp\left(-\frac{mt^2/4}{\sigma^2 + Lt/3}\right).$$

□

Corollary C.1. *Let $x, x' \in \mathcal{X}$. If*

$$\mathbb{P}(E_\omega^c) \leqslant \frac{t}{t + 4\left\|K^{(ij)}(x,x')\right\|} \quad \textit{and} \quad \mathbb{E}[L_{ij}(\omega)1_{E_\omega^c}] \leqslant \frac{t}{4}$$

then $\left\|K_{\bar{E}}^{(ij)}(x,x') - K^{(ij)}(x,x')\right\| \leqslant t/2$.

Proposition C.1. *Let $t > 0$ and assume that*

$$\mathbb{P}(E_\omega^c) \leqslant \frac{t}{t+6} \quad \textit{and} \quad \mathbb{E}[L_{01}(\omega)^2 1_{E_\omega^c}] \leqslant \frac{t}{4s}$$

then $\|\Upsilon - \Upsilon_{\bar{E}}\| \leqslant t/2$ and

$$\mathbb{P}_{\bar{E}}(\left\|\Upsilon - \hat{\Upsilon}\right\| \geqslant t) \leqslant 4(d+1)s\exp\left(-\frac{mt^2/4}{s\bar{L}_{01}^2(3+t/3)}\right)$$

Consequently,

$$\mathbb{P}_{\bar{E}}(\left\|\Upsilon^{-1} - \hat{\Upsilon}^{-1}\right\| \geqslant t) \leqslant 4(d+1)s\exp\left(-\frac{mt^2}{16s\bar{L}_{01}^2(3+2\tilde{t})}\right).$$

Proof. We apply Lemma C.5 to $A_j = \gamma(\omega_j)\gamma(\omega_j)^*$ with the following observations:

- for each ω,
$$\|\gamma(\omega)\gamma(\omega)^*\| \leqslant \|\gamma(\omega)\|^2 \leqslant s\max_{x\in\mathcal{X}}\{\|\mathrm{D}_1[\varphi_\omega](x)\|^2 + |\varphi_\omega(x)|^2\},$$
so under event $\bar{E}$, $\|A_j\| \leqslant s\bar{L}_{01}^2$.
- By Lemma C.1, $\|\mathbb{E}[A_j]\| = \|\Upsilon\| \leqslant 3/2$,
- We may set $\sigma^2 = \bar{L}_{01}(3/2 + t/2)$ since
$$0 \preceq \mathbb{E}_{\bar{E}}[A_1 A_1^*] = \mathbb{E}_{\bar{E}}[A_1^* A_1] = \mathbb{E}_{\bar{E}}[\|\gamma(\omega_j)\|^2 \gamma(\omega_j)\gamma(\omega_j)^*] \preceq \bar{L}_{01}(\|\mathbb{E}[A_j]\| + t/2)\mathrm{Id}.$$

The last claim is because $\left\|\Upsilon - \hat{\Upsilon}\right\| \leqslant t$ implies that $\|\Upsilon\| \leqslant 3/2 + t$, $\left\|\Upsilon^{-1}\right\| \leqslant \frac{\|\Upsilon\|}{1-\left\|\Upsilon-\hat{\Upsilon}\right\|\|\Upsilon^{-1}\|} \leqslant \frac{3}{2-4t}$ and $\left\|\Upsilon^{-1} - \hat{\Upsilon}^{-1}\right\| \leqslant \|\Upsilon^{-1}\|\left\|\Upsilon - \hat{\Upsilon}\right\|\left\|\hat{\Upsilon}^{-1}\right\| \leqslant \frac{3t}{1-2t}$ and writing $\tilde{t} = \frac{3t}{1-2t}$ is equivalent to $t = \tilde{t}/(3+2\tilde{t})$. □

Bounds on $\hat{\mathbf{f}}_X$ applied to a fixed vector

Proposition C.2. *Let $t \in (0,1)$, $r \in \{0,2\}$, $q \in \mathbb{C}^{s(d+1)}$ and $y \in \mathcal{X}_r$, where $\mathcal{X}_0 \overset{\text{def.}}{=} \mathcal{X}$ and $\mathcal{X}_2 \overset{\text{def.}}{=} \mathcal{X}^{\text{near}}$. If*

$$\mathbb{P}(E_\omega^c) \leqslant \frac{t}{t+4B_r} \quad \textit{and} \quad \mathbb{E}[L_{01}(\omega)L_r(\omega)1_{E_\omega^c}] \leqslant \frac{t}{4\sqrt{s}}$$

then

$$\mathbb{P}_{\bar{E}}\left(\left\|\mathrm{D}_r\left[(\hat{\mathbf{f}}_{X_0} - \mathbf{f}_{X_0})^\top q\right](y)\right\| \geqslant t\|q\|\right) \leqslant 2\tilde{d}\exp\left(\frac{-mt^2/4}{2\bar{L}_r^2 + \bar{L}_r\bar{L}_{01}t/(3\sqrt{s})}\right)$$

where $\tilde{d} = 1$ *if* $r = 0$ *and* $\tilde{d} = d$ *if* $r = 2$.

As a consequence, since $\sqrt{2s}\,\|q\|_{*,\infty} \geqslant \|q\|_2$, *we have*

$$\mathbb{P}_E\left(\left\|\mathrm{D}_r\left[(\mathbf{f}_{X_0} - \hat{\mathbf{f}}_{X_0})^\top q\right](y)\right\| \geqslant t\,\|q\|_{*,\infty}\right) \leqslant 2\tilde{d}\exp\left(\frac{-mt^2}{16s(\bar{L}_r^2 + 8\bar{L}_r\bar{L}_{01}t/(3\sqrt{2}))}\right)$$

provided that

$$\mathbb{P}(E_\omega^c) \leqslant \frac{t}{t + 4\sqrt{2s}B_r} \quad \textit{and} \quad \mathbb{E}[L_{01}(\omega)L_r(\omega)1_{E_\omega^c}] \leqslant \frac{t}{4\sqrt{2}s}.$$

Proof. Without loss of generality, assume that $\|q\| = 1$. First note that

$$\mathrm{D}_r\left[(\hat{\mathbf{f}}_{X_0} - \mathbf{f}_{X_0})^\top q\right](y) = \frac{1}{m}\sum_{k=1}^m q^\top\gamma(\omega_k)\mathrm{D}_r\left[\varphi_{\omega_k}\right](y) - \mathbb{E}[q^\top\gamma(\omega_k)\mathrm{D}_r\left[\varphi_{\omega_k}\right](y)].$$

We first consider the case of $r = 0$. We apply Lemma C.5 to $A_k \overset{\text{def.}}{=} q^\top\gamma(\omega_k)\varphi_{\omega_k}(y) \in \mathbb{C}$: Note that $|A_k| \leqslant \sqrt{s}L_{01}(\omega_k)L_0(\omega_k)$ and $|\mathbb{E}[A_k]| \leqslant B_0$.

- Under event E_{ω_k}, $\|A_k\| \leqslant \bar{L}_2\bar{L}_{01}\sqrt{s} \overset{\text{def.}}{=} L$.
- $\mathbb{E}_{\bar{E}}\,|A_k|^2 = \mathbb{E}_{\bar{E}}[\langle\gamma(\omega_k)\gamma(\omega_k)^* q,\, q\rangle\,|\varphi_{\omega_k}(y)|^2] \leqslant \bar{L}_0^2\,\|\Upsilon_{\bar{E}}\| \leqslant (3/2 + t/2)\,\bar{L}_0^2 \leqslant 2\bar{L}_0^2 \overset{\text{def.}}{=} \sigma^2$.

For the case $r = 2$, we apply Lemma C.5 with $A_k \overset{\text{def.}}{=} q^\top\gamma(\omega_k)\mathrm{D}_2\left[\varphi_{\omega_k}\right](y) \in \mathbb{C}^{d\times d}$. Then, $\|A_k\| \leqslant \sqrt{s}L_{01}(\omega_k)L_2(\omega_k)$, $\|\mathbb{E}[A_k]\| \leqslant B_2$, under event E_{ω_k}, $\|A_k\| \leqslant \bar{L}_2\bar{L}_{01}\sqrt{s} \overset{\text{def.}}{=} L$ and

$$\|\mathbb{E}_{\bar{E}}[A_kA_k^*]\| = \|\mathbb{E}_{\bar{E}}[A_k^*A_k]\| = \left\|\mathbb{E}_{\bar{E}}[\mathrm{D}_2\left[\varphi_{\omega_k}\right](y)^*\mathrm{D}_2\left[\varphi_{\omega_k}\right](y)\left|q^\top\gamma(\omega_k)\right|^2]\right\| \leqslant \bar{L}_2^2\mathbb{E}_{\bar{E}}[\left|q^\top\gamma(\omega_k)\right|^2] \leqslant 2\bar{L}_2^2 \overset{\text{def.}}{=} \sigma^2.$$

□

Lemma C.6. *Assume that*

$$\mathbb{P}(E_\omega^c) \leqslant \frac{t}{t + 6\sqrt{2s}} \quad \textit{and} \quad \mathbb{E}[L_{01}(\omega)^2 1_{\bar{E}^c}] \leqslant \frac{t}{4\sqrt{2}s^{3/2}}$$

Let $q \in \mathbb{C}^{s(d+1)}$. *Then, for all* $t \geqslant \frac{2\sqrt{2}s\bar{L}_{01}\bar{L}_1}{m} + \sqrt{\frac{8s^2\bar{L}_{01}^2\bar{L}_1^2}{m^2} + \frac{144s\bar{L}_1^2}{m}}$, *we have for each* $x_i \in X$,

$$\mathbb{P}_E\left(\left\|\mathrm{D}_1\left[q^\top(\mathbf{f}_X - \hat{\mathbf{f}}_X)\right](x_i)\right\|_2 > 2t\,\|q\|_{*,\infty}\right) \leqslant 28\exp\left(-\frac{mt^2/(4s)}{2\bar{L}_1^2 + \sqrt{2}t\bar{L}_1\bar{L}_{01}/3}\right).$$

Proof. For each $x_i \in X$,

$$\left\|\mathrm{D}_1\left[(\mathbb{E}_{\bar{E}}[q^\top\hat{\mathbf{f}}_X] - q^\top\mathbf{f}_X)\right](x_i)\right\| \leqslant \|\Upsilon - \Upsilon_{\bar{E}}\|\,\|q\| \leqslant \frac{t}{\sqrt{2s}}\,\|q\|,$$

by Proposition C.1. For convenience, we drop the subscript X from $\mathbf{f}_X$. Fix $i \in \{1, \ldots, s\}$. Observe that

$$\begin{aligned}\mathbb{P}_E\left(\left\|\mathrm{D}_1\left[q^\top(\mathbf{f} - \hat{\mathbf{f}})\right](x_i)\right\|_2 > 2t\,\|q\|_{*,\infty}\right) &\leqslant \mathbb{P}_E\left(\left\|\mathrm{D}_1\left[q^\top(\mathbf{f} - \hat{\mathbf{f}})\right](x_i)\right\|_2 > \frac{2t}{\sqrt{2s}}\,\|q\|_2\right)\\ &\leqslant \mathbb{P}_E\left(\left\|\mathrm{D}_1\left[q^\top(\mathbb{E}_{\bar{E}}[\hat{\mathbf{f}}] - \hat{\mathbf{f}})\right](x_i)\right\|_2 > \frac{t}{\sqrt{2s}}\,\|q\|_2\right)\end{aligned}$$

The claim of this lemma follows by applying Lemma G.3: Let

$$Y_k = \mathrm{D}_1\left[\varphi_{\omega_k}\right](x_i)\gamma(\omega_k)^\top q - \mathbb{E}_{\bar{E}}\mathrm{D}_1\left[\varphi_{\omega_k}\right](x_i)\gamma(\omega)^\top q \in \mathbb{C}^d,$$

and observe that $\mathrm{D}_1\left[q^\top(\hat{\mathbf{f}} - \mathbb{E}_{\bar{E}}[\hat{\mathbf{f}}])\right](x_i) = \frac{1}{m}\sum_k Y_k$. Without loss of generality, assume that $\|q\|_2 = 1$. We apply Lemma G.3. Observe that conditional on event E,

- $\|Y_k\|_2 \leqslant 2\|q\|_2 \|\gamma(\omega_k)\|_2 \|\mathrm{D}_1[\varphi_{\omega_k}](x_i)\|_2 \leqslant 2\sqrt{s}\bar{L}_{01}\bar{L}_1$.
- $\mathbb{E}_E\|Y_k\|^2 \leqslant \mathbb{E}_E[\left|\gamma(\omega_k)^\top q\right|^2 \mathrm{D}_1[\varphi_{\omega_k}](x_i)\mathrm{D}_1[\varphi_{\omega_k}](x_i)^\top] \leqslant \bar{L}_1^2\|\Upsilon_E\|$. So, $\sigma^2 \leqslant m\bar{L}_1^2\|\Upsilon_E\| \leqslant m\bar{L}_1^2(t + \|\Upsilon\|) \leqslant m\bar{L}_1^2(t/2 + 3/2) \leqslant 2m\bar{L}_1^2$ (here we are talking about the σ^2 in Lemma G.3).

Therefore, for all

$$t \geqslant \frac{2\sqrt{2}s\bar{L}_{01}\bar{L}_1}{m} + \sqrt{\frac{8s^2\bar{L}_{01}^2\bar{L}_1^2}{m^2} + \frac{144s\bar{L}_1^2}{m}}$$
$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{k=1}^m Y_k\right\|_2 \geqslant \frac{t}{\sqrt{2s}}\right) \leqslant 28\exp\left(-\frac{mt^2/(4s)}{2\bar{L}_1^2 + \sqrt{2}t\bar{L}_1\bar{L}_{01}/3}\right)$$

□

Proposition C.3 (Block norm bound on $\hat{\Upsilon}$ applied to a fixed vector)**.** *Suppose that*

$$\mathbb{P}(E_\omega^c) \leqslant \frac{t}{t + 6\sqrt{s}(B_0+1)} \quad \textit{and} \quad \mathbb{E}[L_{01}(\omega)^2 1_{\bar{E}^c}] \leqslant \frac{t}{4s^{3/2}(1+4B_0)}$$

Then, for all

$$t \geqslant \left(\frac{4\sqrt{2}s\bar{L}_{01}\bar{L}_1}{m} + \sqrt{\frac{32s^2\bar{L}_{01}^2\bar{L}_1^2}{m^2} + \frac{576s\bar{L}_1^2}{m}}\right)$$

we have

$$\mathbb{P}_E\left(\left\|(\Upsilon - \hat{\Upsilon})q\right\|_{*,\infty} \geqslant t\|q\|_{*,\infty}\right) \leqslant 32s\exp\left(-\frac{mt^2}{s\left(32\bar{L}_1^2 + 34t\bar{L}_1\bar{L}_{01}\right)}\right). \tag{C.8}$$

Proof. Let $S_0 \overset{\text{def.}}{=} \{1,\dots,s\}$ and $S_j \overset{\text{def.}}{=} \{s+(j-1)d+1,\dots,s+jd\}$ for $j = 1,\dots,s$. Observe that by the union bound

$$\begin{aligned}
&\mathbb{P}_E\left(\left\|(\Upsilon - \hat{\Upsilon})q\right\|_{*,\infty} \geqslant t\|q\|_{*,\infty}\right) \\
&\leqslant \mathbb{P}_E\left(\left\|((\Upsilon - \hat{\Upsilon})q)_{S_0}\right\|_\infty \geqslant t\|q\|_{*,\infty}\right) + \sum_{j=1}^s \mathbb{P}_E\left(\left\|((\Upsilon - \hat{\Upsilon})q)_{S_j}\right\|_2 \geqslant t\|q\|_{*,\infty}\right) \\
&\leqslant \sum_{j=1}^s \mathbb{P}_E\left(\left|((\Upsilon - \hat{\Upsilon})q)_j\right| \geqslant t\|q\|_{*,\infty}\right) + \sum_{j=1}^s \mathbb{P}_E\left(\left\|((\Upsilon - \hat{\Upsilon})q)_{S_j}\right\|_2 \geqslant t\|q\|_{*,\infty}\right).
\end{aligned} \tag{C.9}$$

To bound the first sum, observe that $((\Upsilon - \hat{\Upsilon})q)_j = (\mathbf{f}(x_j) - \hat{\mathbf{f}}(x_j))^\top q$ and $((\Upsilon - \hat{\Upsilon})q)_{S_j} = \mathrm{D}_1\left[q^\top(\mathbf{f} - \hat{\mathbf{f}})\right](x_j)$. So, the first sum can be bounded by applying Proposition C.2. The second sum can be bounded by applying Lemma C.6.

□

Norm bounds for $\hat{\mathbf{f}}$ We will repeatedly make use of the following result on $\hat{\mathbf{f}}_X$. This result is due to concentration bounds on the kernel $\hat{K}$ which are derived subsequently.

Proposition C.4 (Bound on $\hat{\mathbf{f}}_X$)**.** *Let $X \in \mathcal{X}^s$. Let $\rho > 0$. Assume that for all $(i,j) \in \{(0,0),(1,0),(0,2),(1,2)\}$,*

$$\mathbb{P}(E_\omega^c) \leqslant \frac{t}{t + 4\sqrt{s}\max\{B_0, B_2\}}, \quad \mathbb{E}[L_i(\omega)L_j(\omega)1_{E_\omega^c}] \leqslant \frac{t}{4\sqrt{s}}$$

Then, given any $y \in \mathcal{X}$,

$$\mathbb{P}_{\bar{E}}\left(\left\|\hat{\mathbf{f}}_X(y) - \mathbf{f}_X(y)\right\| \geqslant t\right) \leqslant 4sd\exp\left(-\frac{mt^2/8}{3s\bar{L}_{01}^2}\right). \tag{C.10}$$

and given any $y \in \mathcal{X}^{\mathrm{near}}$, writing $\hat{\mathbf{f}}_X = (\hat{f}_j)_{j=1}^p$ and $\mathbf{f}_X = (f_j)_{j=1}^p$ with $p = s(d+1)$, we have

$$\mathbb{P}_{\bar{E}}\left(\sup_{\|q\|=1}\sqrt{\sum_{j=1}^p\left\|\mathrm{D}_2\left[\hat{f}_j - f_j\right](y)q\right\|^2} > t\right) \leqslant s(3d+d^2)\exp\left(-\frac{mt^2/8}{s(\bar{L}_2^2B_{11} + \bar{L}_1^2B_{22} + \bar{L}_{01}\bar{L}_2)}\right). \tag{C.11}$$

Proof. Let $i,j \in \mathbb{N}_0$ with $i+j \leqslant 2$. Let $[s] \overset{\text{def.}}{=} \{1,\dots,s\}$ and $I \overset{\text{def.}}{=} \{(0,0),(1,0)\}$, By Lemma C.7 and the union bound,

$$\mathbb{P}_{\bar{E}}\left(\exists(i,j)\in I, \exists \ell \in [s], \left\|\hat{K}^{(ij)}(x_\ell,y) - K^{(ij)}(x_\ell,y)\right\| \geqslant \frac{t}{\sqrt{s}}\right) \leqslant 4sd\exp\left(-\frac{mt^2/4}{3s\bar{L}_{01}^2}\right). \tag{C.12}$$

So, (C.10) follows because

$$\left\|\hat{\mathbf{f}}_X(y) - \mathbf{f}_X(y)\right\| \leqslant \sqrt{\sum_{i=1}^s\left|\hat{K}(x_i,y) - K(x_i,y)\right|^2 + \left\|\hat{K}^{(10)}(x_i,y) - K^{(10)}(x_i,y)\right\|^2} \leqslant \sqrt{2}t.$$

By Lemma C.7, Lemma C.9 and the union bound, letting $I_2 \overset{\text{def.}}{=} \{(0,2),(1,2)\}$, we have

$$\begin{aligned}\mathbb{P}_{\bar{E}}\left(\exists(i,j)\in I_2, \exists \ell \in [s], \left\|\hat{K}^{(ij)}(x_\ell,y) - K^{(ij)}(x_\ell,y)\right\| \geqslant \frac{t}{\sqrt{s}}\right) &\leqslant 2sd\exp\left(-\frac{mt^2/4}{2s(\bar{L}_2^2 + \bar{L}_0\bar{L}_2)}\right)\\ &+ s(d+d^2)\exp\left(-\frac{mt^2/4}{s(\bar{L}_2^2B_{11} + \bar{L}_1^2B_{22} + \bar{L}_1\bar{L}_2)}\right).\end{aligned} \tag{C.13}$$

and (C.11) follows since given $q \in \mathbb{C}^d$, $\|q\| = 1$, we have

$$\sum_{j=1}^p\left\|\mathrm{D}_2\left[\hat{f}_j - f_j\right](y)q\right\|^2 \leqslant \sum_{j=1}^s\left(\left\|\hat{K}^{(02)}(x_j,y) - K^{(02)}(x_j,y)\right\|^2 + \left\|\hat{K}^{(12)}(x_j,y) - K^{(12)}(x_j,y)\right\|^2\right) \leqslant 2t^2$$

□

Lemma C.7 (Concentration on kernel)**.** *Let $t > 0$, $x, x' \in \mathcal{X}$. Let $i,j \in \mathbb{N}_0$ with $i+j \leqslant 2$. Assume*

$$\mathbb{P}(E_\omega^c) \leqslant \frac{t}{t + 4\left\|K^{(ij)}(x,x')\right\|}, \quad \mathbb{E}[L_i(\omega)L_j(\omega)1_{E_\omega^c}] \leqslant \frac{t}{4}$$

then

$$\mathbb{P}_{\bar{E}}\left(\left\|\hat{K}^{(ij)}(x,x') - K^{(ij)}(x,x')\right\| \geqslant t\right) \leqslant 2d\exp\left(-\frac{mt^2}{\bar{L}_p^2(b_{ij}+1) + \bar{L}_i\bar{L}_jt/3}\right)$$

where $p = \max(i,j)$ and $b_{ij} = 1$ if $\min(i,j) = 0$ and $b_{ij} \overset{\text{def.}}{=} \left\|K^{(11)}(x,x')\right\|$ otherwise.

Proof. It is an immediate application of Lemma C.5 with $A_k = \mathrm{Re}\left(\overline{\mathrm{D}_i\left[\varphi_{\omega_k}\right](x)}\mathrm{D}_j\left[\varphi_{\omega_k}\right](x')^\top\right)$ for $k = 1,\dots,m$. Note that $A_k \in (\mathbb{R}^d)^{i+j}$ if $(i,j) \in \{(0,0),(0,1),(1,0)\}$ and $A_k \in \mathbb{R}^{d\times d}$ if $\max(i,j) = 2$. noting that under $\bar{E}$, $\|A_k\| \leqslant \bar{L}_i\bar{L}_j$. Next, we need to bound $\|\mathbb{E}_{\bar{E}}[A_kA_k^*]\|$ and $\|\mathbb{E}_{\bar{E}}[A_k^*A_k]\|$. We present only the argument for

$(i,j) = (0,2)$, since all the other cases are similar:

$$\begin{aligned} 0 \preceq \mathbb{E}_{\bar{E}} A_k A_k^* &\preceq \mathbb{E}_{\bar{E}}[\|\varphi_{\omega_k}(x')\|^2 \mathrm{D}_2[\varphi_{\omega_k}](x)\mathrm{D}_2[\varphi_\omega](x)^*] \\ &\preceq \bar{L}_2^2 \mathbb{E}_{\bar{E}} \|\varphi_{\omega_k}(x')\|^2 \mathrm{Id} = \bar{L}_2^2 |K_{\bar{E}}(x',x')| \mathrm{Id} \preceq (1+t/2)\bar{L}_2^2 \mathrm{Id} \end{aligned}$$

so $\|\mathbb{E}_{\bar{E}} A_k A_k^*\| \leqslant (1+t/2)\bar{L}_2^2$. Similarly, $\|\mathbb{E}_{\bar{E}} A_k^* A_k\| \leqslant (1+t/2)\bar{L}_2^2$ and

$$\|\mathbb{E}_{\bar{E}} A_k^* A_k\|, \|\mathbb{E}_{\bar{E}} A_k A_k^*\| \leqslant L_p^2(B_{qq} + t/2)$$

where $p = \max(i,j)$ and $q = \min(i,j)$.

□

Applying a grid on $\mathcal{X}^{\text{near}}$, we get a uniform version.

Lemma C.8. *Let $i,j \in \mathbb{N}_0$ with $i+j \leqslant 2$, and assume that*

$$\mathbb{P}(E_\omega^c) \leqslant \frac{t}{t + 16 B_{ij}}, \quad \mathbb{E}[L_i(\omega) L_j(\omega) 1_{E_\omega^c}] \leqslant \frac{t}{16}.$$

Then

$$\begin{aligned} &\mathbb{P}_{\bar{E}}\left(\exists\, x,x' \in \mathcal{X}^{\text{near}},\ \left\|\hat{K}^{(ij)}(x,x') - K^{(ij)}(x,x')\right\| \geqslant t\right) \\ &\qquad \leqslant 2ds^2 \exp\left(-\frac{mt^2/16}{L_p^2(B_{qq}+1) + \bar{L}_i\bar{L}_j t/12} + 2d\log\left(\frac{4(\mathcal{L}_i\bar{L}_j + \bar{L}_i\mathcal{L}_j)}{t}\right)\right). \end{aligned}$$

where $p = \max(i,j)$ and $q = \min(i,j)$ and $\mathcal{L}_i, \mathcal{L}_j$ are as in Lemma C.2

Proof. We define a δ-covering of $\mathcal{X}^{\text{near}}$ for the metric $d_{\mathbf{H}}$ with $\delta = \min\left(r_{\text{near}}, \frac{t}{4(\mathcal{L}_i\bar{L}_j + \bar{L}_i\mathcal{L}_j)}\right)$ of size $s\left(\frac{r_{\text{near}}}{\delta}\right)^d$. Let this covering be denoted by $\mathcal{X}^{\text{grid}}$.

By the union bound and Lemma C.7,

$$\mathbb{P}_{\bar{E}}\left(\exists x,x' \in \mathcal{X}^{\text{grid}} \text{ s.t. } \left\|\hat{K}^{(ij)}(x,x') - K^{(ij)}(x,x')\right\| \geqslant t/4\right) \leqslant 2ds^2\left(\frac{r_{\text{near}}}{\delta}\right)^{2d} \exp\left(-\frac{mt^2/16}{L_p^2(B_{qq}+1) + \bar{L}_i\bar{L}_j t/12}\right)$$

where $p = \max(i,j)$ and $q = \min(i,j)$. This gives the required upper bound: Given any $x,x' \in \mathcal{X}$, let $x_{\text{grid}}, x'_{\text{grid}} \in \mathcal{X}^{\text{grid}}$ be such that $d_{\mathbf{H}}(x,x_{\text{grid}}), d_{\mathbf{H}}(x',x'_{\text{grid}}) \leqslant \delta$. Then, under event $\bar{E}$, by Lemma C.2,

$$\left\|\hat{K}^{(ij)}(x,x') - \hat{K}^{(ij)}(x_{\text{grid}}, x'_{\text{grid}})\right\| \leqslant (\mathcal{L}_i\bar{L}_j + \bar{L}_i\mathcal{L}_j)\delta \leqslant t/4.$$

By Jensen's inequality and since $\left\|K_{\bar{E}}^{(ij)}(x,x') - K^{(ij)}(x,x')\right\| \leqslant t/4$ for all x,x', we have

$$\left\|K^{(ij)}(x,x') - K^{(ij)}(x_{\text{grid}}, x'_{\text{grid}})\right\| \leqslant t/2.$$

□

We now derive analogous results for the kernel differentiated 3 times.

Lemma C.9 (Concentration on order 3 kernel)**.** *Let $x,x' \in \mathcal{X}^{\text{near}}$. Assume that*

$$\mathbb{P}(E_\omega^c) \leqslant \frac{t}{t + 4\max\{B_{12}, B_{22}\}}, \quad \mathbb{E}[(L_1(\omega)L_2(\omega) + L_2^2(\omega))1_{E_\omega^c}] \leqslant \frac{t}{4}$$

For $j = 1, \dots, m$, let $a_i = (\mathrm{D}_1 \left[\overline{\varphi_{\omega_j}}\right](x))_i \in \mathbb{C}$, $D \overset{\text{def.}}{=} \mathrm{D}_2 \left[\varphi_\omega\right](x') \in \mathbb{C}^{d\times d}$ and

$$A_j \overset{\text{def.}}{=} \begin{pmatrix} a_1 D & a_2 D & \cdots & a_d D \end{pmatrix}^\top \in \mathbb{C}^{d^2\times d} \tag{C.14}$$

Let $Z \overset{\text{def.}}{=} \frac{1}{m}\sum_{j=1}^m (A_j - \mathbb{E}[A_j])$. Then, given

$$\begin{aligned} g(x') \overset{\text{def.}}{=} (g_i(x'))_{i=1}^d &\overset{\text{def.}}{=} \sum_{k=1}^m \left(\overline{\mathrm{D}_1\left[\varphi_{\omega_k}\right](x)}\varphi_\omega(x') - \mathbb{E}[\overline{\mathrm{D}_1\left[\varphi_{\omega_k}\right](x)}\varphi_\omega(x')]\right) \\ &= \hat{K}^{(10)}(x,x') - K^{(10)}(x,x'), \end{aligned}$$

(i) $\sup_{q\in\mathbb{C}^d,\|q\|\leqslant 1}\sum_{i=1}^d \left\|\mathrm{D}_2\left[g_i\right](x')q\right\|^2 = \|Z\|^2$,

(ii) $\sup_{q\in\mathbb{C}^d,\|q\|\leqslant 1}\left\|\mathrm{D}_2\left[q^\top g\right](x')\right\| = \left\|\hat{K}^{(12)}(x,x') - K^{(12)}(x,x')\right\| \leqslant \|Z\|$.

and

$$\mathbb{P}_{\bar{E}}\left(\|Z\| \geqslant t\right) \leqslant (d+d^2)\exp\left(-\frac{mt^2/4}{\tilde{B} + \bar{L}_1\bar{L}_2 t/3}\right)$$

where $\tilde{B} \overset{\text{def.}}{=} \max\{\bar{L}_2^2(B_{11}+t/2), \bar{L}_1^2(B_{22}+t/2)\}$.

Proof. The claim (i) is simply by definition, since $Zq = (\mathrm{D}_2\left[g_i\right](x')q)_{i=1}^d \in \mathbb{C}^{d^2}$. For (ii), the first equality is simply be definition, and for the inequality, observe that

$$\begin{aligned} \sup_{q\in\mathbb{C}^d,\|q\|\leqslant 1}\left\|\mathrm{D}_2\left[q^\top g\right](x')\right\| &= \sup_{q\in\mathbb{C}^d,\|q\|\leqslant 1}\sup_{p\in\mathbb{C}^d,\|p\|\leqslant 1}\left\|\sum_{i=1}^d q_i \mathrm{D}_2\left[g_i\right](x')p\right\| \\ &\leqslant \sup_{q\in\mathbb{C}^d,\|q\|\leqslant 1}\sup_{p\in\mathbb{C}^d,\|p\|\leqslant 1}\|q\|\sqrt{\sum_{i=1}^d \left\|\mathrm{D}_2\left[g_i\right](x')p\right\|^2} \leqslant \|Z\|. \end{aligned}$$

Finally, the probability bound follows by applying Lemma C.5: First note that under $\bar{E}$, $\|A_j\| \leqslant \bar{L}_1\bar{L}_2$. It remains to bound $\left\|\mathbb{E}_{\bar{E}}[A_j^* A_j]\right\|$ and $\left\|\mathbb{E}_{\bar{E}}[A_j A_j^*]\right\|$:

$$\begin{aligned} \sup_{\|q\|\leqslant 1}\mathbb{E}_{\bar{E}}\langle A_j^* A_j q,\, q\rangle &= \sup_{\|q\|\leqslant 1}\mathbb{E}_E\sum_{i=1}^d \left|(\mathrm{D}_1\left[\varphi_{\omega_j}\right](x))_i\right|^2 \left\|\mathrm{D}_2\left[\varphi_\omega\right](x')q\right\|^2 \\ &\leqslant \sup_{\|q_k\|\leqslant 1}\bar{L}_1^2\mathbb{E}_{\bar{E}}\overline{\mathrm{D}_2\left[\varphi_\omega\right](x')[q_1,q_2]}\mathrm{D}_2\left[\varphi_\omega\right](x')[q_3,q_4] \\ &\leqslant \bar{L}_1^2\left\|K_{\bar{E}}^{(22)}(x,x)\right\| \leqslant \bar{L}_1^2(B_{22}+t/2). \end{aligned}$$

Given $p_i \in \mathbb{C}^d$ for $i = 1,\dots,d$ such that $\sum_i \|p_i\|^2 \leqslant 1$, write $P = \begin{pmatrix} p_1 & p_2 & \cdots p_d \end{pmatrix} \in \mathbb{C}^{d\times d}$ and $\bar{p} = \begin{pmatrix} p_1^\top & p_2^\top & \cdots p_d^\top \end{pmatrix}^\top \in \mathbb{C}^{d^2}$. Then,

$$\begin{aligned} &\mathbb{E}_E\langle A_j A_j^*\bar{p},\, \bar{p}\rangle = \mathbb{E}_E\left\|\sum_{i=1}^d (\mathrm{D}_1\left[\varphi_{\omega_j}\right](x))_i \mathrm{D}_2\left[\varphi_{\omega_j}\right](x')p_i\right\|^2 \\ &= \mathbb{E}_E\left\|\mathrm{D}_2\left[\varphi_{\omega_j}\right](x')P\mathrm{D}_1\left[\varphi_{\omega_j}\right](x)\right\|^2 \\ &\leqslant \bar{L}_2^2\mathbb{E}_E\sum_i\left|\sum_k p_{i,k}(\mathrm{D}_1\left[\varphi_{\omega_j}\right](x))_k\right|^2 \\ &= \bar{L}_2^2\sum_i\langle\hat{K}_{\bar{E}}^{(11)}(x,x)p_i,\, p_i\rangle \leqslant \bar{L}_2^2\left\|\hat{K}_{\bar{E}}^{(11)}(x,x)\right\|^2\sum_i\|p_i\|^2 \leqslant \bar{L}_2^2(B_{11}+t/2). \end{aligned}$$

□

Lemma C.10 (Uniform concentration on order 3 kernel)**.** *Assume*

$$\mathbb{P}(E_\omega^c) \leqslant \frac{t}{t + 16\max\{B_{12}, B_{22}\}}, \quad \mathbb{E}[L_1(\omega)L_2(\omega)1_{E_\omega^c}] \leqslant \frac{t}{16}$$

then

$$\begin{aligned}\mathbb{P}_{\bar{E}}\left(\exists x, x' \in \mathcal{X}^{\text{near}},\ \left\|\hat{K}^{(12)}(x,x') - K^{(12)}(x,x')\right\| \geqslant t\right) \\ \leqslant s^2(d+d^2)\exp\left(-\frac{mt^2/16}{\tilde{B} + \bar{L}_1\bar{L}_2 t/6} + 2d\log\left(\frac{8(\mathcal{L}_1\bar{L}_2 + \bar{L}_2\mathcal{L}_2)}{t}\right)\right)\end{aligned}$$

where $\tilde{B} \overset{\text{def.}}{=} \max\{\bar{L}_2^2(B_{11} + t/2), \bar{L}_1^2(B_{22} + t/2)\}$*,* $\mathcal{L}_1$*,* $\mathcal{L}_2$ *are as in Lemma C.2.*

Proof. Let $\mathcal{X}^{\text{grid}}$ be a δ-covering of $\mathcal{X}^{\text{near}}$ for the metric $d_{\mathbf{H}}$ with $\delta = \min\left(r_{\text{near}}, \frac{t}{8(\mathcal{L}_1\bar{L}_2 + \mathcal{L}_2\bar{L}_2)}\right)$ of size at most $s\left(\frac{8(\mathcal{L}_1\bar{L}_2 + \mathcal{L}_2\bar{L}_2)}{t}\right)^d$. By Lemma C.9 and the union bound,

$$\begin{aligned}&\mathbb{P}_{\bar{E}}\left(\exists x, x' \in \mathcal{X}^{\text{grid}},\ \left\|\hat{K}^{(ij)}(x,x') - K^{(ij)}(x,x')\right\| \geqslant t/2\right) \\ &\leqslant s^2(d+d^2)\left(\frac{8(\bar{L}_1\bar{L}_2 + \bar{L}_2^2)}{t}\right)^{2d}\exp\left(-\frac{mt^2/16}{\bar{L}_2^2(B_{11} + t/4) + \bar{L}_1\bar{L}_2 t/6}\right) \overset{\text{def.}}{=} \rho.\end{aligned}$$

Moreover, under event $\bar{E}$, given any $x, x' \in \mathcal{X}^{\text{near}}$, there exists grid points x_{grid}, x'_{grid} such that

$$d_{\mathbf{H}}(x, x_{\text{grid}}), d_{\mathbf{H}}(x', x'_{\text{grid}}) \leqslant \delta$$

and

$$\begin{aligned}\left\|\left(\hat{K}^{(12)}(x,x') - K^{(12)}(x,x')\right)\right\| \leqslant &\left\|\left(\hat{K}^{(12)}(x_{\text{grid}}, x'_{\text{grid}}) - K^{(12)}(x_{\text{grid}}, x'_{\text{grid}})\right)\right\| \\ &+ \left\|\left(\hat{K}^{(12)}(x,x') - \hat{K}^{(12)}(x_{\text{grid}}, x'_{\text{grid}})\right)\right\| \\ &+ \left\|\left(K^{(12)}(x,x') - K^{(12)}(x_{\text{grid}}, x'_{\text{grid}})\right)\right\|,\end{aligned}$$

and by Lemma C.2, under event $\bar{E}$,

$$\left\|\left(\hat{K}^{(12)}(x,x') - \hat{K}^{(12)}(x_{\text{grid}}, x'_{\text{grid}})\right)\right\| \leqslant (\mathcal{L}_1\bar{L}_2 + \mathcal{L}_2\bar{L}_2)\delta \leqslant t/8.$$

and by Jensen's inequality and since $\left\|K^{(12)}(x,y) - K^{(12)}_{\bar{E}}(x,y)\right\| \leqslant t/8$,

$$\left\|\left(K^{(12)}(x,y) - K^{(12)}(x_{\text{grid}}, y)\right)\right\| \leqslant 3t/8.$$

Therefore, conditional on $\bar{E}$, $\left\|\left(\hat{K}^{(12)}(x,y) - K^{(12)}(x,y)\right)\right\| < t$ with probability at least $1 - \rho$.

□

D Proof of Theorem 3

In all the rest of the proofs we fix $X_0 \in \mathcal{X}^s$ to be Δ-separated points, $a_0 \in \mathbb{C}^s$, and let $\mathbf{u} = (\text{sign}(a_0), 0_{sd})$. We denote $\mathcal{X}_i^{\text{near}} = \{x \in \mathcal{X}\ ;\ d_{\mathbf{H}}(x, x_{0,i}) \leqslant r_{\text{near}}\}$ and $\mathcal{X}^{\text{near}} = \cup_i \mathcal{X}_i^{\text{near}}$ and $\mathcal{X}^{\text{far}} = \mathcal{X}\backslash\mathcal{X}^{\text{near}}$.

Since K is an admissible kernel, from (B.2) and (B.1) in the proof of Theorem 2 η_{X_0} satisfies

(i) for all $y \in \mathcal{X}^{\mathrm{far}}$, $|\eta_{X_0}(y)| \leqslant 1 - \frac{1}{2}\varepsilon_0$,

(ii) for all $y \in \mathcal{X}^{\mathrm{near}}(i)$, $-\mathrm{Re}\left(\mathrm{sign}(a_i)\mathrm{D}_2\left[\eta_{X_0}\right](y)\right) \succcurlyeq \frac{1}{2}\varepsilon_2\mathrm{Id}$ and $\left\|\mathrm{Im}\left(\mathrm{sign}(a_i)\mathrm{D}_2\left[\eta_{X_0}\right](y)\right)\right\| \leqslant (\frac{p}{2})\frac{1}{2}\varepsilon_2$.

$$p \stackrel{\mathrm{def.}}{=} \sqrt{(1-\varepsilon_2 r_{\mathrm{near}}^2/2)/(\varepsilon_2 r_{\mathrm{near}}^2/2)} \geqslant 1,$$

since $\varepsilon_2 r_{\mathrm{near}}^2 \leqslant 1$ by assumption of K being admissible. We aim to show that, for X close to X_0, $\hat{\eta}_X$ is nondegenerate by showing that $\left\|\mathrm{D}_r\left[\hat{\eta}_X\right] - \mathrm{D}_r\left[\eta_{X_0}\right]\right\| \leqslant c\varepsilon_r$ for some positive constant c sufficiently small.

D.1 Nondegeneracy of $\hat{\eta}_{X_0}$

We first establish the nondegeneracy of $\hat{\eta}_{X_0}$, our proof can be seen as a generalisation of the techniques in Tang et al. (2013) to the multidimensional setting with general sampling operators:

Theorem D.1. *Let $\rho > 0$ and assume that the assumptions in Section 2.3 hold. Assume also that either (a) or (b) holds:*

(a) $\mathrm{sign}(a_0)$ *is a Steinhaus sequence and*

$$m \gtrsim C \cdot s \cdot \log\left(\frac{N^d}{\rho}\right)\log\left(\frac{s}{\rho}\right)$$

(b) $\mathrm{sign}(a_0)$ *is an arbitrary sequence from the complex unit circle, and*

$$m \gtrsim C \cdot s^{3/2} \cdot \log\left(\frac{N^d}{\rho}\right)$$

where C, N are defined in the main paper. Then with probability at least $1-\rho$, the following hold: For all $y \in \mathcal{X}^{far}$, $|\hat{\eta}_{X_0}(y)| \leqslant 1 - \frac{7}{16}\varepsilon_0$, and for all $y \in \mathcal{X}^{near}(i)$, $-\mathrm{Re}\left(\mathrm{sign}(a_i)\mathrm{D}_2\left[\hat{\eta}_{X_0}\right](y)\right) \succcurlyeq \frac{7}{16}\varepsilon_2\mathrm{Id}$ and $\left\|\mathrm{Im}\left(\mathrm{sign}(a_i)\mathrm{D}_2\left[\hat{\eta}_{X_0}\right](y)\right)\right\| \leqslant (\frac{p}{2}+\frac{p}{8})\frac{1}{2}\varepsilon_2$ and hence, $\hat{\eta}_{X_0}$ is $(\frac{7}{16}\varepsilon_0, \frac{7}{16}\varepsilon_2)$-nondegenerate.

Proof. Note that

$$\frac{8}{7}\left(\frac{p}{2}+\frac{p}{8}\right) = \frac{5}{8}p < \sqrt{\frac{1-7\varepsilon_2 r_{\mathrm{near}}^2/16}{7\varepsilon_2 r_{\mathrm{near}}^2/16}}$$

so $\hat{\eta}_{X_0}$ is $(\frac{7}{16}\varepsilon_0, \frac{7}{16}\varepsilon_2)$-nondegenerate by Lemma B.1

Let $c \stackrel{\mathrm{def.}}{=} 1/32$. Observe that by assumption and Lemma C.4, $\mathbb{P}(\bar{E}) \leqslant \rho/2$. Therefore, it is sufficient to prove that conditional on $\bar{E}$, with probability at least $1-\delta$ with $\delta \stackrel{\mathrm{def.}}{=} \rho/2$, $\hat{\eta}_{X_0}$ is nondegenerate.

We will repeatedly use the fact that our assumptions (by Lemma C.4) also imply that

$$\mathbb{P}(E_\omega^c) \leqslant \frac{\varepsilon}{m}, \quad \mathbb{E}[L_i(\omega)L_j(\omega)1_{E_\omega^c}] \leqslant \frac{\varepsilon}{m}$$

for all $(i,j) \in \{(0,0),(1,0),(0,2),(1,2)\}$,

Step I: Proving nondegeneracy on a finite grid.

Let $\mathcal{X}_{\mathrm{grid}}^{\mathrm{far}} \subset \mathcal{X}^{\mathrm{far}}$ and $\mathcal{X}_{\mathrm{grid}}^{\mathrm{far}} \subset \mathcal{X}^{\mathrm{near}}$, be finite point sets. Let

$$Q_r(y) \stackrel{\mathrm{def.}}{=} \left\|\mathrm{D}_r\left[\hat{\eta}_{X_0}\right](y) - \mathrm{D}_r\left[\eta_{X_0}\right](y)\right\|, \qquad r = 0, 2.$$

We first prove that conditional on $\bar{E}$, with probability at least $1-\delta$ where $\delta \stackrel{\mathrm{def.}}{=} \rho/2$, that $Q_0(y) \leqslant c\varepsilon_0$ for all $y \in \mathcal{X}_{\mathrm{grid}}^{\mathrm{far}}$ and $Q_2(y) \leqslant c\varepsilon_2$ for all $y \in \mathcal{X}_{\mathrm{grid}}^{\mathrm{far}}$.

Let us first recall some facts which were proven in the previous section: Let $a, t \in (0,1)$ and write $\mathbf{f} = (\bar{f}_j)_{j=1}^{s(d+1)}$ and $\hat{\mathbf{f}} = (f_j)_{j=1}^{s(d+1)}$. Let $q_0 \stackrel{\mathrm{def.}}{=} \Upsilon^{-1}\mathbf{u}$, so $\|q_0\| \leqslant 2\sqrt{s}$. Let F be the event that

(a) $\left\|\Upsilon^{-1} - \hat{\Upsilon}^{-1}\right\| \leqslant t$,

(b) $\forall y \in \mathcal{X}^{\text{far}}_{\text{grid}}$, $\left\|\hat{\mathbf{f}}_{X_0}(y) - \mathbf{f}_{X_0}(y)\right\| \leqslant a\varepsilon_0$,

(c) $\forall y \in \mathcal{X}^{\text{near}}_{\text{grid}}$, $\sup_{q\in\mathbb{C}^d,\, \|q\|=1} \sqrt{\sum_{j=1}^p \left\|\mathrm{D}_2\left[f_j - \bar{f}_j\right](y)q\right\|^2} \leqslant a\varepsilon_2$,

Let G be the event that

(d) $\forall y \in \mathcal{X}^{\text{far}}_{\text{grid}}$, $\left|(\hat{\mathbf{f}}_{X_0}(y) - \mathbf{f}_{X_0}(y))^\top q_0\right| \leqslant 2a\varepsilon_0$

(e) $\forall y \in \mathcal{X}^{\text{near}}_{\text{grid}}$, $\left\|\mathrm{D}_2\left[(\hat{\mathbf{f}}_{X_0} - \mathbf{f}_{X_0})^\top q_0\right](y)\right\| \leqslant 2a\varepsilon_2$

then provided that

$$\mathbb{P}(E_\omega^c) \leqslant \frac{u}{u + \max\{4\sqrt{s}B_{ij}, 6\}}, \quad \mathbb{E}[L_i(\omega)L_j(\omega)1_{E_\omega^c}] \leqslant \frac{u}{4s} \tag{D.1}$$

where $u = \min\{a\varepsilon_i, t\}$, we have

$$\begin{aligned}
\mathbb{P}_{\bar{E}}(F^c) \leqslant & 4(d+1)s\exp\left(-\frac{mt^2}{16s\bar{L}_{01}^2(3+2t)}\right) \\
& + 4sd\left|\mathcal{X}^{\text{far}}_{\text{grid}}\right|\exp\left(-\frac{m(a\varepsilon_0)^2/8}{s(\bar{L}_{01}^2(B_{11}+1)+\bar{L}_{01}^2)}\right) \\
& + s(3d+d^2)\left|\mathcal{X}^{\text{near}}_{\text{grid}}\right|\exp\left(-\frac{m(a\varepsilon_2)^2/8}{s(\bar{L}_2^2B_{11}+\bar{L}_1^2B_{22})+\bar{L}_{01}\bar{L}_2)}\right) \\
\mathbb{P}_{\bar{E}}(G^c) \leqslant & 2\left|\mathcal{X}^{\text{far}}_{\text{grid}}\right|\exp\left(-\frac{ma^2\varepsilon_0^2}{s(8\bar{L}_0^2+\frac{4}{3}\bar{L}_0\bar{L}_{01}a\varepsilon_0)}\right) \\
& + 2d\left|\mathcal{X}^{\text{near}}_{\text{grid}}\right|\exp\left(-\frac{ma^2\varepsilon_2^2}{s(8\bar{L}_2^2+\frac{4}{3}\bar{L}_2\bar{L}_{01}a\varepsilon_2)}\right),
\end{aligned} \tag{D.2}$$

where for $\mathbb{P}_{\bar{E}}(F^c)$, the first term on the right is due to Proposition C.1, the second and third are due to Proposition C.4 while the bound on $\mathbb{P}_{\bar{E}}(G^c)$ is due to Proposition C.2 (noting that, since this probability bound over the ω_j is valid for all fixed $\mathbf{u}$, and the ω_j and the signs are independent, it is valid with the same probability over both ω_j and $\mathbf{u}$).

Observe that

$$\begin{aligned}
\|\mathrm{D}_j\left[\hat{\eta}_{X_0}\right](y) - \mathrm{D}_j\left[\eta_{X_0}\right](y)\| &= \left\|\mathrm{D}_j\left[(\hat{\alpha}_{X_0} - \alpha_{X_0})^\top\hat{\mathbf{f}}_{X_0}\right](y) + \mathrm{D}_j\left[\alpha_{X_0}^\top(\hat{\mathbf{f}}_{X_0} - \mathbf{f}_{X_0})\right](y)\right\| \\
&\leqslant \left\|\mathrm{D}_j\left[\mathbf{u}^\top\left((\hat{\Upsilon}^{-1} - \Upsilon^{-1})\hat{\mathbf{f}}_{X_0} + \Upsilon^{-1}(\hat{\mathbf{f}}_{X_0} - \mathbf{f}_{X_0})\right)\right](y)\right\|
\end{aligned} \tag{D.3}$$

Step I (a): Random signs

We first bound (D.3) in the case where $\mathbf{u}$ is a Steinhaus sequence.

Let $\beta_1(y) \overset{\text{def.}}{=} (\hat{\Upsilon}^{-1} - \Upsilon^{-1})\hat{\mathbf{f}}_{X_0}(y)$ and $\beta_2(y) \overset{\text{def.}}{=} \Upsilon^{-1}(\hat{\mathbf{f}}_{X_0}(y) - \mathbf{f}_{X_0}(y))$. Then, event F implies that $\|\beta_1(y)\| \leqslant t(B_0 + a\varepsilon_0)$ for all $y \in \mathcal{X}^{\text{far}}_{\text{grid}}$, and event G implies that $\left|\mathbf{u}^\top\beta_2(y)\right| \leqslant 2a\varepsilon_0$. So,

$$\begin{aligned}
&\mathbb{P}_{\bar{E}}\left(\left|\exists y \in \mathcal{X}^{\text{far}}_{\text{grid}},\ \mathbf{u}^\top(\beta_1+\beta_2)(y)\right| > c\varepsilon_0\right) \\
&\quad\leqslant \mathbb{P}_{F\cap\bar{E}}\left(\exists y \in \mathcal{X}^{\text{far}}_{\text{grid}},\ \left|\mathbf{u}^\top\beta_1(y)\right| > \frac{c}{2}\varepsilon_0\right)\mathbb{P}_{\bar{E}}(F) + \mathbb{P}_{\bar{E}}\left(F^c\right) \\
&\qquad + \mathbb{P}_{G\cap\bar{E}}\left(\exists y \in \mathcal{X}^{\text{far}}_{\text{grid}},\ \left|\mathbf{u}^\top\beta_2(y)\right| > \frac{c}{2}\varepsilon_0\right)\mathbb{P}_{\bar{E}}(G) + \mathbb{P}_{\bar{E}}\left(G^c\right) \\
&\quad\leqslant \mathbb{P}_{F\cap\bar{E}}\left(\exists y \in \mathcal{X}^{\text{far}}_{\text{grid}},\ \left|\mathbf{u}^\top\beta_1\right| > \frac{c}{2}\varepsilon_0\right) + \mathbb{P}_{\bar{E}}\left(F^c\right) + \mathbb{P}_{\bar{E}}\left(G^c\right) \\
&\quad\leqslant 4\left|\mathcal{X}^{\text{far}}_{\text{grid}}\right| e^{-\frac{(c/4)^2\varepsilon_0^2}{8t^2(B_0+a\varepsilon_0)^2}} + \mathbb{P}_{\bar{E}}(F^c) + \mathbb{P}_{\bar{E}}\left(G^c\right).
\end{aligned} \tag{D.4}$$

where we set $a = c/4$ for the second inequality and the last inequality follows from Lemma G.4 and because $\mathbf{u}$ consists if random signs.

Now consider $Q_2(y) = \mathrm{D}_2\left[\mathbf{u}^\top \beta\right](y)$. Under event G, $\left\|\mathrm{D}_2\left[\mathbf{u}^\top \beta_2\right](y)\right\| \leqslant \frac{c}{2}\varepsilon_2$. Writing $M = (\hat{\Upsilon}^{-1} - \Upsilon^{-1})$, we have

$$\mathrm{D}_2\left[\mathbf{u}^\top \beta_1\right](y) = \mathrm{D}_2\left[\mathbf{u}^\top \left(M\hat{\mathbf{f}}_{X_0}\right)\right](y) = \sum_{\ell=1}^{p} \mathbf{u}_\ell \left(\sum_{j=1}^{p} M_{\ell j}\mathrm{D}_2\left[f_j\right](y)\right). \tag{D.5}$$

We aim to bound (D.5) by applying the Matrix Hoeffding's inequality (Corollary G.1): let

$$Y_\ell \overset{\text{def.}}{=} \mathrm{Re}\left(\sum_{j=1}^{p} M_{\ell j}\mathrm{D}_2\left[f_j\right](y)\right) \in \mathbb{R}^{d\times d}$$

which is a symmetric matrix. Note that

$$\left\|\sum_{\ell=1}^{p} Y_\ell^2\right\| = \sup_{q\in\mathbb{R}^d, \|q\|=1} \sum_{\ell=1}^{p} \langle Y_\ell^2 q,\, q\rangle = \sup_{q\in\mathbb{R}^d, \|q\|=1} \sum_{\ell=1}^{d} \|Y_\ell q\|^2 \leqslant \sup_{q\in\mathbb{R}^d, \|q\|=1} \left\|\sum_{j=1}^{p} M_{\ell,j}(\mathrm{D}_2\left[f_j\right](y)q)\right\|^2.$$

Then, for a vector q of unit norm, let $V_{j,n} \overset{\text{def.}}{=} (\mathrm{D}_2\left[f_j\right](y)q)_n$ for $j = 1,\dots,p$ and $n = 1,\dots,d$, then

$$\begin{aligned}\sum_{\ell=1}^{p}\left\|\sum_{j=1}^{p} M_{\ell,j}(\mathrm{D}_2\left[f_j\right](y)q)\right\|^2 &= \sum_{\ell=1}^{p}\sum_{n=1}^{d}\left|\sum_{j=1}^{p} M_{\ell,j}V_{j,n}\right|^2 = \sum_{n=1}^{d}\|MV_{\cdot,n}\|^2 \leqslant \|M\|^2\sum_{n=1}^{d}\|V_{\cdot,n}\|^2 \\ &= \|M\|^2\sum_{n=1}^{d}\sum_{j=1}^{p}|V_{j,n}|^2 = \|M\|^2\sum_{j=1}^{p}\|\mathrm{D}_2\left[f_j\right](y)q\|^2.\end{aligned}$$

Under event F, we have $\|M\|^2\sum_{j=1}^{p}\|\mathrm{D}_2\left[f_j\right](y)q\|^2 \leqslant t^2(B_2 + a\varepsilon_2)^2$. Then,

$$\mathbb{P}_{F\cap\bar{E}}\left(\left\|\mathrm{D}_2\left[\mathbf{u}^\top \mathrm{Re}\left(M\hat{\mathbf{f}}_{X_0}\right)\right](y)\right\| \geqslant \frac{c\varepsilon_2}{\sqrt{2}}\right) \leqslant 2d\exp\left(-\frac{(c/2)^2\varepsilon_2^2}{4t^2(B_2+a\varepsilon_2)^2}\right).$$

By repeating this argument for the imaginary part, we obtain

$$\mathbb{P}_{F\cap\bar{E}}\left(\left\|\mathrm{D}_2\left[\mathbf{u}^\top \mathrm{Im}\left(M\hat{\mathbf{f}}_{X_0}\right)\right](y)\right\| \geqslant \frac{c\varepsilon_2}{\sqrt{2}}\right) \leqslant 2d\exp\left(-\frac{(c/2)^2\varepsilon_2^2}{4t^2(B_2+a\varepsilon_2)^2}\right).$$

So,

$$\begin{aligned}&\mathbb{P}_{\bar{E}}\left(\exists y\in\mathcal{X}_{\text{grid}}^{\text{near}},\ \left\|\mathrm{D}_2\left[\mathbf{u}^\top\beta(y)\right]\right\| > c\varepsilon_2\right) \\ &\quad\leqslant \mathbb{P}_{F\cap\bar{E}}\left(\exists y\in\mathcal{X}_{\text{grid}}^{\text{near}},\ \left\|\mathrm{D}_2\left[\mathbf{u}^\top\mathrm{Re}\left(M\hat{\mathbf{f}}_{X_0}\right)\right](y)\right\| \geqslant \frac{c}{2}\varepsilon_2\right) + \mathbb{P}_{\bar{E}}(F^c) + \mathbb{P}_{\bar{E}}(G^c) \\ &\quad\leqslant 4d\left|\mathcal{X}_{\text{grid}}^{\text{near}}\right|\exp\left(-\frac{(c/2)^2\varepsilon_2^2}{4t^2(B_2+a\varepsilon_2)^2}\right) + \mathbb{P}_{\bar{E}}(F^c) + \mathbb{P}_{\bar{E}}(G^c).\end{aligned} \tag{D.6}$$

Therefore,

$$\begin{aligned}&1 - \mathbb{P}\left(Q_0(y_0) \leqslant c\varepsilon_0 \text{ and } Q_2(y_2) \leqslant c\varepsilon_2, \forall y_0\in\mathcal{X}_{\text{grid}}^{\text{far}}, \forall y_2\in\mathcal{X}_{\text{grid}}^{\text{near}}\right) \\ &\leqslant 4\left|\mathcal{X}_{\text{grid}}^{\text{far}}\right|\exp\left(-\frac{(c/2)^2\varepsilon_0^2}{32t^2(B_0+a\varepsilon_0)^2}\right) + 4d\left|\mathcal{X}_{\text{grid}}^{\text{near}}\right|\exp\left(-\frac{(c/2)^2\varepsilon_2^2}{16t^2(B_2+a\varepsilon_2)^2}\right) + 2\mathbb{P}_{\bar{E}}(F^c) + 2\mathbb{P}_{\bar{E}}(G^c).\end{aligned}$$

The first 2 terms are each bounded by $\delta/7$ by setting t such that

$$\frac{1}{t^2} = 2^{13} \log\left(\frac{112\bar{N}d}{\delta}\right) \frac{\left(\bar{B}+1\right)}{c^2\varepsilon^2}$$

where $\bar{B} \overset{\text{def.}}{=} \max\{B_0, B_2\}$, $\varepsilon \overset{\text{def.}}{=} \min\{\varepsilon_0, \varepsilon_2\}$ and $\bar{N} = \max\left(\left|\mathcal{X}_{\text{grid}}^{\text{near}}\right|, \left|\mathcal{X}_{\text{grid}}^{\text{far}}\right|\right)$. The first term of (D.2) is bounded by $\delta/7$ if

$$m \geqslant \frac{1}{t^2} \log\left(\frac{28(d+1)s}{\delta}\right) 64s\bar{L}_{01}^2 = s\bar{L}_{01}^2 \frac{2^{19}\left(\bar{B}+1\right)}{c^2\varepsilon^2} \log\left(\frac{112\bar{N}d}{\delta}\right) \log\left(\frac{28(d+1)s}{\delta}\right)$$

and the last 4 terms of (D.2) are each bounded by $\delta/7$ provided that

$$m \gtrsim \log\left(\frac{28(s+d)d\bar{N}}{\delta}\right) \frac{16s(\bar{L}_2^2 B_{11} + \bar{L}_1^2 B_{22} + \bar{L}_{01}\bar{L}_2)}{c^2\varepsilon^2}$$

So, to summarise, recalling that $\delta = \rho/2$, $\hat{\eta}_{X_0}$ is nondegenerate on $\mathcal{X}_{\text{grid}}^{\text{near}}$ and $\mathcal{X}_{\text{grid}}^{\text{far}}$ with probability at least $1-\delta$ (conditional on $\bar{E}$) provided that

$$m \gtrsim \log\left(\frac{sdN}{\rho}\right) \log\left(\frac{sd}{\rho}\right) \frac{s(\bar{L}_2^2 B_{11} + \bar{L}_1^2 B_{22} + \bar{B}\bar{L}_{01}^2 + \bar{L}_{01}\bar{L}_2)}{\varepsilon^2}$$

and

$$\mathbb{P}(E_\omega^c) \lesssim \frac{\varepsilon}{\bar{B}^{3/2}\sqrt{s}\sqrt{\log(\bar{N}d/\rho)}} \quad \text{and} \quad , \quad \mathbb{E}[L_i(\omega)L_j(\omega)1_{E_\omega^c}] \lesssim \frac{\varepsilon}{4s\sqrt{\bar{B}}\sqrt{\log(\bar{N}d/\rho)}}$$

Step I (b): Deterministic signs Assume now that $\mathbf{u}$ consists of arbitrary signs. We will show that (D.3) can be bounded by $c\varepsilon$ when m is chosen as in condition (b) of this theorem. Let F' be the event that

(a') $\left\|\Upsilon - \hat{\Upsilon}\right\| \leqslant \frac{t}{s^{1/4}}$ and $\left\|\Upsilon^{-1} - \hat{\Upsilon}^{-1}\right\| \leqslant \frac{t}{s^{1/4}}$

(b') $\forall y \in \mathcal{X}_{\text{grid}}^{\text{far}}$, $\left\|(\hat{\mathbf{f}}_{X_0}(y) - \mathbf{f}_{X_0}(y))\right\| \leqslant \frac{a\varepsilon_0}{s^{1/4}}$

(c') $\forall y \in \mathcal{X}_{\text{grid}}^{\text{near}}$, $\sup_{\|q\|=1} \left\|\mathrm{D}_2\left[(\hat{\mathbf{f}}_{X_0} - \mathbf{f}_{X_0})^\top q\right](y)\right\| \leqslant \frac{a\varepsilon_2}{s^{1/4}}$

(f) $\left\|(\Upsilon - \hat{\Upsilon})\Upsilon^{-1}\mathbf{u}\right\|_{*,\infty} \leqslant a\varepsilon \left\|\Upsilon^{-1}\mathbf{u}\right\|_{*,\infty} \leqslant 2a\varepsilon.$

Then, provided that

$$\mathbb{P}(E_\omega^c) \leqslant \frac{u}{u + 6s(B_0+B_2)} \quad \text{and} \quad \mathbb{E}[L_{01}(\omega)^2 1_{\bar{E}^c}] \leqslant \frac{u}{4\bar{B}s^{3/2}},$$

with $u = \min\{a\varepsilon_i, t\}$ as before, we have

$$\begin{aligned}
\mathbb{P}_{\bar{E}}((F')^c) \leqslant & 4(d+1)s \exp\left(-\frac{mt^2}{16s^{3/2}\bar{L}_{01}^2(3+2t)}\right) \\
& + 4sd\left|\mathcal{X}_{\text{grid}}^{\text{far}}\right| \exp\left(-\frac{m(a\varepsilon_0)^2/8}{s^{3/2}(\bar{L}_{01}^2(B_{11}+1) + \bar{L}_{01}^2)}\right) \\
& + s(3d+d^2)\left|\mathcal{X}_{\text{grid}}^{\text{near}}\right| \exp\left(-\frac{m(a\varepsilon_2)^2/8}{s^{3/2}(\bar{L}_2^2 B_{11} + \bar{L}_1^2 B_{22} + \bar{L}_{01}\bar{L}_2)}\right) \\
& + 32s \exp\left(-\frac{m4a^2\varepsilon^2}{s\left(32L_1^2 + 68a\varepsilon L_1 \bar{L}_{01}\right)}\right).
\end{aligned}$$

where the first bound is from Proposition C.1, the second and third are from Proposition C.4 and the final bound is due to Proposition C.3.

To bound (D.3), we first observe that if event G holds, then just as observed previously, $\left|\mathrm{D}_r\left[\mathbf{u}^\top\beta_2\right](y)\right| \leqslant 2a\varepsilon_r$. To bound $\left|\mathbf{u}^\top\beta_1(y)\right|$, observe that

$$\begin{aligned}\mathbf{u}^\top\beta_1(y) &= \mathbf{u}^\top(\Upsilon^{-1}-\hat{\Upsilon}^{-1})(\hat{\mathbf{f}}_{X_0}-\mathbf{f}_{X_0}) + \mathbf{u}^\top(\Upsilon^{-1}-\hat{\Upsilon}^{-1})\mathbf{f}_{X_0}\\ &= \mathbf{u}^\top(\Upsilon^{-1}-\hat{\Upsilon}^{-1})(\hat{\mathbf{f}}_{X_0}-\mathbf{f}_{X_0}) + \mathbf{u}^\top\Upsilon^{-1}(\hat{\Upsilon}-\Upsilon)\hat{\Upsilon}^{-1}\mathbf{f}_{X_0}\\ &= \mathbf{u}^\top(\Upsilon^{-1}-\hat{\Upsilon}^{-1})(\hat{\mathbf{f}}_{X_0}-\mathbf{f}_{X_0}) + \mathbf{u}^\top\Upsilon^{-1}(\hat{\Upsilon}-\Upsilon)(\hat{\Upsilon}^{-1}-\Upsilon^{-1})\mathbf{f}_{X_0} + \mathbf{u}^\top\Upsilon^{-1}(\hat{\Upsilon}-\Upsilon)\Upsilon^{-1}\mathbf{f}_{X_0}\end{aligned}$$

Under event F',

- $\left|\mathbf{u}^\top(\Upsilon^{-1}-\hat{\Upsilon}^{-1})(\hat{\mathbf{f}}_{X_0}-\mathbf{f}_{X_0})\right| \leqslant \sqrt{s}\left\|\Upsilon^{-1}-\hat{\Upsilon}^{-1}\right\|\left\|\hat{\mathbf{f}}_{X_0}-\mathbf{f}_{X_0}\right\| \leqslant ta\varepsilon$
- $\left|\mathbf{u}^\top\Upsilon^{-1}(\hat{\Upsilon}-\Upsilon)(\hat{\Upsilon}^{-1}-\Upsilon^{-1})\mathbf{f}_{X_0}\right| \leqslant \sqrt{s}\cdot 2\cdot\left\|\hat{\Upsilon}-\Upsilon\right\|\left\|\hat{\Upsilon}^{-1}-\Upsilon^{-1}\right\|B_0 \leqslant 2t^2B_0$
- $\left\|\Upsilon^{-1}(\hat{\Upsilon}-\Upsilon)\Upsilon^{-1}\mathbf{u}\right\|_{*,\infty} \leqslant \left\|\Upsilon^{-1}\right\|_{*,\infty}\left\|(\hat{\Upsilon}-\Upsilon)\Upsilon^{-1}\mathbf{u}\right\|_{*,\infty} \leqslant 4a\varepsilon.$

Finally, given any vector q such that $\|q\|_{*,\infty} \leqslant 4a\varepsilon$, we have $\left|q^\top\mathbf{f}_{X_0}\right| \leqslant 4a\varepsilon B_0$. Therefore,

$$\left|\mathbf{u}^\top\beta_1(y)\right| \leqslant ta + 2t^2 + 4a\varepsilon B_0,$$

and in a similar manner, we can show that the same upper bound holds for $\left\|\mathrm{D}_2\left[\mathbf{u}^\top\beta_1\right](y)\right\|$.

Therefore,

$$\left\|\mathrm{D}_r\left[\mathbf{u}^\top\beta\right](y)\right\| \leqslant c\varepsilon_r \tag{D.7}$$

if both F' and G hold, so conditional on $\bar{E}$, (D.7) holds with probability at least $1-\delta$ provided that

$$m \gtrsim s^{3/2}\cdot\frac{(\bar{L}_2^2B_{11}+\bar{L}_1^2B_{22}+\bar{B}\bar{L}_{01}^2+\bar{L}_{01}\bar{L}_2)}{\varepsilon^2}\cdot\log\left(\frac{\bar{N}ds}{\rho}\right)$$

and

$$\mathbb{P}(E_\omega^c) \lesssim \frac{\varepsilon}{\bar{B}^{3/2}s\sqrt{\log(\bar{N}d/\rho)}} \quad \text{and} \quad , \quad \mathbb{E}[L_i(\omega)L_j(\omega)1_{E_\omega^c}] \lesssim \frac{\varepsilon}{s^{3/2}\sqrt{\bar{B}}\sqrt{\log(\bar{N}d/\rho)}}$$

Step II: Extending to the entire space To prove that $\hat{\eta}_{X_0}$ is nondegenerate on the entire space $\mathcal{X}$, we first show that $\hat{\eta}_{X_0}$ is locally Lipschitz (and hence determine how fine our grids $\mathcal{X}_{\mathrm{grid}}^{\mathrm{near}}$, $\mathcal{X}_{\mathrm{grid}}^{\mathrm{far}}$ need to be): for $x, x' \in \mathcal{X}$ with $d_{\mathbf{H}}(x,x') \leqslant r_{\mathrm{near}}$,

$$\begin{aligned}\left\|\mathrm{D}_r\left[\hat{\eta}_{X_0}\right](x) - \mathrm{D}_r\left[\hat{\eta}_{X_0}\right](x')\right\| &= \Bigg\|\frac{1}{m}\sum_{k=1}^m \mathrm{D}_r\left[\mathrm{Re}\left((\hat{\Upsilon}_X^{-1}\mathbf{u})^\top\gamma(\omega_k)\varphi_{\omega_k}\right)\right](x) \qquad \text{(D.8)}\\ &\qquad - \mathrm{D}_r\left[\mathrm{Re}\left((\hat{\Upsilon}_X^{-1}\mathbf{u})^\top\gamma(\omega_k)\varphi_{\omega_k}\right)\right](x')\Bigg\|\\ &= \left\|\frac{1}{m}\sum_{j=1}^m \mathrm{Re}\left(\left((\hat{\Upsilon}_X^{-1}\mathbf{u})^\top\gamma(\omega_k)\right)\cdot\left(\mathrm{D}_r\left[\varphi_{\omega_k}\right](x) - \mathrm{D}_r\left[\varphi_{\omega_k}\right](x')\right)\right)\right\|\\ &\leqslant \left\|\hat{\Upsilon}_X^{-1}\right\|\|\mathbf{u}\|\sqrt{s}\bar{L}_{01}\left\|\mathrm{D}_r\left[\varphi_{\omega_k}\right](x) - \mathrm{D}_r\left[\varphi_{\omega_k}\right](x')\right\| \qquad \text{(D.9)}\\ &\leqslant 4s\bar{L}_{01}d_{\mathbf{H}}(x,x')\mathcal{L}_r \leqslant c\varepsilon_r. \qquad \text{(D.10)}\end{aligned}$$

where we have applied Lemma C.2 to obtain the last line.

Choosing $\mathcal{X}_{\mathrm{grid}}^{\mathrm{far}}$ to be a $\delta_0 \overset{\text{def.}}{=} \frac{c\varepsilon_0}{4\mathcal{L}_0\bar{L}_{01}s}$-covering of $\mathcal{X}^{\mathrm{near}}$ (of size at most $\mathcal{O}(R_{\mathcal{X}}/\delta_0)$), $\mathcal{X}_{\mathrm{grid}}^{\mathrm{far}}$ to be a $\delta_2 \overset{\text{def.}}{=} \frac{c\varepsilon_2}{4\mathcal{L}_2\bar{L}_{01}s}$-covering of $\mathcal{X}^{\mathrm{far}}$ (of size at most $\mathcal{O}(R_{\mathcal{X}}/\delta_2)$). Then for any $x \in \mathcal{X}^{\mathrm{near}}$ and $x' \in \mathcal{X}_{\mathrm{grid}}^{\mathrm{near}}$ such that $d_{\mathbf{H}}(x,x') \leqslant \delta_0$,

$$|\hat{\eta}_{X_0}(x)| \leqslant |\hat{\eta}_{X_0}(x')| + |\hat{\eta}_{X_0}(x) - \hat{\eta}_{X_0}(x')| \leqslant 1 - \varepsilon_0 + 2c\varepsilon_0.$$

and given any $x \in \mathcal{X}^{\text{far}}$, let $x' \in \mathcal{X}^{\text{far}}_{\text{grid}}$ be such that $d_{\mathbf{H}}(x, x') \leqslant \delta_2$, so

$$\operatorname{Re}\left(\overline{\operatorname{sign}(a_i)}\mathrm{D}_2\left[\hat{\eta}_{X_0}\right](x)\right) \preceq \operatorname{Re}\left(\overline{\operatorname{sign}(a_i)}\mathrm{D}_2\left[\hat{\eta}_{X_0}\right](x')\right) + \left\|\mathrm{D}_2\left[\hat{\eta}_X\right](x) - \mathrm{D}_2\left[\hat{\eta}_X\right](x')\right\| \mathrm{Id} \preceq (-\varepsilon_2 + 2c\varepsilon_2)\mathrm{Id},$$

and

$$\left\|\operatorname{Im}\left(\overline{\operatorname{sign}(a_i)}\mathrm{D}_2\left[\hat{\eta}_{X_0}\right](x)\right)\right\| \leqslant \left\|\operatorname{Im}\left(\overline{\operatorname{sign}(a_i)}\mathrm{D}_2\left[\hat{\eta}_{X_0}\right](x')\right)\right\| + c\varepsilon_2 \leqslant (c_2 + c)\varepsilon_2.$$

□

D.2 Nondegeneracy transfer to $\hat{\eta}_X$.

We are now ready to prove Theorem 3, which we restate below for clarity.

Theorem D.2. *Under the assumptions of Theorem D.1, the following holds with probability at least $1-\rho$: for all X such that*

$$d_{\mathbf{H}}(X, X_0) \lesssim \min\left(r_{\text{near}}, \varepsilon_r (C_{\mathbf{H}} B\sqrt{s})^{-1}, \varepsilon_r (C_{\mathbf{H}} \bar{L}_{12} \bar{L}_r \sqrt{s})^{-1}\right), \tag{D.11}$$

we have

(i) for all $y \in \mathcal{X}^{far}$, $|\hat{\eta}_X(y)| \leqslant 1 - \frac{13}{32}\varepsilon_0$

(ii) for all $y \in \mathcal{X}^{near}(i)$, $-\operatorname{Re}\left(\overline{\operatorname{sign}(a_i)}\mathrm{D}_2\left[\hat{\eta}_X\right](y)\right) \succcurlyeq \frac{13\varepsilon_2}{32}\mathrm{Id}$ and $\left\|\operatorname{Im}\left(\overline{\operatorname{sign}(a_i)}\mathrm{D}_2\left[\hat{\eta}_X\right](y)\right)\right\| \leqslant (\frac{p}{2} + \frac{3p}{16})\frac{1}{2}\varepsilon_2$.

Hence, $\hat{\eta}_X$ is $(\frac{13}{32}\varepsilon_0, \frac{13}{32}\varepsilon_2)$-nondegenerate.

The proof essentially exploits the fact that $\hat{\Upsilon}_X$, $\hat{\mathbf{f}}_X$ are locally Lipschitz in X with respect to the metric $d_{\mathbf{H}}$, and consequently nondegeneracy of $\hat{\eta}_{X_0}$ implies nondegeneracy of $\hat{\eta}_X$ whenever $d_{\mathbf{H}}(X, X_0)$ is sufficiently small.

D.2.1 Proof of Theorem D.2

We begin with a lemma which shows that $\hat{\Upsilon}_X$ is locally Lipschitz in X.

Lemma D.1 (Lipschitz bound of $\hat{\Upsilon}_X$)**.** *Let $X_0 \in \mathcal{X}^s$ be Δ-separated points. Assume that for all $i + j \leqslant 3$*

$$\mathbb{P}(E_\omega^c) \leqslant \frac{1}{1 + 16\sqrt{s}B_{ij}}, \quad \mathbb{E}[L_i(\omega)L_j(\omega)1_{E_\omega^c}] \leqslant \frac{1}{16\sqrt{s}}$$

for all $i, j = 0, ..., 2$. Let $\rho > 0$ and

$$m \gtrsim s(\bar{L}_2^2 B_{11} + \bar{L}_1^2 B_{22} + \bar{L}_{01}\bar{L}_2)\left(\log\left(\frac{sd}{\rho}\right) + d\log\left(sC_{\mathbf{H}} \max_{i=0}^{3} \bar{L}_i\right)\right)$$

Then, conditional on event $\bar{E}$, with probability at least $1-\rho$, the following hold:

- *(i) for all X such that $d_{\mathbf{H}}(x_i, x_{0,i}) \leqslant r_{\text{near}}$, we have*

$$\left\|\hat{\Upsilon}_X - \hat{\Upsilon}_{X_0}\right\| \lesssim C_{\mathbf{H}} B d_{\mathbf{H}}(X, X_0).$$

- *(ii) for all X such that $d_{\mathbf{H}}(X, X_0) \lesssim \min\left(r_{\text{near}}, \frac{1}{C_{\mathbf{H}}B}\right)$, we have $\left\|\mathrm{Id} - \hat{\Upsilon}_X\right\| \leqslant \frac{3}{4}$ and $\left\|\mathbf{G}_X^{-\frac{1}{2}}\Gamma_X^*\right\| \lesssim 1$.*

Proof. By Lemma C.8 and Lemma C.10, with probability at least $1-\rho$ conditonal on $\bar{E}$, for all $(i,j) \in \{(0,0),(0,1),(1,1),(1,2)\}$ and all $x, y \in \mathcal{X}^{\text{near}}$,

$$\left\|\hat{K}^{(ij)}(x,y)\right\| \leqslant \left\|K^{(ij)}(x,y)\right\| + \frac{1}{\sqrt{s}},$$

note that this also holds for $\hat{K}^{(ji)}(x,y)$ since $\hat{K}^{(ij)}(x,y) = \overline{\hat{K}^{(ij)}(y,x)}$.

In particular, for all x, x' such that $d_{\mathbf{H}}(x,x') \geqslant \Delta/4$, we have $\left\|\hat{K}^{(ij)}(x,x')\right\| \leqslant \frac{2}{\sqrt{s}}$. Take any X such that $d_{\mathbf{H}}(x_i, x_{0,i}) \leqslant r_{\mathrm{near}}$, we have that both $x_i, x_{0,i}$ are at least $\Delta/4$-separated from x_j and $x_{0,j}$. Therefore, for $k, \ell \in \{0,1\}$, using Lemma C.3:

$$\begin{aligned}\left\|\hat{K}^{(k\ell)}(x_i,x_j) - \hat{K}^{(k\ell)}(x_{i,0},x_{j,0})\right\| &\lesssim \frac{C_{\mathbf{H}}}{\sqrt{s}}\sqrt{d_{\mathbf{H}}(x_i,x_{0,i})^2 + d_{\mathbf{H}}(x_j,x_{0,j})^2} \\ \left\|\hat{K}^{(k\ell)}(x_i,x_i) - \hat{K}^{(k\ell)}(x_{i,0},x_{i,0})\right\| &\lesssim C_{\mathbf{H}}\left(B_{k+1,\ell} + B_{k,\ell+1}\right) d_{\mathbf{H}}(x_i,x_{0,i})\end{aligned} \tag{D.12}$$

and therefore by Lemma G.6:

$$\begin{aligned}\left\|\hat{\Upsilon}_X - \hat{\Upsilon}_{X_0}\right\|^2 &\leqslant \sum_{i,j=1}^{s}\sum_{k,\ell=0}^{1}\left\|\hat{K}^{(k\ell)}(x_i,x_j) - \hat{K}^{(k\ell)}(x_{0,i},x_{0,j})\right\|^2 \\ &\leqslant 2\sum_{i,j=1}^{s}\sum_{k,\ell=0}^{1}\left\|\hat{K}^{(k\ell)}(x_i,x_j) - \hat{K}^{(k\ell)}(x_{0,i},x_j)\right\|^2 + \left\|\hat{K}^{(\ell k)}(x_j,x_{0,i}) - \hat{K}^{(\ell k)}(x_{0,j},x_{0,i})\right\|^2 \\ &\lesssim C_{\mathbf{H}}^2\left(\sum_{\substack{k,l\in\{0,1,2\}\\ k+\ell\leqslant 3}} B_{k\ell}\right)^2 \sum_i d_{\mathbf{H}}(x_i,x_{0,i})^2 + \frac{1}{s}\sum_{j\neq i} d_{\mathbf{H}}(x_j,x_{0,j})^2\end{aligned}$$

which yields the desired result.

For the second statement, using Proposition C.1, $\mathbb{P}_{\bar{E}}(\left\|\hat{\Upsilon}_{X_0} - \Upsilon_{X_0}\right\| > \frac{1}{8}) \leqslant \rho$, so conditional on $\bar{E}$, we have with probability $1-\rho$, $\left\|\hat{\Upsilon}_X - \hat{\Upsilon}_{X_0}\right\| \leqslant \frac{1}{8}$ and the claim follows since $\|\mathrm{Id} - \Upsilon_{X_0}\| \leqslant \frac{1}{2}$ (due to Lemma C.1) implies that $\left\|\mathrm{Id} - \hat{\Upsilon}_X\right\| \leqslant \frac{3}{4}$ and

$$\left\|\hat{\Upsilon}_X\right\| \leqslant 7/4 \quad \text{and} \quad \left\|\mathbf{G}_X^{-\frac{1}{2}}\Gamma_X^*\right\| = \sqrt{\left\|\hat{\Upsilon}_X\right\|} \lesssim \sqrt{7}/2.$$

□

Proof of Theorem D.2. Since $\hat{\eta}_{X_0}$ is nondegenerate with probability at least $1-\rho$, the conclusion follows if we prove that for all $x \in \mathcal{X}^{\mathrm{far}}$ and all $y \in \mathcal{X}^{\mathrm{near}}$,

$$\left\|\mathrm{D}_2\left[\hat{\eta}_X - \hat{\eta}_{X_0}\right](x)\right\| \leqslant \varepsilon_0/32 \quad \text{and} \quad \left\|\mathrm{D}_2\left[\hat{\eta}_X - \hat{\eta}_{X_0}\right](y)\right\| \leqslant p\varepsilon_2/32 \tag{D.13}$$

with probability at least $1-\rho$. We first write

$$\hat{\eta}_X(y) - \hat{\eta}_{X_0}(y) = \hat{\alpha}_X^\top(\hat{\mathbf{f}}_X - \hat{\mathbf{f}}_{X_0}) + (\hat{\alpha}_X - \hat{\alpha}_{X_0})^\top\hat{\mathbf{f}}_{X_0}(y).$$

Conditional on $\bar{E}$, with probability at least $1-\rho/2$, we have by Lemma D.1 (note that our assumptions imply the assumptions of Lemma D.1), $\|\Upsilon_X - \Upsilon_{X_0}\| \lesssim C_{\mathbf{H}} B d_{\mathbf{H}}(X, X_0)$ and $\left\|\Upsilon_X^{-1}\right\| \leqslant 4$. So,

$$\left\|\mathrm{D}_r\left[(\hat{\alpha}_X - \hat{\alpha}_{X_0})^\top\hat{\mathbf{f}}_{X_0}\right](y)\right\| \leqslant \sqrt{s}\left\|\Upsilon_X^{-1} - \Upsilon_{X_0}^{-1}\right\| \leqslant 8\sqrt{s}\left\|\hat{\Upsilon}_X - \hat{\Upsilon}_{X_0}\right\| \lesssim \sqrt{s}C_{\mathbf{H}}Bd_{\mathbf{H}}(X,X_0).$$

By Lemma C.2, if $\bar{E}$ occurs, then

$$\left\|\mathrm{D}_r\left[\hat{\alpha}_X^\top(\hat{\mathbf{f}}_X - \hat{\mathbf{f}}_{X_0})\right](y)\right\| \leqslant C_r\left\|\hat{\alpha}_X\right\| d_{\mathbf{H}}(X,X_0) \leqslant C_r\left\|\hat{\Upsilon}_X^{-1}\right\|\sqrt{s}d_{\mathbf{H}}(X,X_0) \leqslant 4C_r\sqrt{s}d_{\mathbf{H}}(X,X_0),$$

where $C_r \lesssim (1 + C_{\mathbf{H}})\bar{L}_r\bar{L}_{12}$. Finally, since $\mathbb{P}(\bar{E}^c) \leqslant \rho/2$, we have with probability at least $1-\rho$, for all $y \in \mathcal{X}$, (D.13) holds provided that (D.11) holds. Combining with the nondegeneracy of $\hat{\eta}_{X_0}$, the conclusion follows with probability $1-2\rho$.

□

E Supplementary results to the proof Theorem 1

Recall that in the proof of Theorem 1, we defined the function $f : \mathbb{C}^s \times \mathcal{X}^s \times \mathbb{R}_+ \times \mathbb{C}^m$ by

$$f(u,v) \stackrel{\text{def.}}{=} \Gamma_X^*(\Phi_X a - \Phi_{X_0} a_0 - w) + \lambda \begin{pmatrix} \operatorname{sign}(a_0) \\ 0_{sd} \end{pmatrix}$$

where $u = (a, X)$ and $v = (\lambda, w)$. This function f is differentiable with

$$\partial_v f(u,v) = \left(\begin{pmatrix} \operatorname{sign}(a_0) \\ 0_{sd} \end{pmatrix},\ -\Gamma_X^* \right) \in \mathbb{C}^{s(d+1)\times m}, \tag{E.1}$$

and $\partial_u f(u,v)$ is

$$\Gamma_X^* \Gamma_X J_a + \begin{pmatrix} 0_{1\times s} & A_{11} & 0 & \cdots & 0 \\ 0_{1\times s} & 0 & A_{12} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{1\times s} & 0 & 0 & \cdots & A_{1s} \\ 0_{d\times s} & A_{21} & 0 & \cdots & 0 \\ 0_{d\times s} & 0 & A_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{d\times s} & 0 & 0 & \cdots & A_{2s} \end{pmatrix} \tag{E.2}$$

where $A_{1j} \stackrel{\text{def.}}{=} \nabla_x \langle \varphi(x_j), z\rangle^\top$, $A_{2j} \stackrel{\text{def.}}{=} \nabla_x^2 \langle \varphi(x_j), z\rangle$, $z \stackrel{\text{def.}}{=} (\Phi_X a - \Phi_{X_0} a_0 - w)$ and $J_a \in \mathbb{R}^{s(d+1)\times s(d+1)}$ is a the diagonal matrix:

$$J_a = \begin{pmatrix} \mathrm{Id}_{s\times s} & & & 0 \\ & a_1 \mathrm{Id}_{d\times d} & & \\ & & \ddots & \\ 0 & & & a_s \mathrm{Id}_{d\times d} \end{pmatrix}.$$

Letting $u_0 = (a_0, X_0)$ and $v_0 = (0,0)$, $\partial_u f(u_0, v_0) = \Gamma_{X_0}^* \Gamma_{X_0} J_a$ is invertible and $f(u_0, v_0) = 0$. Hence, by the Implicit Function Theorem, there exists a neighbourhood V of v_0 in $\mathbb{C} \times \mathbb{C}^m$, a neighbourhood U of u_0 in $\mathbb{C}^s \times \mathcal{X}^s$ and a Fréchet differentiable function $g : V \to U$ such that for all $(u,v) \in U \times V$, $f(u,v) = 0$ if and only if $u = g(v)$. To conclude, we simply need to bound the size of the region on which g is well defined, and to bound the error between $g(v)$ and $g(0)$. Let us first remark that our assumptions imply that $\mathbb{P}(\bar{E}^c) \leqslant \rho/2$ and

$$\mathbb{P}(E_\omega^c) \leqslant \frac{1}{1 + 16\sqrt{s} B_{ij}}, \quad \mathbb{E}[L_i(\omega) L_j(\omega) 1_{E_\omega^c}] \leqslant \frac{1}{16\sqrt{s}}, \tag{E.3}$$

for all $i, j = 0, ..., 2$. Therefore, it is sufficient to prove the existence of g conditional on event $\bar{E}$:

Theorem E.1. *Assume that for all $i + j \leqslant 3$*

$$\mathbb{P}(E_\omega^c) \leqslant \frac{1}{1 + 16\sqrt{s} B_{ij}}, \quad \mathbb{E}[L_i(\omega) L_j(\omega) 1_{E_\omega^c}] \leqslant \frac{1}{16\sqrt{s}}$$

for all $i, j = 0, ..., 2$. Let $\rho > 0$ and suppose that

$$m \gtrsim s(\bar{L}_2^2 B_{11} + \bar{L}_1^2 B_{22} + \bar{L}_{01} \bar{L}_2) \left(\log\left(\frac{sd}{\rho}\right) + d \log\left(s C_{\mathbf{H}} \mathbb{L}_3\right) \right)$$

where $\mathbb{L}_r \stackrel{\text{def.}}{=} \max_{i \leqslant r} L_r$. Then, conditional on event $\bar{E}$, with probability at least $1 - \rho$: there exists a $\mathscr{C}^1$ function g such that, for all $v = (\lambda, w)$ such that $\|v\| \leqslant r$ with r satisfying

$$r = \mathcal{O}\left(\frac{1}{\sqrt{s}} \min\left(\tfrac{\min\{r_{\text{near}}, (C_{\mathbf{H}} B)^{-1}\}}{\min_i |a_{0,i}|},\ \tfrac{1}{\bar{L}_{01} \bar{L}_{12} (1 + \|a_0\|)}, \right) \right) \tag{E.4}$$

we have $f(g(v), v) = 0$ *and* $g(0) = u_0$*. Furthermore, given* (λ, w) *in this ball,* $(a, X) \overset{\text{def.}}{=} g((\lambda, w))$ *satisfies*

$$\|a - a_0\| + d_{\mathbf{H}}(X, X_0) \leqslant \frac{\sqrt{s}(\lambda + \|w\|)}{\min_i |a_{0,i}|}. \tag{E.5}$$

We begin with some preliminary results before presenting the proof of this theorem in Section E.2.

E.1 Preliminary results

Theorem E.2 (Quantitative implicit function theorem, adapted from Denoyelle et al. (2017))**.** *Let* $F : \mathcal{H} \times \mathcal{Y} \to \mathbb{C}^n$ *be a differentiable mapping where* $\mathcal{H}$ *is a Hilbert space,* $\mathcal{Y} \subseteq \mathbb{C}^s \times \mathbb{R}^{sd}$*,* $n = s(d+1)$*,* $\|\cdot\|$ *be a norm on* $\mathcal{H}$*. For each* $y \in \mathcal{Y}$*, suppose that there exists a positive definite matrix* $\mathbf{G}_y$*, and let* d_G *be the associated metric. Assume that* $F(x_0, y_0) = 0$*, and that for* $x \in \mathcal{B}_{\|\cdot\|}(x_0, r_1), y \in \mathcal{B}_{d_G}(y_0, r_2)$*,* $\partial_y F(x, y)$ *is invertible and we have*

$$\left\|\mathbf{G}_y^{-\frac{1}{2}} \partial_x F(x, y)\right\| \leqslant D_1 \quad \textit{and} \quad \left\|\mathbf{G}_y^{\frac{1}{2}} \partial_y F(x, y)^{-1} \mathbf{G}_x^{\frac{1}{2}}\right\| \leqslant D_2\,.$$

Then, defining $R = \min\left(\frac{r_2}{D_1 D_2}, r_1\right)$*, there exists a unique Fréchet differentiable mapping* $g : \mathcal{B}_{\|\cdot\|}(x_0, R) \to \mathcal{B}_{d_G}(y_0, r_2)$ *such that* $g(x_0) = y_0$ *and for all* $x \in \mathcal{B}_{\|\cdot\|}(x_0, R)$*,* $F(x, g(x)) = 0$*, and furthermore*

$$\mathrm{d}g(x) = -(\partial_y F(x, g(x)))^{-1} \partial_x F(x, g(x))$$

and consequently $\left\|\mathbf{G}_{g(x)}^{\frac{1}{2}} \mathrm{d}g(x)\right\| \leqslant D_1 D_2$*.*

Proof. Let $V^* = \cup_{V \in \mathcal{V}} V$, where $\mathcal{V}$ is the collection of all open sets $V \in \mathbb{R}^m$ such that

1. $x_0 \in V$,
2. V is star-shaped with respect to x_0,
3. $V \subset \mathcal{B}_{\|\cdot\|}(x_0, r_1)$,
4. there exists a $\mathcal{C}^1$ function $g : V \to \mathcal{B}_{d_G}(y_0, r_2)$ such that $g(x_0) = y_0$ and $F(x, g(x)) = 0$ for all $x \in V$.

Observe that $\mathcal{V}$ is non-empty by the (classical) Implicit Function Theorem. Moreover, $\mathcal{V}$ is stable by union: indeed, all conditions expect the last one are easy to check. Now, let $V, \tilde{V} \in \mathcal{V}$ and $g, \tilde{g}$ be corresponding functions. The set $\overline{V} = \{x \in V \cap \tilde{V},\ g(x) = \tilde{g}(x)\}$ is non-empty (it contains x_0), and closed in $V \cap \tilde{V}$. Moreover, it is open: for any $x \in \overline{V}$, by our assumptions $\partial_y F(x, g(x))$ is invertible and the Implicit Function theorem applies at $(x, g(x))$, and by the uniqueness of the mapping resulting from it we obtain an open set around x in which g and $\tilde{g}$ coincide. Hence $\overline{V}$ is both closed and open in $V \cap \tilde{V}$, and by the connectedness of it $\overline{V} = V \cap \tilde{V}$. Therefore, there exists a function g' defined on $V \cup \tilde{V}$ that satisfies condition 4. above (it is defined as g on V and $\tilde{g}$ on $\tilde{V}$, which is well-posed for their intersection), and $\mathcal{V}$ is indeed stable by union.

Hence $V^* \in \mathcal{V}$, let g^* be its corresponding function. It is unique by the arguments above, satisfies $F(x, g^*(x)) = 0$ and

$$\begin{aligned}\mathbf{G}_{g^*(x)}^{\frac{1}{2}} \mathrm{d}g^*(x) &= -\mathbf{G}_{g^*(x)}^{\frac{1}{2}} (\partial_y F(x, g^*(x)))^{-1} \partial_x F(x, g^*(x)) \\ &= -(\mathbf{G}_{g^*(x)}^{-\frac{1}{2}} \partial_y F(x, g^*(x)) \mathbf{G}_{g^*(x)}^{-\frac{1}{2}})^{-1} \mathbf{G}_{g^*(x)}^{-\frac{1}{2}} \partial_x F(x, g^*(x))\end{aligned}$$

for all $x \in V^*$. Note that by our assumptions $\left\|\mathbf{G}_{g^*(x)}^{\frac{1}{2}} \mathrm{d}g^*(x)\right\| \leqslant D_1 D_2$.

We finish the proof by showing that V^* contains a ball of radius $r_2/(D_1 D_2)$. Let $x \in \mathbb{R}^m$ with $\|x\| = 1$, $R_x = \sup\{R,\ x_0 + Rx \in V^*\}$, and $x^* = x_0 + R_x x \in \partial V^*$. Clearly $0 < R_x \leqslant r_1$ since V^* is open, assume $R_x < r_1$. Our goal is to show that in that case $R_x \geqslant \frac{r_1}{D_1 D_2}$. Since $\mathrm{d}g^*$ is bounded, g^* is uniformly continuous on V^* and it can be extended on ∂V^*, and by continuity $F(x^*, g^*(x^*)) = 0$. By contradiction, if $g^*(x^*) \in \mathcal{B}_{d_G}(y_0, r_2)$, by our assumptions we can apply the Implicit Function Theorem at $(x^*, g^*(x^*))$, and therefore extend g^* on an

open set V that is not included in V^* such that $V \cup V^* \in \mathcal{V}$, which contradicts the maximality of V^*. Hence $d_G(g^*(x^*), y_0) = r_2$. Let $\gamma : [0,1] \to \mathcal{Y}$ be defined by $\gamma(t) \overset{\text{def.}}{=} g^*(x^* + t(x_0 - x^*))$, so $\gamma'(t) = \mathrm{d}g^*(\gamma(t))(x_0 - x^*)$. Then,

$$\begin{aligned} r_2 = d_G(g^*(x^*), g^*(x_0)) &\leqslant \sqrt{\int_0^1 \langle \mathbf{G}_{g^*(\gamma(t))}\gamma'(t),\, \gamma'(t)\rangle \mathrm{d}t} \\ &= \sqrt{\int_0^1 \left\| \mathbf{G}^{\frac{1}{2}}_{g^*(\gamma(t))} \mathrm{d}g^*(\gamma(t))(x_0 - x^*) \right\|^2 \mathrm{d}t} \leqslant D_1 D_2 R_x. \end{aligned}$$

□

Lemma E.1. *Asssume that event $\bar{E}$ occurs. Then, for all X such that $d_{\mathbf{H}}(x_i, x_{0,i}) \leqslant r_{\text{near}}$,*

$$\|\Pi_X \Gamma_{X_0} a\| \lesssim \begin{cases} \bar{L}_2 \|a\|_1 \max_i d_{\mathbf{H}}(x_i, x_{0,i})^2 \\ \bar{L}_2 \|a\|_\infty d_{\mathbf{H}}(X, X_0)^2 \end{cases}$$

Proof. Recall that $\mathrm{Im}(\Gamma_X) = \{\varphi(x_i), J_\varphi(x_i)\}_i$, and Π_X is a projector on $\mathrm{Im}(\Gamma_X)^\perp$. Also note that for $d_{\mathbf{H}}(x_i, x_{0,i}) \leqslant r_{\text{near}}$, we have $\left\| \mathbf{H}^{-\frac{1}{2}}_{x_{0,i}} \mathbf{H}^{\frac{1}{2}}_{x_i} \right\| \lesssim 1$, and therefore under $\bar{E}$:

$$\left\| \mathbf{H}^{-\frac{1}{2}}_{x_{0,i}} \nabla^2 \varphi_{\omega_j}(x_i) \mathbf{H}^{-\frac{1}{2}}_{x_{0,i}} \right\| \lesssim \left\| \mathrm{D}_2 \left[\varphi_{\omega_j}\right](x_i) \right\| \leqslant \bar{L}_2$$

Let $\gamma_i : [0,1] \to \mathcal{X}$ be any piecewise smooth curve such that $\gamma_i(1) = x_{0,i}$ and $\gamma_i(0) = x_i$. Then, by Taylor expanding $\varphi(\gamma_i(t))$ about $t = 0$, we obtain

$$\varphi(x_{0,i}) = \varphi(x_i) + \langle \nabla\varphi(x_i),\, \gamma_i'(0)\rangle + \int_0^1 \frac{1}{2} \langle \nabla^2 \varphi(\gamma_i(t))\gamma_i'(t),\, \gamma_i'(t)\rangle \mathrm{d}t.$$

Therefore,

$$\Pi_X \Gamma_{X_0} a = \Pi_X \left(\sum_{i=1}^s a_i \varphi(x_{0,i}) \right) = \Pi_X \left(\sum_{i=1}^s \frac{a_i}{2} \int_0^1 \langle \nabla^2 \varphi(\gamma_i(t))\gamma_i'(t),\, \gamma_i'(t)\rangle \mathrm{d}t \right)$$

Taking the norm implies

$$\|\Pi_X \Gamma_{X_0} a\| \leqslant \sum_{i=1}^s \frac{|a_i|}{2} \int_0^1 \bar{L}_2 \left\| \mathbf{H}_{\gamma_i(t)} \gamma_i'(t) \right\|^2 \mathrm{d}t$$

and taking the infimum over all paths γ_i yields

$$\|\Pi_X \Gamma_{X_0} a\| \leqslant \bar{L}_2 \sum_i |a_i|\, d_{\mathbf{H}}(x_i, x_{0,i})^2.$$

□

E.2 Proof of Theorem E.1

Our goal is to apply Theorem E.2. Let $u = (a, X)$, $u_0 = (a_0, X_0)$, $v = (\lambda, w)$ and $v_0 = (0, 0)$. We must control $\left\| \mathbf{G}_X^{-\frac{1}{2}} \partial_v f(u, v) \right\|$ and $\left\| \mathbf{G}_X^{\frac{1}{2}} \partial_u f(u, v)^{-1} \mathbf{G}_X^{\frac{1}{2}} \right\|$ for (u, v) sufficiently close to (u_0, v_0). Using Lemma D.1, conditional on event $\bar{E}$, with probability $1 - \rho$ we have

$$\left\| \mathbf{G}_X^{-\frac{1}{2}} \partial_v f(u, v) \right\| \leqslant \|\mathbf{u}\| + \left\| \mathbf{G}_X^{-\frac{1}{2}} \Gamma_X \right\| \lesssim \sqrt{s}$$

To control $\left\|\mathbf{G}_X^{\frac{1}{2}}\partial_u f(u,v)^{-1}\mathbf{G}_X^{\frac{1}{2}}\right\|$, first observe that

$$\mathbf{G}_X^{-1/2}\partial_u f(u,v)\mathbf{G}_X^{-1/2} = \left(\mathbf{G}_X^{-1/2}\Gamma_X^*\Gamma_X\mathbf{G}_X^{-1/2} + M(u,v)\right) J_a$$

where

$$M(u,v) \overset{\text{def.}}{=} \begin{pmatrix} 0_{1\times s} & \frac{1}{a_1}\left(\mathbf{H}_{x_1}^{-\frac{1}{2}}\nabla[\langle\varphi,z\rangle](x_1)\right)^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0_{1\times s} & 0 & \cdots & \frac{1}{a_s}\left(\mathbf{H}_{x_s}^{-\frac{1}{2}}\nabla[\langle\varphi,z\rangle](x_s)\right)^\top \\ 0_{d\times s} & \frac{1}{a_1}\mathbf{H}_{x_1}^{-\frac{1}{2}}\nabla^2[\langle\varphi,z\rangle](x_1)\mathbf{H}_{x_{0,1}}^{-\frac{1}{2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0_{d\times s} & 0 & \cdots & \frac{1}{a_s}\mathbf{H}_{x_s}^{-\frac{1}{2}}\nabla^2[\langle\varphi,z\rangle](x_s)\mathbf{H}_{x_{0,s}}^{-\frac{1}{2}} \end{pmatrix}, \tag{E.6}$$

where $z = (\Phi_X a - \Phi_{X_0}a_0 - w)$. Now, let us study the invertibility of $\mathbf{G}_X^{-\frac{1}{2}}\Gamma_X^*\Gamma_X\mathbf{G}_X^{-\frac{1}{2}} + M(u,v)$ and bound the norm of its inverse.

Lemma E.2 (Bound on $M(u,v)$)**.** *Let $u=(a,X)$, $v=(\lambda,w)$ and let $M(u,v)$ be as defined in* (E.6)*. Assume that $\bar{E}$ occurs and given $\varepsilon>0$, let $c_\varepsilon \overset{\text{def.}}{=} \frac{\varepsilon\,\min_i|a_{0,i}|}{2\bar{L}_{12}}$. Then, for all $X\in\mathcal{X}^s$, $a\in\mathbb{R}^s$ and $w\in\mathbb{C}^m$ such that*

$$\|a-a_0\| \leqslant \frac{c_\varepsilon}{3\bar{L}_0}, \quad \|w\| \leqslant c_\varepsilon/3 \quad \textit{and} \quad d_{\mathbf{H}}(X,X_0) \leqslant \min\left(r_{\text{near}}, \frac{c_\varepsilon}{3\bar{L}_1\|a_0\|}\right),$$

we have

$$\|M(u,v)\| \leqslant \varepsilon \quad \textit{and} \quad \|M(u,v)\|_{*,\infty} \leqslant \varepsilon$$

Proof. First note that for $r\in\mathbb{N}_0$,

$$\left\|\mathrm{D}_r\left[\varphi^\top z\right](x_i)\right\| \leqslant \frac{1}{\sqrt{m}}\sum_{j=1}^m \left\|z_j\mathrm{D}_r\left[\varphi_{\omega_j}\right](x_i)\right\| \leqslant L_r\|z\|$$

Now, for $\bar{q}=[q_1,\dots,q_s,Q_1,\dots,Q_s]\in\mathbb{C}^{s(d+1)}$, where $q_i\in\mathbb{C}$ and $Q_i\in\mathbb{C}^d$, and $\|\bar{q}\|=1$, we have

$$\begin{aligned}
\|M(u,v)\bar{q}\|^2 &= \sum_{i=1}^s \left|\frac{1}{a_i}\left(\mathbf{H}_{x_i}^{-\frac{1}{2}}\nabla[\varphi^\top z](x_i)\right)^\top Q_i\right|^2 + \left\|\frac{1}{a_i}\mathbf{H}_{x_i}^{-\frac{1}{2}}\nabla^2[\varphi^\top z](x_i)\mathbf{H}_{x_i}^{-\frac{1}{2}}Q_i\right\|^2 \\
&\leqslant \frac{4}{\min_i|a_{0,i}|^2}\|q\|^2\max_i\left(\left\|\mathbf{H}_{x_i}^{-\frac{1}{2}}\nabla[\varphi^\top z](x_i)\right\|^2 + \left\|\mathbf{H}_{x_i}^{-\frac{1}{2}}\nabla^2[\varphi^\top z](x_i)\mathbf{H}_{x_i}^{-\frac{1}{2}}\right\|^2\right) \\
&= \frac{4}{\min_i|a_{0,i}|^2}\max_i\left(\left\|\mathrm{D}_1\left[\varphi^\top z\right](x_i)\right\|^2 + \left\|\mathrm{D}_2\left[\varphi^\top z\right](x_i)\right\|^2\right) \\
&\leqslant \frac{4}{\min_i|a_{0,i}|^2}(\bar{L}_1^2+\bar{L}_2^2)\|z\|^2
\end{aligned}$$

where we have used the fact that $\min_i|a_i| \geqslant \min_i|a_{0,i}|/2$. If $\|\bar{q}\|_{*,\infty}=1$, then

$$\begin{aligned}
\|M(u,v)\bar{q}\|_{*,\infty} &= \max_i\left\{\left|\left(\mathbf{H}_{x_i}^{-\frac{1}{2}}\nabla[\varphi^\top z](x_i)\right)^\top Q_i\right|, \left\|\mathbf{H}_{x_i}^{-\frac{1}{2}}\nabla[\varphi^\top z](x_i)\mathbf{H}_{x_i}^{-\frac{1}{2}}Q_i\right\|^2\right\} \\
&\leqslant \max_i\left\{\left\|\mathbf{H}_{x_i}^{-\frac{1}{2}}\nabla[\varphi^\top z](x_i)\right\|, \left\|\mathbf{H}_{x_i}^{-\frac{1}{2}}\nabla[\varphi^\top z](x_i)\mathbf{H}_{x_i}^{-\frac{1}{2}}\right\|^2\right\}
\end{aligned}$$

and the same bound holds.

Now it remains to bound $\|z\|$. Writing $\varphi(x) \overset{\text{def.}}{=} (\varphi_{\omega_k}(x))_{k=1}^m$, we have

$$\begin{aligned}\|z\| &= \left\| \sum_i (a_i \varphi(x_i) - a_{0,i} \varphi(x_{0,i})) - w \right\| \\ &\leqslant \bar{L}_0 \|a - a_0\| + \|a_0\| \max_k \sqrt{\sum_i |\varphi_{\omega_k}(x_i) - \varphi_{\omega_k}(x_{0,i})|^2} + \|w\| \\ &\leqslant \bar{L}_0 \|a - a_0\| + \|a_0\| \bar{L}_1 d_{\mathbf{H}}(X, X_0) + \|w\|\end{aligned}$$

where the last inequality follows from Lemma C.2.

□

The bound on $\|M(u,v)\|$ from Lemma E.2 allows us to conclude that under event $\bar{E}$, taking

$$c \overset{\text{def.}}{=} \frac{\min_i |a_{0,i}|}{16 \bar{L}_{12}} \tag{E.7}$$

for all $X \in \mathcal{X}^s$, $a \in \mathbb{R}^s$ and $w \in \mathbb{C}^m$ such that

$$\|a - a_0\| \leqslant \frac{c}{3\bar{L}_0}, \quad \|w\| \leqslant c/3 \quad \text{and} \quad d_{\mathbf{H}}(X, X_0) \leqslant \min\left(r_{\text{near}}, \frac{c}{3\bar{L}_1 \|a_0\|} \right),$$

we have $\|M(u,v)\| \leqslant \frac{1}{8}$. Combining this with Lemma D.1 gives

$$\left\| \mathrm{Id} - (\mathbf{G}_X^{-\frac{1}{2}} \Gamma_X^* \Gamma_X \mathbf{G}_X^{-\frac{1}{2}} + M(u,v)) \right\| \leqslant \left\| \mathrm{Id} - \mathbf{G}_X^{-\frac{1}{2}} \Gamma_X^* \Gamma_X \mathbf{G}_X^{-\frac{1}{2}} \right\| + \|M(u,v)\| < \frac{7}{8}$$

and therefore it is invertible and

$$\left\| (\mathbf{G}_X^{-\frac{1}{2}} \Gamma_X^* \Gamma_X \mathbf{G}_X^{-\frac{1}{2}} + M(u,v))^{-1} \right\| \leqslant \frac{1}{1 - \left\| \mathrm{Id} - (\mathbf{G}_X^{-\frac{1}{2}} \Gamma_X^* \Gamma_X \mathbf{G}_X^{-\frac{1}{2}} + M(u,v)) \right\|} = \mathcal{O}(1).$$

In this case, $\partial_u f(u,v)$ is invertible, and we have

$$\left\| (\mathbf{G}_X^{-\frac{1}{2}} \partial_u f(u,v) \mathbf{G}_X^{-\frac{1}{2}})^{-1} \right\| = \left\| J_a^{-1} (\mathbf{G}_X^{-\frac{1}{2}} \Gamma_X^* \Gamma_X \mathbf{G}_X^{-\frac{1}{2}} + M(u,v))^{-1} \right\| \lesssim \frac{1}{\min_i |a_{0,i}|}$$

since $\|a - a_0\| \lesssim \min_i |a_{0,i}|$ by assumption.

Therefore we can apply Theorem E.2 with (recalling the definition of c in (E.7))

$$r_1 = c,\ D_1 = \mathcal{O}(\sqrt{s}),\ r_2 = \mathcal{O}\left(\min\left(r_{\text{near}},\ \tfrac{c}{\bar{L}_1 \|a_0\|}, \tfrac{c}{\bar{L}_0}, \tfrac{1}{C_{\mathbf{H}} B} \right) \right),\ D_2 = \mathcal{O}\left(\tfrac{1}{\min_i |a_{0,i}|} \right)$$

with $B = \sum_{i+j \leqslant 3} B_{ij}$, we obtain that $g(v)$ is defined for $v \in V \overset{\text{def.}}{=} \mathcal{B}_{\|\cdot\|_2}(0, r)$ with

$$r \overset{\text{def.}}{=} \min\left(\tfrac{r_2}{D_1 D_2}, r_1 \right) = \tfrac{r_2}{D_1 D_2} = \mathcal{O}\left(\min\left(\tfrac{r_{\text{near}}}{\sqrt{s} \min_i |a_{0,i}|},\ \tfrac{1}{\sqrt{s} \bar{L}_1 \bar{L}_{12} \|a_0\|}, \tfrac{1}{\sqrt{s} \bar{L}_{12} \bar{L}_0}, \tfrac{1}{\sqrt{s} \min_i |a_{0,i}| C_{\mathbf{H}} B} \right) \right)$$

such that g is $\mathcal{C}^1$, $f(g(v), v) = 0$, $g(v_0) = u_0$, where we recall that $u_0 = (a_0, X_0)$ and $v_0 = (0,0)$.

Finally, from Theorem E.2 we also have that

$$\|\mathbf{G}_X \mathrm{d} g(v)\| \leqslant D_1 D_2 \lesssim \frac{\sqrt{s}}{\min_i |a_{0,i}|}$$

and by defining $\gamma(t) = g(v_0 + t(v - v_0))$ for $t \in [0,1]$, we have the following error bound between $u = g(v)$ and

$u_0 = g(v_0)$:

$$d_G(u,u_0) = \sqrt{\|a-a_0\|_2^2 + d_{\mathbf{H}}(X,X_0)^2} \leqslant \sqrt{\int_0^1 \langle \mathbf{G}_{\gamma(t)}\gamma'(t),\, \gamma'(t)\rangle \mathrm{d}t}$$
$$= \sqrt{\int_0^1 \langle \mathbf{G}_{\gamma(t)} \mathrm{d}g(tv)v,\, \mathrm{d}g(tv)v\rangle \mathrm{d}t}$$
$$\leqslant \frac{\sqrt{s}}{\min_i |a_{0,i}|}\|v\|\,.$$

F Examples

F.1 Fejér kernel

Let $f \in \mathbb{N}$ and $\mathcal{X} \in \mathbb{T}^d$ the d-dimensional torus. We consider the Fejér kernel

$$K(x,x') = \prod_{i=1}^d \kappa(x_i - x_i'),$$

where $\kappa(x) \overset{\text{def.}}{=} \left(\frac{\sin\left(\left(\frac{f}{2}+1\right)\pi x\right)}{\left(\frac{f}{2}+1\right)\sin(\pi x)}\right)^4$, with constant metric tensor

$$\mathbf{H}_x = C_f \mathrm{Id} \quad \text{and} \quad d_{\mathbf{H}}(x,x') = C_f^{-\frac{1}{2}}\|x-x'\|_2\,.$$

where $C_f \overset{\text{def.}}{=} -\kappa''(0) = \frac{\pi^2}{3}f(f+4) \sim f^2$. Note that $K^{(ij)} = C_f^{-(i+j)/2}\nabla_1^i\nabla_2^j K$ and since the metric is constant, we can set $C_{\mathbf{H}} \overset{\text{def.}}{=} 0$.

F.1.1 Discrete Fourier sampling

A random feature expansion associated with the Fejér kernel is obtained by choosing $\Omega = \left\{\omega \in \mathbb{Z}^d\,;\, \|\omega\|_\infty \leqslant f\right\}$, $\varphi_\omega(x) \overset{\text{def.}}{=} e^{\mathrm{i}2\pi\omega^\top x}$, and $\Lambda(\omega) = \prod_{j=1}^d g(\omega_j)$ where $g(j) = \frac{1}{f}\sum_{k=\max(j-f,-f)}^{\min(j+f,f)}(1-|k/f|)(1-|(j-k)/f|)$. Note that this corresponds to sampling *discrete* Fourier frequencies. In this case, the derivatives of the random features are uniformly bounded with $\left\|\nabla^j\varphi_\omega(x)\right\| = \|\omega\|^j = \mathcal{O}(C_f^{j/2}d^{j/2})$. So, we can set $\bar{L}_i = \mathcal{O}(d^{i/2})$.

F.1.2 Admissibility of the kernel

Theorem F.1. *Suppose that $f \geqslant 128$. Then, K is an admissible kernel with $r_{\text{near}} = 1/(8\sqrt{2})$, $\varepsilon_2 = 0.941$, $\varepsilon_0 = 0.00097$, $h = \mathcal{O}(d^{-1/2})$ and $\Delta = \mathcal{O}(d^{1/2}s_{\max}^{1/4})$, $B_{00} = B_{11} = B_{20} = \mathcal{O}(1)$, $B_{01} = \mathcal{O}(d^{1/2})$ and $B_{22} = \mathcal{O}(d)$.*

The remainder of this section is dedicated to proving this theorem. The uniform bounds on B_{ij} are due to Lemma F.4 (uniform bounds), and the bound on Δ and h are due to Lemma F.3. From Lemma F.1, we see that by setting $r_{\text{near}} \overset{\text{def.}}{=} \frac{1}{8\sqrt{2}}$, for all $d_{\mathbf{H}}(x,x') \leqslant r_{\text{near}}$, $K^{(20)}(x,x') \prec -\varepsilon_2\mathrm{Id}$ with $\varepsilon_2 = (1-6r_{\text{near}}^2)(1-r_{\text{near}}^2/(2-r_{\text{near}}^2)-r_{\text{near}}^2) \geqslant 0.941$. Finally, from Lemma F.2, we have that for for all $d_{\mathbf{H}}(x,x') \geqslant r_{\text{near}}$, $|K| \leqslant 1 - 1/(8^3\cdot 2)$, so we can set $\varepsilon_0 \overset{\text{def.}}{=} 0.00097$.

Before proving these lemmas, we first summarise in Section F.1.3 some key properties of the univariate Fejér kernel κ when $f \geqslant 128$ which were derived in Candès and Fernandez-Granda (2014).

For notational convenience, write $t_i \overset{\text{def.}}{=} x_i - x_i'$, $\kappa_i \overset{\text{def.}}{=} \kappa(t_i)$, $\kappa_i' \overset{\text{def.}}{=} \kappa'(t_i)$, and so on. Let

$$K_i \overset{\text{def.}}{=} \prod_{\substack{k=1\\k\neq i}}^d \kappa_k, \quad K_{ij} \overset{\text{def.}}{=} \prod_{\substack{k=1\\k\neq i,j}}^d \kappa_k \quad \text{and} \quad K_{ij\ell} \overset{\text{def.}}{=} \prod_{\substack{k=1\\k\neq i,j,\ell}}^d \kappa_k.$$

With this, we have:

$$\begin{aligned}\partial_{1,i}K(x,x') &= \kappa_i' K_i \\ \partial_{1,i}\partial_{2,i}K(x,x') &= -\kappa_i'' K_i, \quad \text{and} \quad \forall i \neq j,\ \partial_{1,i}\partial_{2,j}K(x,x') = -\kappa_i'\kappa_j' K_{ij}.\end{aligned}$$

Where convenient, we sometimes write $K(t) = K(x-x') \overset{\text{def.}}{=} K(x,x')$.

F.1.3 Properties of κ

From (Candès and Fernandez-Granda, 2014, Equations (2.20)-(2.24) and (2.29)), for all $t \in [-1/2, 1/2]$ and $\ell = 0,1,2,3$:

$$\begin{aligned} 1 - \frac{C_f}{2}t^2 \leqslant \kappa(t) \leqslant 1 - \frac{C_f}{2}t^2 + 8\left(\frac{1+2/f}{1+2/(2+f)}\right)^2 C_f^2 t^4 \leqslant 1 - \frac{C_f}{2}t^2 + 8C_f^2 t^4 \\ |\kappa'(t)| \leqslant C_f t, \quad |\kappa''(t)| \leqslant C_f, \quad |\kappa'''(t)| \leqslant 3\left(\frac{1+2/f}{1+2/(2+f)}\right)^2 C_f^2 t \leqslant 12 C_f^2 t \\ \kappa'' \leqslant -C_f + \frac{3}{2}\left(\frac{1+2/f}{1+2/(2+f)}\right)^2 C_f^2 t^2 \leqslant -C_f + 6C_f^2 t^2. \end{aligned} \tag{F.1}$$

By (Candès and Fernandez-Granda, 2014, Lemma 2.6),

$$\left|\kappa^{(\ell)}(t)\right| \leqslant \begin{cases} \frac{\pi^\ell H_\ell(t)}{(f+2)^{4-\ell}t^4}, & t \in [\frac{1}{2f}, \frac{\sqrt{2}}{\pi}] \\ \frac{\pi^\ell H_\ell^\infty}{(f+2)^{4-\ell}t^4}, & t \in [\frac{\sqrt{2}}{\pi}, \frac{1}{2}), \end{cases}$$

where $H_0^\infty \overset{\text{def.}}{=} 1$, $H_1^\infty \overset{\text{def.}}{=} 4$, $H_2^\infty \overset{\text{def.}}{=} 18$ and $H_3^\infty \overset{\text{def.}}{=} 77$, and $H_\ell(t) \overset{\text{def.}}{=} \alpha^4(t)\beta_\ell(t)$, with

$$\alpha(t) \overset{\text{def.}}{=} \frac{2}{\pi(1-\frac{\pi^2 t^2}{6})}, \quad \bar{\beta}(t) \overset{\text{def.}}{=} \frac{\alpha(t)}{ft} = \frac{2}{ft\pi(1-\pi^2t^2/6)}$$

and $\beta_0(t) \overset{\text{def.}}{=} 1$, $\beta_1(t) \overset{\text{def.}}{=} 2 + 2\bar{\beta}(t)$, $\beta_2 \overset{\text{def.}}{=} 4 + 7\bar{\beta}(t) + 6\bar{\beta}(t)^2$ and $\beta_3(t) \overset{\text{def.}}{=} 8 + 24\bar{\beta} + 30\bar{\beta}(t)^2 + 15\bar{\beta}(t)^3$. Let us first remark that $\bar{\beta}$ is decreasing on $I \overset{\text{def.}}{=} [\frac{1}{2f}, \frac{\sqrt{2}}{\pi}]$, so $\left|\bar{\beta}(t)\right| \leqslant \left|\bar{\beta}(1/(2f))\right| \approx 1.2733$, and $a(t) \leqslant a(\sqrt{2}/\pi) = \frac{3}{\pi}$ on I. Therefore, on I, $H_0(t) \leqslant \frac{3}{\pi}$, $H_1(t) \leqslant 3.79$, $H_2(t) \leqslant 18.83$ and $H_3(t) \leqslant 98.26$, and we can conclude that on $[\frac{1}{2f}, \frac{1}{2})$, we have

$$\left|\kappa^{(\ell)}(t)\right| \leqslant \frac{\pi^\ell \bar{H}_\ell^\infty}{(f+2)^{4-\ell}t^4}$$

where $\bar{H}_0^\infty = 1$, $\bar{H}_1^\infty \overset{\text{def.}}{=} 4$, $\bar{H}_2^\infty \overset{\text{def.}}{=} 19$, $\bar{H}_3^\infty \overset{\text{def.}}{=} 99$. Combining with (F.1), we have $\left\|\kappa^{(\ell)}\right\|_\infty \leqslant \kappa_\ell^\infty$ where $\kappa_0^\infty \overset{\text{def.}}{=} 1$, $\kappa_2^\infty \overset{\text{def.}}{=} C_f$,

$$\begin{aligned} \kappa_1^\infty &\overset{\text{def.}}{=} \sqrt{C_f}\max\left(\frac{2\pi^4}{(\frac{1}{2}+\frac{1}{f})^3}\frac{f}{\sqrt{C_f}}, \frac{\sqrt{C_f}}{2f}\right) = \mathcal{O}(\sqrt{C_f}) \\ \kappa_3^\infty &\overset{\text{def.}}{=} (C_f)^{3/2}\max\left(\frac{99\pi^3}{(\frac{1}{2}+\frac{1}{f})}\left(\frac{2f}{\sqrt{C_f}}\right)^4, \frac{6\sqrt{C_f}}{f}\right) = \mathcal{O}((C_f)^{3/2}). \end{aligned}$$

Finally, given $p \in (0,1)$,

$$(f+2)^4 t^4 \geqslant (1 + p(f+2)^2 t^2)^2, \qquad \forall\, t \geqslant \frac{1}{\sqrt{(1-p)}(f+2)}.$$

Choosing $p = \frac{1}{2}$ and using $(f+2)^2 = (\frac{3}{\pi^2}C_f + 4) \geqslant \frac{3}{\pi^2}C_f$, we have

$$\left|\kappa^{(\ell)}(t)\right| \leqslant \frac{\kappa_\ell^\infty}{(1+\frac{3}{2\pi^2}C_f t^2)^2}, \qquad \forall\, t^2 \geqslant \frac{2\pi^2}{3C_f}, \tag{F.2}$$

F.1.4 Bounds in neighbourhood of $x' = x$

Lemma F.1. *Suppose that $C_f \|t\|_2^2 \leqslant c$ with $c > 0$ such that*

$$\varepsilon \stackrel{\text{def.}}{=} (1-6c)\left(1 - \frac{c}{2-c}\right) - c > 0$$

Then, $\hat{K}^{02}(t) \preceq -\varepsilon \mathrm{Id}$.

Proof. We need to show that $\lambda_{\min}(-K^{(02)}(t)) \geqslant b$. Let $q \in \mathbb{R}^d$, and note that

$$\begin{aligned}
-\langle \nabla_2^2 K q,\, q\rangle &= -\sum_i \left(q_i \kappa_i'' K_i - \kappa_i' \sum_{j\neq i} q_j \kappa_j' K_{ij} \right) q_i \\
&= -\left(\sum_i q_i^2 \kappa_i'' K_i - \sum_i q_i \kappa_i \sum_{j\neq i} q_j \kappa_j K_{ij} \right) \\
&\geqslant \|q\|^2 \left(-\max_i\{\kappa_i'' K_i\} - \sum_j \left|\kappa_j'\right|^2 \right).
\end{aligned} \tag{F.3}$$

We first consider $\kappa_i'' K_i$:

$$\begin{aligned}
&\kappa_i'' \leqslant -C_f + 6C_f^2 t_i^2, \\
&K_i \geqslant \prod_{j\neq i}\left(1 - \frac{C_f}{2}t_i^2\right) \geqslant 1 - \frac{C_f}{2}\|t\|_2^2 - \left(\frac{C_f}{2}\|t\|_2^2\right)^3 - \left(\frac{C_f}{2}\|t\|_2^2\right)^5 - \cdots \\
&\qquad \geqslant 1 - \frac{C_f\|t\|_2^2}{2(1-\frac{C_f}{2}\|t\|_2^2)}.
\end{aligned}$$

and hence,

$$\kappa_i'' K_i \leqslant \left(-C_f + 6C_f^2\|t\|_2^2\right)\left(1 - \frac{C_f\|t\|_2^2}{2(1-\frac{C_f}{2}\|t\|_2^2)}\right)$$

For the second term,

$$\sum_j \left|\kappa_j'\right|^2 \leqslant C_f^2 \|t\|_2^2 .$$

Therefore,

$$\lambda_{\min}(-K^{(02)}(t)) \geqslant \left(1 - 6C_f\|t\|_2^2\right)\left(1 - \frac{C_f\|t\|_2^2}{2(1-\frac{C_f}{2}\|t\|_2^2)}\right) - C_f\|t\|_2^2$$

□

Lemma F.2. *Assume that $\frac{1}{8\sqrt{C_f}} \geqslant \|t\|_2$ Then,*

$$K(t) \leqslant 1 - \frac{C_f}{4}\|t\|_2^2 + 16C_f^2\|t\|_2^4 .$$

Consequently, for all

$$0 < c \leqslant \frac{1}{8\sqrt{2C_f}},$$

and all t such that $\|t\|_2 \geqslant c$,

$$|K(t)| \leqslant 1 - \frac{C_f}{8}c^2.$$

Proof. First note that

$$|\kappa(u)| \leqslant 1 - \frac{C_f}{2}u^2 + 32C_f^2u^4 = 1 - u^2 g(u)$$

where

$$g(u) \overset{\text{def.}}{=} C_f\left(\frac{1}{2} - 32C_f u^2\right),$$

and note that $g(u) \in (0, \frac{C_f}{2})$ for $u \in (0, 1/(8\sqrt{C_f}))$. So, writing $t = (t_i)_{i=1}^d$ and $g_j \overset{\text{def.}}{=} g(t_j)$, we have

$$\begin{aligned} K(t) &= \prod_{j=1}^d \kappa(t_i) \leqslant \prod_{j=1}^d \left(1 - t_j^2 \cdot g(t_j)\right) \\ &= 1 - \sum_{j=1}^d t_j^2 g_j + \sum_{j\neq k} t_j^2 t_k^2 g_j g_k - \sum_{j\neq k\neq \ell} t_j^2 t_k^2 t_\ell^2 g_j g_k g_\ell + \cdots \end{aligned}$$

Note that

$$\begin{aligned} &- \sum_{j\neq k\neq \ell} t_j^2 t_k^2 t_\ell^2 \cdot g_j g_k g_\ell + \sum_{j\neq k\neq \ell\neq n} t_j^2 t_k^2 t_\ell^2 t_n^2 \cdot g_j g_k g_\ell g_n \\ &\leqslant - \sum_{j\neq k\neq \ell} t_j^2 t_k^2 t_\ell^2 \cdot g_j g_k g_\ell + \left(\sum_{j\neq k\neq \ell} t_j^2 t_k^2 t_\ell^2 \cdot g_j g_k g_\ell\right)\left(\sum_n t_n^2 g_n\right) \\ &\leqslant - \sum_{j\neq k\neq \ell} t_j^2 t_k^2 t_\ell^2 \cdot g_j g_k g_\ell \left(1 - \frac{C_f}{2}\|t\|_2^2\right) < 0 \end{aligned}$$

since $\left(1 - \frac{C_f}{2}\|t\|_2^2\right) > 0$. Also,

$$\sum_{j=1}^d t_j^2 g_j \leqslant \frac{C_f}{2}\sum_{j=1}^d t_j^2 < 1,$$

by assumption. So,

$$\begin{aligned} K(t) &\leqslant 1 - \sum_{j=1}^d t_j^2 g_j + \sum_{j\neq k} t_j^2 t_k^2 g_j g_k \\ &\leqslant 1 - \sum_{j=1}^d t_j^2 g_j + \frac{1}{2}\left(\sum_j t_j^2 g_j\right)^2 \leqslant 1 - \frac{1}{2}\sum_{j=1}^d t_j^2 g_j \\ &\leqslant 1 - \frac{C_f}{2}\left(\frac{1}{2}\sum_{j=1}^d t_j^2 - 32C_f\sum_{j=1}^d t_j^4\right) \leqslant 1 - \frac{C_f}{4}\|t\|_2^2 + 16C_f^2\|t\|_2^4. \end{aligned}$$

Finally, observe that the function

$$q(z) \overset{\text{def.}}{=} \frac{C_f}{4}z^2 - 16C_f^2 z^4$$

is positive and increasing on the interval $[0, \frac{1}{8\sqrt{2C_f}}]$. So, for t satisfing

$$c \leqslant \|t\|_2 \leqslant \frac{1}{8\sqrt{2C_f}}, \tag{F.4}$$

we have $|K(t)| \leqslant 1 - q(c) \leqslant 1 - \frac{C_f}{8}c^2$. Finally, since $|K(t)|$ is decreasing as t increases, we trivially have that $|K(t)| \leqslant 1 - q(c)$ for all t with $\|t\|_2 \geqslant c$.

□

F.1.5 Bounds under separation

Lemma F.3. *Let $i, j \in \{0, 1, 2\}$ with $i + j \leqslant 3$. Let $\bar{A} \geqslant \sqrt{\frac{4\pi^2}{3}}$ and $\|t\|_2 \geqslant \bar{A}\sqrt{d}s_{\max}^{1/4}/\sqrt{C_f}$. Then, we have $\left\|K^{(ij)}(t)\right\| \leqslant d^{\frac{i+j-4}{2}}(\bar{A}^4 s_{\max})^{-1}$.*

Proof. Write $t = (t_j)_{j=1}^d$. To bound $K(t) = \prod_{j=1}^d \kappa(a_j)$, we want to make use of the form (F.2). We can do this for each t_j such that $|t_j| \geqslant \sqrt{\frac{2\pi^2}{3C_f}}$. Note that there exists at least one such t_j since $\|t\|_\infty \geqslant \|t\|_2/\sqrt{d} \geqslant \bar{A}s_{\max}^{1/4}/\sqrt{C_f} \geqslant \sqrt{\frac{2\pi^2}{3C_f}}$. If $\{|t_j|\}_{j=1}^k \subset [0, \sqrt{\frac{2\pi^2}{3C_f}})$ for $k \leqslant d-1$, then

$$k\frac{2\pi^2}{3C_f} + \sum_{j=k+1}^{d} t_j^2 \geqslant \|t\|_2^2 \geqslant \frac{\bar{A}^2 d s_{\max}^{1/2}}{C_f},$$

which implies that $\sum_{j=k+1}^d t_j^2 \geqslant \frac{1}{C_f}\left(\bar{A}^2 d s_{\max}^{1/2} - \frac{2\pi^2(d-1)}{3}\right) \geqslant \frac{\bar{A}^2 d s_{\max}^{1/2}}{2C_f}$, by our assumptions on $\bar{A}$. Therefore, we may assume that we have some $d \geqslant p \geqslant 1$ such that $\{b_j\}_{j=1}^p \subseteq \{t_j\}$ with $|b_j| \geqslant \sqrt{\frac{2\pi^2}{3C_f}}$ and $\|b\|_2 \geqslant \frac{\bar{A}\sqrt{d}\sqrt[4]{s_{\max}}}{\sqrt{2C_f}}$. Observe that

$$\prod_{j=1}^{p}(1 + \frac{3C_f}{2\pi^2}b_j^2) \geqslant 1 + \frac{3C_f}{2\pi^2}\sum_{j=1}^{p} b_j^2 = 1 + \frac{3C_f}{2\pi^2}\|b\|_2^2 \geqslant 1 + \frac{3}{4\pi^2}\bar{A}^2 d\sqrt{s_{\max}}.$$

So, by applying the fact that $|\kappa| \leqslant 1$, $\kappa_0^\infty = 1$ and (F.2), we have

$$|K(t)| \leqslant \prod_{j=1}^{p}|\kappa(b_j)| \leqslant \prod_{j=1}^{p}\frac{1}{\left(1 + \frac{3C_f}{2\pi^2}b_j^2\right)^2} \leqslant \frac{1}{\left(1 + \frac{3}{4\pi^2}\bar{A}^2 d\sqrt{s_{\max}}\right)^2}.$$

For $|\kappa_i' K_i|$, if $i \notin \left\{j \ ; \ |t_j| > \sqrt{\frac{2\pi^2}{3C_f}}\right\}$, then

$$|\kappa_i' K_i| \leqslant \|\kappa_i'\|_\infty \prod_{j=1}^{p}|\kappa(b_j)| \leqslant \frac{\|\kappa_i'\|_\infty}{\left(1 + \frac{3}{4\pi^2}\bar{A}^2 d\sqrt{s_{\max}}\right)^2},$$

and otherwise, we have $|\kappa_i' K_i| \leqslant |\kappa'(t_i)|\prod_{j\neq i}|\kappa(b_j)| \leqslant \frac{\kappa_1^\infty}{\left(1+\frac{3}{4\pi^2}\bar{A}^2 d\sqrt{s_{\max}}\right)^2}$, In a similar manner, writing $V \overset{\text{def.}}{=} \left(1 + \frac{3}{4\pi^2}\bar{A}^2 d\sqrt{s_{\max}}\right)^{-2}$, we can deduce that

$$\begin{aligned}
&|\kappa_i' K_i| \leqslant \kappa_1^{\max} V, \qquad |\kappa_i'' K_i| \leqslant \kappa_2^{\max} V, \qquad \left|\kappa_i'\kappa_j' K_{ij}\right|^2 \leqslant (\kappa_1^{\max})^2 V \\
&|\kappa_i''' K_i|^3 \leqslant \kappa_3^{\max} V, \qquad \left|\kappa_i''\kappa_j' K_{ij}\right|^3 \leqslant \kappa_2^{\max}\kappa_1^{\max} V, \qquad \left|\kappa_i'\kappa_j'\kappa_\ell' K_{ij\ell}\right| \leqslant (\kappa_1^{\max})^3 V.
\end{aligned}$$

Therefore,

$$\left\|K^{(10)}\right\| = \frac{1}{\sqrt{C_f}}\|\nabla_1 K\| \leqslant \frac{1}{\sqrt{C_f}}\sqrt{\sum_{j=1}^{d}\left|\kappa_j' K_j\right|^2} \leqslant \frac{\kappa_1^\infty}{\sqrt{C_f}} V\sqrt{d} \lesssim \frac{1}{\bar{A}^4 d^{3/2} s_{\max}}.$$

Using Gershgorin theorem, we have

$$\left\|\nabla_2^2 K(x,x')\right\| \leqslant \max_{1\leqslant i\leqslant d}\{|\kappa_i'' K_i| + |\kappa_i'|\sum_{j\neq i}\left|\kappa_j'\right||K_{ij}|\}$$

and hence,

$$\begin{aligned}\left\|K^{(02)}\right\| = \frac{1}{C_f}\left\|\nabla_2^2 K\right\| &\leqslant \frac{1}{C_f}\overset{d}{\max_{i=1}}\{|\kappa_i'' K_i| + |\kappa_i'|\sum_{j\neq i}\left|\kappa_j' K_{ij}\right|\} \\ &\leqslant \frac{1}{C_f}V\left(\kappa_2^{\max} + (\kappa_1^{\max})^2(d-1)\right) \leqslant \frac{\max\{\kappa_2^\infty, (\kappa_1^\infty)^2\}}{C_f}Vd \lesssim \frac{1}{\bar{A}^4 d s_{\max}}.\end{aligned}$$

Note also that $\left\|K^{(11)}\right\| = \left\|K^{(02)}\right\|$. Finally, since

$$\begin{aligned}\left\|\partial_{1,i}\nabla_2^2 K(x,x')\right\| \leqslant \max\Bigg\{ &|\kappa_i''' K_i| + |\kappa_i''|\sum_{j\neq i}\left|\kappa_j'\right||K_{ij}|, \\ &\max_{j\neq i}\{\left|\kappa_j''\kappa_i' K_{ij}\right| + \left|\kappa_j'\kappa_i'' K_{ij}\right| + |\kappa_i'|\left|\kappa_j'\right|\sum_{l\neq i,j}|\kappa_l'||K_{ij\ell}|\}\Bigg\},\end{aligned}$$

we have

$$\begin{aligned}\left\|K^{(12)}\right\| &= \frac{1}{C_f^{3/2}}\left\|\nabla_1\nabla_2^2 K\right\| \\ &\leqslant \frac{1}{C_f^{3/2}}\sqrt{d}V\max\left(\kappa_3^{\max} + \kappa_2^{\max}\kappa_1^{\max}(d-1), 2\kappa_2^{\max}\kappa_1^\infty + (d-1)(\kappa_1^\infty)^3\right) \\ &\leqslant d^{3/2}\max\{\kappa_3^\infty, \kappa_1^\infty\kappa_2^\infty, (\kappa_1^\infty)^3\}\frac{1}{C_f^{3/2}}V \lesssim \frac{1}{\bar{A}^4 d^{1/2} s_{\max}}\end{aligned}$$

□

F.1.6 Uniform bounds

Lemma F.4. *If $r_{\text{near}} \sim 1/\sqrt{C_f}$, then $B_0 = \mathcal{O}(1)$, $B_{01} = \mathcal{O}(\sqrt{d})$, $B_{02} = B_{12} = B_{11} = \mathcal{O}(1)$ and $B_{22} = \mathcal{O}(d)$.*

Proof. We have $|K| \leqslant 1$, and

$$\|\nabla K\|^2 \leqslant \sum_i |\kappa_i|^2 |K_i|^2 \leqslant d(\kappa_1^\infty)^2 \lesssim C_f d,$$

so $B_{01} = \mathcal{O}(\sqrt{d})$.

From (F.3), for all $\|q\| = 1$,

$$\langle \nabla_2^2 K(t)q,\, q\rangle \leqslant \max_i |\kappa_i''|\,\|q\|_2^2 + \|q\|_2^2\sum_i |\kappa_i|^2 \leqslant C_f + C_f^2\|t\|^2 = \mathcal{O}(C_f),$$

for $\|t\| \lesssim 1/\sqrt{C_f}$. So, since $r_{\text{near}} \leqslant 2/\sqrt{C_f}$, $\left\|K^{02}(t)\right\| \leqslant 2 \overset{\text{def.}}{=} B_{02}$. The norm bound for K^{11} is the same.

$$\begin{aligned}
\left\|K^{(12)}\right\| &= \sup_{\|q\|=\|p\|=1} \frac{1}{C_f^{3/2}} \Bigg(\sum_k \sum_{k\neq i} \partial_{1,i} \left(\partial_{2,k}^2 K p_i q_k^2 + \partial_{1,i}\partial_{2,i}\partial_{2,k} K p_i q_i q_k \right) \\
&\quad + \sum_i \sum_k \sum_j \partial_{1,i}\partial_{2,j}\partial_{2,k} K p_i p_j p_k + \sum_i \sum_{j\neq i} \partial_{1,i}\partial_{2,i}\partial_{2,j} K p_i q_i q_j + \sum_i \partial_{1,i}\partial_{2,j}^2 K p_i q_i^2 \Bigg) \\
&= \sup_{\|q\|=\|p\|=1} \frac{1}{C_f^{3/2}} \Bigg(\sum_k \sum_{k\neq i} \kappa_i' \kappa_k'' K_{ik} p_i q_k^2 + \kappa_i'' \kappa_k' K_{ik} p_i q_i q_k \\
&\quad + \sum_i \sum_k \sum_j \kappa_i' \kappa_k' \kappa_j' K_{ijk} p_i p_j p_k + \sum_i \sum_{j\neq i} \kappa_i'' \kappa_j' K_{ij} p_i q_i q_j + \sum_i \kappa_i' \kappa_j'' K_{ij} p_i q_i^2 \Bigg) \\
&\leqslant \frac{1}{C_f^{3/2}} \left(3 \|\kappa''\|_\infty \sqrt{\sum_i |\kappa_k'|^2} + \left(\sum_i |\kappa_k'|^2 \right)^{3/2} + \|\kappa'\|_\infty \|\kappa''\|_\infty \right) \\
&\leqslant \frac{1}{C_f^{3/2}} \left(3C_f^2 \|t\| + C_f^3 \|t\|^3 + \mathcal{O}(C_f^{3/2}) \right) = \mathcal{O}(1)
\end{aligned}$$

for $\|t\| \leqslant 1/C_f^{1/2}$.

We finally consider $K^{(22)}(x,x)$: for $\|p\| = 1$,

$$\begin{aligned}
\sum_i \sum_k \sum_j \partial_{1,k}\partial_{1,i}\partial_{2,j}\partial_{2,i} K p_j p_k &= \sum_i \sum_{k\neq i} \kappa_i'' \kappa_k'' p_j^2 K_{ik} + \sum_i \sum_{k\neq i} \kappa_i''' \kappa_k' p_i p_k K_{ik} \\
&\quad + \sum_i \sum_k \sum_j \kappa_i'' \kappa_j' \kappa_k' K_{ijk} p_j p_k + \sum_i \sum_j \kappa_i''' \kappa_j' p_j p_i K_{ij} + \sum_i \kappa_i'''' p_i^2 K_i \\
&= \sum_i \sum_{k\neq i} \kappa_i'' \kappa_k'' p_j^2 K_{ik} + \sum_i \kappa_i'''' p_i^2 \\
&= d\mathcal{O}(C_f^2)
\end{aligned}$$

since $\kappa'(0) = \kappa'''(0) = 0$ and $|\kappa''(0)| = \mathcal{O}(C_f)$, $|\kappa''''(0)| = \mathcal{O}(C_f^2)$. So, $B_{22} = \mathcal{O}(d)$. □

F.2 The Gaussian kernel

We consider the Gaussian kernel $K(x,x') = \exp\left(-\frac{1}{2}\|x-x'\|_{\Sigma^{-1}}^2\right)$ in $\mathbb{R}^d$. Note that K is translation invariant, so that $\mathbf{H}_x$ will be constant and equal to $-\nabla^2 K(x,x)$. For simplicity define $t = x - x'$, $\hat{K}_\Sigma(t) = \exp\left(-\frac{1}{2}\|t\|_{\Sigma^{-1}}^2\right)$ and for $u \in \mathbb{R}$, $\kappa(u) = \exp\left(-\frac{1}{2}u^2\right)$. Denote by $\{e_i\}$ the canonical basis of $\mathbb{R}^d$, and by $f_i = \Sigma^{-1}e_i$ the i^{th} row of Σ^{-1}. We have the following:

$$\begin{aligned}
\nabla \hat{K}_\Sigma(t) &= -\Sigma^{-1} t \hat{K}_\Sigma(t) \\
\nabla^2 \hat{K}_\Sigma(t) &= \left(-\Sigma^{-1} + \Sigma^{-1} t t^\top \Sigma^{-1}\right) \hat{K}_\Sigma(t) \\
\partial_{1,i}\nabla^2 \hat{K}_\Sigma(t) &= \left(\Sigma^{-1} t f_i^\top + f_i t^\top \Sigma^{-1} - (-\Sigma^{-1} + \Sigma^{-1} t t^\top \Sigma^{-1})(t^\top f_i)\right) \hat{K}_\Sigma(t)
\end{aligned}$$

Hence we have $\mathbf{H}_x = -\nabla^2 \hat{K}_\Sigma(0) = \Sigma^{-1}$, and, defining $d_{\mathbf{H}}(x,x') = \|x-x'\|_{\Sigma^{-1}} = \left\|\Sigma^{-\frac{1}{2}}(x-x')\right\|$, we have $C_{\hat{K}} = 1, C_{\mathbf{H}} = 0$ (that is, the metric tensor of the kernel is constant, and $d_{\mathbf{H}}$ is defined as the corresponding normalized norm).

Then, we have

$$\left\|K^{(10)}(x,x')\right\| = \left\|K^{(01)}(x,x')\right\| = d_{\mathbf{H}}(x,x')\kappa(d_{\mathbf{H}}(x,x'))$$
$$\left\|K^{(02)}(x,x')\right\| = \left\|K^{(11)}(x,x')\right\| \leqslant (d_{\mathbf{H}}(x,x')^2+1)\kappa(d_{\mathbf{H}}(x,x'))$$
$$K^{(02)}(x,x') \preccurlyeq (d_{\mathbf{H}}(x,x')^2-1)\kappa(d_{\mathbf{H}}(x,x'))\mathrm{Id}$$

and for $q \in \mathbb{R}^d$ with $\|q\| = 1$, since

$$\sum_i (\Sigma^{\frac{1}{2}}\nabla\varphi_\omega)_i q_i = \nabla\varphi_\omega^\top(\Sigma^{\frac{1}{2}}q) = \sum_i \partial_i\varphi_\omega(q^\top\Sigma^{\frac{1}{2}}e_i)$$

we can write

$$K^{(12)}(x,x')q = \sum_{i=1}^d (q^\top\Sigma^{\frac{1}{2}}e_i)\Sigma^{\frac{1}{2}}\partial_{1,i}\nabla^2\hat{K}_\Sigma(t)\Sigma^{\frac{1}{2}}$$

Thus we examine each term in $\partial_{1,i}\nabla^2\hat{K}_\Sigma$. We have

$$\sum_i (q^\top\Sigma^{\frac{1}{2}}e_i)\Sigma^{\frac{1}{2}}\Sigma^{-1}tf_i^\top\Sigma^{\frac{1}{2}} = \Sigma^{-\frac{1}{2}}t\left(\sum_i q^\top\Sigma^{\frac{1}{2}}e_ie_i^\top\Sigma^{-\frac{1}{2}}\right) = \Sigma^{-\frac{1}{2}}tq^\top$$

and similarly $\sum_i (q^\top\Sigma^{\frac{1}{2}}e_i)\Sigma^{\frac{1}{2}}f_it^\top\Sigma^{-1}\Sigma^{\frac{1}{2}} = qt^\top\Sigma^{\frac{1}{2}}$. Then

$$\sum_i (q^\top\Sigma^{\frac{1}{2}}e_i)(t^\top\Sigma^{-1}e_i)\Sigma^{\frac{1}{2}}\Sigma^{-1}\Sigma^{\frac{1}{2}} = t^\top\Sigma^{-1}(\sum_i e_ie_i^\top)\Sigma^{\frac{1}{2}}q = (t^\top\Sigma^{\frac{1}{2}}q)\mathrm{Id}$$

and similarly $\sum_i\sum_i (q^\top\Sigma^{\frac{1}{2}}e_i)(t^\top\Sigma^{-1}e_i)\Sigma^{\frac{1}{2}}\Sigma^{-1}tt^\top\Sigma^{-1}\Sigma^{\frac{1}{2}} = (t^\top\Sigma^{\frac{1}{2}}q)\Sigma^{-\frac{1}{2}}tt^\top\Sigma^{-\frac{1}{2}}$.
Hence at the end of the day

$$\left\|K^{(12)}(x,x')\right\| \leqslant (3d_{\mathbf{H}}(x,x') + d_{\mathbf{H}}(x,x')^3)\kappa(d_{\mathbf{H}}(x,x'))$$

and this bound is automatically valid for $K^{(21)}$ as well.

Finally, note that

$$\left\|K^{(22)}(x,x)\right\| = \sup_{\|p\|\leqslant 1} \langle \Sigma^{1/2}\nabla_2\nabla_2 \cdot \left(\Sigma^{1/2}K^{(2,0)}(x,x)p\right), p\rangle$$

where $\nabla_2\cdot$ is the divergence operator on the 2nd variable, and one can show that $\left\|K^{(22)}(x,x)\right\| = (d+1)$.

We are then going to use the fact that for any $q \geqslant 1$ the function $f(r) = r^q e^{-\frac{1}{2}r^2}$ defined on $\mathbb{R}_+$ is increasing on $[0,\sqrt{q}]$ and decreasing after, and its maximum value is $f(\sqrt{q}) = \left(\frac{q}{e}\right)^{q/2}$. Furthermore, it is easy to see that we have $f(r) = r^q e^{-r^2/2} \leqslant \left(\frac{2q}{2}\right)^{\frac{q}{2}} e^{-r^2/4}$ and therefore $f(r) \leqslant \varepsilon$ if $r \geqslant 2\left(\log\left(\frac{1}{\varepsilon}\right) + \frac{q}{2}\log\left(\frac{2q}{e}\right)\right)$.
We define $r_{\text{near}} = 1/\sqrt{2}$ and $\Delta = C_1\sqrt{\log(s_{\max})} + C_2$ for some C_1 and C_2.

1. *Global Bounds.* From what preceeds, we have

$$\left\|K^{(10)}\right\| \leqslant \frac{1}{\sqrt{e}}, \quad \left\|K^{(02)}\right\| \leqslant \frac{2}{e} + 1, \quad \left\|K^{(12)}\right\| \leqslant \frac{3}{\sqrt{e}} + \left(\frac{3}{e}\right)^{\frac{3}{2}}$$

   and note that $\left\|K^{(11)}\right\| = \left\|K^{(02)}\right\|$, so for all $i+j \leqslant 3$ $B_{ij} = \mathcal{O}(1)$.

2. *Near* 0 For $d_{\mathbf{H}}(x,x') \leqslant r_{\text{near}}$, we have

$$K^{(02)} \preccurlyeq -\frac{e^{-\frac{1}{4}}}{2}\mathrm{Id}$$

and for $d_{\mathbf{H}}(x,x') \geqslant \frac{1}{2}$,

$$|K| \leqslant e^{-\frac{1}{4}} = 1 - (1 - e^{-\frac{1}{4}})$$

and $\left\|K^{(22)}(x,x)\right\| = d+1$, so we have also $\varepsilon_i = \mathcal{O}(1)$, so $B_i = B_{0i} + B_{1i} + 1 = \mathcal{O}(1)$ and $B_{22} = d+1$.

3. *Separation.* Since $\varepsilon_i = \mathcal{O}(1)$ and $B_{ij} = \mathcal{O}(1)$, every condition $\left\|K^{(ij)}\right\| \lesssim \frac{1}{s_{\max}}$ is satisfied if $\Delta \geqslant C_1\sqrt{\log(s_{\max})} + C_2$ for some constant C_1 and C_2.

F.2.1 Fourier measurements with Gaussian frequencies

The random feature expansion for K is $\varphi_\omega(x) = e^{i\omega^\top x}$ and $\Lambda = \mathcal{N}(0, \Sigma^{-1})$. We have immediately $L_0 = 1$. For $j \geqslant 1$, we have $\mathrm{D}_j[\varphi_\omega](x)[q_1,\dots,q_j] = \left(\prod_i \omega^\top(\Sigma^{\frac{1}{2}}q_i)\right)\varphi_\omega(x)$ and therefore

$$\|\mathrm{D}_j[\varphi_\omega]\| \leqslant \|\omega\|_\Sigma^j$$

Now, we use $\|\omega\|_\Sigma^j = (\left\|\Sigma^{\frac{1}{2}}\omega\right\|^2)^{\frac{j}{2}} = W^{\frac{j}{2}}$ where W is a χ^2 variable with d degrees of freedom. Then, we use the following Chernoff bound (Dasgupta and Gupta, 2003): for $x \geqslant d$, we have

$$\mathbb{P}(W \geqslant x) \leqslant \left(\frac{ex}{d}e^{-\frac{x}{d}}\right)^{\frac{d}{2}} \leqslant \left(e\left(\sqrt{\frac{x}{d}}\right)^2 e^{-\frac{1}{2}\cdot\left(\sqrt{\frac{x}{d}}\right)^2} e^{-\frac{x}{2d}}\right)^{\frac{d}{2}} \leqslant 2^{\frac{d}{2}}e^{-\frac{x}{4}}$$

by using $x^2 e^{-\frac{x^2}{2}} \leqslant \frac{2}{e}$.

Hence we can define the F_j such that, for all $t \geqslant d^{j/2}$, $\mathbb{P}(L_j(\omega) \geqslant t) \leqslant F_j(t) = 2^{\frac{d}{2}}\exp\left(-\frac{t^{\frac{2}{j}}}{4}\right)$, and $F_j(\bar{L}_j)$ is smaller than some δ if $\bar{L}_j \propto \left(d + \log\frac{1}{\delta}\right)^{\frac{j}{2}}$. Then we must choose the L_j such that $\int_{\bar{L}_j} tF_j(t)\mathrm{d}t$ is bounded by some δ. Taking $L_j \geqslant d^{j/2}$ in any case, we have

$$\begin{aligned}
\int_{\bar{L}_j} tF_j(t)\mathrm{d}t &= 2^{\frac{d}{2}}\int_{\bar{L}_j} t\exp\left(-\frac{t^{\frac{2}{j}}}{4}\right)\mathrm{d}t = 2^{\frac{d}{2}}\int_{\bar{L}_j^{\frac{2}{j}}}(j/2)t^{j-1}\exp\left(-\frac{t}{4}\right)\mathrm{d}t \\
&= 2^{\frac{d}{2}}(j/2)\int_{\bar{L}_j^{\frac{2}{j}}}\left(t^{j-1}\exp\left(-\frac{t}{8}\right)\right)\exp\left(-\frac{t}{8}\right)\mathrm{d}t \leqslant 2^{\frac{d}{2}}(j/2)\left(\frac{8(j-1)}{e}\right)^{j-1}\int_{\bar{L}_j^{\frac{2}{j}}}\exp\left(-\frac{t}{8}\right)\mathrm{d}t \\
&= 2^{\frac{d}{2}}j\left(\frac{8(j-1)}{e}\right)^{j-1} 8\exp\left(-\bar{L}_j^{\frac{2}{j}}/8\right)
\end{aligned}$$

Hence this quantity is bounded by δ if $\bar{L}_j \propto \left(d + \log\left(\frac{1}{\delta}\right)\right)^{\frac{j}{2}}$. Then we have $\bar{L}_j^2 F_i(\bar{L}_i) = \bar{L}_j^2 2^{\frac{d}{2}}\exp\left(-\frac{\bar{L}_i^{\frac{2}{i}}}{4}\right)$ which is also bounded by δ if $\bar{L}_j \propto \left(d + \left(\log\frac{d}{\delta}\right)^2\right)^{\frac{j}{2}}$. At the end of the day, our assumptions are satisfied for

$$\bar{L}_j \propto \left(d + \left(\log\frac{dm}{\rho}\right)^2\right)^{\frac{j}{2}}$$

F.2.2 Gaussian mixture model learning

We apply the mixture model framework with the base distribution:

$$P_\theta = \mathcal{N}(\theta, \Sigma)$$

The random features on the data space are $\varphi'_\omega(x) = Ce^{i\omega^\top x}$ with Gaussian distribution $\omega \sim \Lambda = \mathcal{N}(0, A)$ for some constant C and matrix A. Then, the features on the parameter space are $\varphi_\omega(\theta) = \mathbb{E}_{x\sim P_\theta}\varphi'_\omega(x) = Ce^{i\omega^\top\theta}e^{-\frac{1}{2}\|\omega\|_\Sigma^2}$

(that is, the characteristic function of Gaussians). Then, it is possible to show (Gribonval et al., 2017) that the kernel is

$$K(\theta,\theta') = C^2 \frac{\left|A^{-1}\right|^{\frac{1}{2}}}{\left|2\Sigma + A^{-1}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left\|\theta-\theta'\right\|^2_{(2\Sigma+A^{-1})^{-1}}}$$

Hence we choose $A = c\Sigma^{-1}$, $C = (1+2c)^{\frac{d}{4}}$, and we come back to the previous case $K(\theta,\theta') = e^{-\frac{1}{2}\left\|\theta-\theta'\right\|^2_{\tilde{\Sigma}^{-1}}}$ with covariance $\tilde{\Sigma} = (2+1/c)\Sigma$. Hence $\varepsilon_i = \mathcal{O}(1)$, $B_{ij} = \mathcal{O}(1)$, $d_{\mathbf{H}}(\theta,\theta') = \|\theta-\theta'\|_{\tilde{\Sigma}^{-1}} = \frac{1}{\sqrt{2+1/c}}\|\theta-\theta'\|_{\Sigma^{-1}}$.

Admissible features. Unlike the previous case, the features are directly bounded and Lipschitz. We have

$$|\varphi_\omega(\theta)| \leqslant C \overset{\text{def.}}{=} L_0,$$
$$\|\mathrm{D}_j\left[\varphi_\omega(\theta)\right]\| = C\left\|\tilde{\Sigma}^{\frac{1}{2}}\omega\right\|^j e^{-\frac{\|\omega\|^2_\Sigma}{2}} = C\left(2+1/c\right)^{\frac{j}{2}}\left\|\Sigma^{\frac{1}{2}}\omega\right\|^j e^{-\frac{\|\omega\|^2_\Sigma}{2}} \leqslant C\left(2+1/c\right)^{\frac{j}{2}}\left(\frac{j}{e}\right)^{\frac{j}{2}} \overset{\text{def.}}{=} L_j$$

Hence all constants L_j are in $\mathcal{O}\left(C(2+1/c)^{\frac{j}{2}}\right)$ by choosing $c = \frac{1}{d}$ they are in $\mathcal{O}\left(d^{\frac{j}{2}}\right)$.

F.3 The Laplace transform kernel

Let $\alpha \in \mathbb{R}^d_+$ and let $\mathcal{X} \subset \mathbb{R}^d_+$ be a compact domain. Define for $x \in \mathcal{X}$ and $\omega \in \mathbb{R}^d_+$,

$$\varphi_\omega(x) \overset{\text{def.}}{=} \exp(-\langle x,\,\omega\rangle)\prod_{i=1}^d \sqrt{\frac{(x_i+\alpha_i)}{\alpha_i}} \quad \text{and} \quad \Lambda(\omega) \overset{\text{def.}}{=} \exp(-\langle 2\alpha,\,\omega\rangle)\prod_{i=1}^d (2\alpha_i),$$

The associated kernel is $K(x,x') = \prod_{i=1}^d \kappa(x_i+\alpha_i, x'_i+\alpha_i)$ where κ is the 1D Laplace kernel

$$\kappa(u,v) \overset{\text{def.}}{=} 2\frac{\sqrt{uv}}{(u+v)}.$$

A direct computation shows that $\mathbf{H}_x \in \mathbb{R}^{d\times d}$ is the diagonal matrix with $(h_{x_i+\alpha_i})_{i=1}^d$ where $h_x \overset{\text{def.}}{=} \partial_x\partial_{x'}\kappa(x,x) = (2x)^{-2}$. Note that

$$d_\kappa(s,t) = \int_{\min\{s,t\}}^{\max\{s,t\}} (2x+2\alpha)^{-1}\mathrm{d}x = \left|\log\left(\frac{t+\alpha}{s+\alpha}\right)\right| \tag{F.5}$$

and so, $d_{\mathbf{H}}(x,x') = \sqrt{\sum_{i=1}^d \left|\log\left(\frac{x_i+\alpha_i}{x'_i+\alpha_i}\right)\right|^2}$.

We have the following results concerning the boundedness of $\|\mathrm{D}_j\left[\varphi_\omega\right]\|$ and the admissiblity of K:

Theorem F.2 (Stochastic gradient bounds)**.** *Assume that the α_i's are all distinct. Then, $\bar{L}_0(\omega) \leqslant \bar{L}_0 \overset{\text{def.}}{=} \left(1+\frac{R_{\mathcal{X}}}{\min_i \alpha_i}\right)^d$ and for $j = 1,2,3$,*

$$\mathbb{P}(L_j(\omega) \geqslant t) \leqslant F_j(t) \overset{\text{def.}}{=} \sum_{i=1}^d \beta_i \exp\left(-\alpha_i\left(\frac{1}{2(R_{\mathcal{X}}+\|\alpha\|_\infty)}\left(\frac{t}{\bar{L}_0}\right)^{1/j} - \sqrt{d}\right)\right)$$

and we have that $\sum_i F_j(\bar{L}_j) \leqslant \delta$ and $\bar{L}_j^2\sum_i F_i(\bar{L}_i) + 2\int_{\bar{L}_j}^\infty tF_j(t)\mathrm{d}t \leqslant \delta$ provided that

$$\bar{L}_j \propto \bar{L}_0(R_{\mathcal{X}}+\|\alpha\|_\infty)^j\left(\sqrt{d} + \max_i \frac{1}{\alpha_i}\log\left(\frac{d\beta_i\bar{L}_0(R_{\mathcal{X}}+\|\alpha\|_\infty)}{\delta\alpha_i}\right)\right)^j.$$

where $\beta_i = \prod_{j\neq i}\frac{\alpha_j}{\alpha_j-\alpha_i}$. Note that $\alpha_i \sim d$ implies that $\bar{L}_0 \sim (1+R_{\mathcal{X}}/d)^d \sim e^{R_{\mathcal{X}}}$.

Theorem F.3 (Admissiblity of K)**.** *The Laplace transform kernel K is admissible with $r_{\text{near}} = 0.2$, $C_{\mathbf{H}} = 1.25$,*

$\varepsilon_0 = 0.005$, $\varepsilon_2 = 1.52$. *For all* $i+j \leqslant 3$, $B_{ij} = \mathcal{O}(1)$, $B_{22} = \mathcal{O}(d)$, $\Delta = \mathcal{O}(d + \log\left(d^{3/2} s_{\max}\right))$ *and* $h = \mathcal{O}(1)$.

The first result Theorem F.2 is proved in Section F.3.1 and the second result, Theorem F.4 is a direct consequence of Theorem F.4 and Lemma F.5 in Section F.3.2.

F.3.1 Stochastic gradient bounds

Proof of Theorem F.2. Let $V \stackrel{\text{def.}}{=} (1 - 2(x_i + \alpha_i)\omega_i)_{i=1}^d \in \mathbb{R}^d$. Then,

$$\begin{aligned}\|V\| &= \sqrt{\sum_i (1 - 2(x_i + \alpha_i)\omega_i)^2} \\ &\leqslant \sqrt{\sum_i 1 + 4(x_i + \alpha_i)^2 \omega_i^2} \leqslant \sqrt{d + 4(R_{\mathcal{X}} + \|\alpha\|_\infty)^2 \|w\|^2} \\ &\leqslant \sqrt{d} + 2(R_{\mathcal{X}} + \|\alpha\|_\infty)\|w\|\end{aligned}$$

We have the following bounds:

$$\begin{aligned}|\varphi_\omega(x)| &\leqslant \prod_{i=1}^d \sqrt{1 + \frac{x_i}{\alpha_i}} \leqslant \left(1 + \frac{R_{\mathcal{X}}}{\min_i \alpha_i}\right)^d \stackrel{\text{def.}}{=} \bar{L}_0, \\ \mathrm{D}_1[\varphi_\omega](x) = \varphi_\omega(x) V &\implies \|\mathrm{D}_1[\varphi_\omega](x)\| \leqslant \bar{L}_0 \|V\| \\ \mathrm{D}_2[\varphi_\omega](x) = \varphi_\omega(x)(VV^\top - 2\mathrm{Id}) &\implies \|\mathrm{D}_2[\varphi_\omega](x)\| \leqslant \bar{L}_0 \min\{\|V\|^2, 2\}.\end{aligned}$$

and given $u, q \in \mathbb{R}^d$,

$$\mathrm{D}_3[\varphi_\omega](x)[q,q,u] = \varphi_\omega(x)\left(\langle u, V\rangle\langle q, V\rangle^2 - 2\|q\|^2 - 4\langle u, q\rangle\langle q, V\rangle + 8\sum_i q_i^2 u_i\right),$$

so

$$\|\mathrm{D}_3[\varphi_\omega](x)\| \leqslant |\varphi_\omega(x)|\left(\|V\|^3 + 10 + 4\|V\|\right) \leqslant \bar{L}_0 5(\|V\|^3 + 3),$$

And therefore, in general,

$$\|\mathrm{D}_j[\varphi_\omega](x)\| \leqslant L_j(\omega) \stackrel{\text{def.}}{=} \bar{R}_{\mathcal{X}}^{j+1}\left(\sqrt{d} + \|\omega\|\right)^j$$

$$\|\mathrm{D}_j[\varphi_\omega](x)\| \lesssim L_j(\omega) \stackrel{\text{def.}}{=} \bar{L}_0\left(\sqrt{d} + 2(R_{\mathcal{X}} + \|\alpha\|_\infty)\|w\|\right)^j$$

Assuming for simplicity that all α_j are distinct, we have Akkouchi:

$$\mathbb{P}(\|w\| \geqslant t) \leqslant \mathbb{P}(\|\omega\|_1 \geqslant t) = \sum_{i=1}^d \beta_i e^{-\alpha_i t}$$

where $\beta_i = \prod_{j \neq i} \frac{\alpha_j}{\alpha_j - \alpha_i}$, using the fact that $\|\omega\|_1$ is a sum of independent exponential random variable.

Hence, for all $1 \leqslant j \leqslant 3$ and $t \geqslant d^{\frac{j}{2}}$ we have

$$\begin{aligned}\mathbb{P}(L_j(\omega) \geqslant t) &\leqslant \mathbb{P}\left(\|w\| \geqslant \frac{1}{2(R_{\mathcal{X}} + \|\alpha\|_\infty)}\left(\frac{t}{\bar{L}_0}\right)^{1/j} - \sqrt{d}\right) \\ &\leqslant F_j(t) \stackrel{\text{def.}}{=} \sum_{i=1}^d \beta_i \exp\left(-\alpha_i\left(\frac{1}{2(R_{\mathcal{X}} + \|\alpha\|_\infty)}\left(\frac{t}{\bar{L}_0}\right)^{1/j} - \sqrt{d}\right)\right) \leqslant \delta\end{aligned}$$

and $F_j(\bar{L}_j) \leqslant \delta$ if

$$\bar{L}_j \geqslant \bar{L}_0 \left(2^j (R_{\mathcal{X}} + \|\alpha\|_\infty)^j \left(\sqrt{d} + \max_i \frac{1}{\alpha_i} \log\left(\frac{d\beta_i}{\delta} \right) \right)^j \right)$$

Next, in a similar manner to the Gaussian case, we compute

$$\begin{aligned}
\int_{\bar{L}_j} t F_j(t) \mathrm{d}t &= \sum_{i=1}^d \beta_i \int_{\bar{L}_j} t \exp\left(-\alpha_i \left(\frac{1}{2(R_{\mathcal{X}} + \|\alpha\|_\infty)} \left(\frac{t}{\bar{L}_0} \right)^{1/j} - \sqrt{d} \right) \right) \mathrm{d}t \\
&= \bar{L}_0^2 j \sum_{i=1}^d e^{\alpha_i \sqrt{d}} \beta_i \int_{(\bar{L}_j/\bar{L}_0)^{1/j}} \exp\left(\frac{-\alpha_i u}{2(R_{\mathcal{X}} + \|\alpha\|_\infty)} \right) u^{2j-1} \mathrm{d}u \\
&\leqslant \left(\frac{(2j-1)4(R_{\mathcal{X}} + \|\alpha\|_\infty)}{e\alpha_i} \right)^{2j-1} \bar{L}_0^2 j \sum_{i=1}^d e^{\alpha_i \sqrt{d}} \beta_i \int_{(\bar{L}_j/\bar{L}_0)^{1/j}} \exp\left(\frac{-\alpha_i u}{4(R_{\mathcal{X}} + \|\alpha\|_\infty)} \right) \mathrm{d}u \\
&\leqslant \left(\frac{4(R_{\mathcal{X}} + \|\alpha\|_\infty)}{\alpha_i} \right)^{2j} \left(\frac{2j-1}{e} \right)^{2j-1} \bar{L}_0^2 j \sum_{i=1}^d e^{\alpha_i \sqrt{d}} \beta_i \exp\left(\frac{-\alpha_i (\bar{L}_j/\bar{L}_0)^{1/j}}{4(R_{\mathcal{X}} + \|\alpha\|_\infty)} \right) \leqslant \delta
\end{aligned}$$

if for all $i = 1, \dots, d$,

$$\frac{4(R_{\mathcal{X}} + \|\alpha\|_\infty)}{\alpha_i} \left(2j \log\left(\frac{4(2j-1)(R_{\mathcal{X}} + \|\alpha\|_\infty)}{e\alpha_i} \right) + \log(\bar{L}_0^2 j) + \alpha_i \sqrt{d} + \log\left(\frac{d\beta_i}{\delta} \right) \right) \leqslant \left(\frac{\bar{L}_j}{\bar{L}_0} \right)^{1/j}$$

that is,

$$\bar{L}_j \gtrsim \bar{L}_0 \left(2^j (R_{\mathcal{X}} + \|\alpha\|_\infty)^j \left(\sqrt{d} + \max_i \frac{1}{\alpha_i} \log\left(\frac{d\beta_i}{\delta} \right) \right)^j \right).$$

It remains to bound $\bar{L}_j F_\ell(\bar{L}_\ell)$ with $\ell, j \in \{0,1,2,3\}$: Let $\bar{L}_\ell \geqslant \bar{L}_0 M^\ell$ for some M to be determined. Then,

$$\begin{aligned}
\bar{L}_j F_\ell(\bar{L}_\ell) &\leqslant \bar{L}_0 M^j \sum_{i=1}^d \beta_i \exp\left(\frac{-\alpha_i}{2(R_{\mathcal{X}} + \|\alpha\|_\infty)} M + \alpha_i \sqrt{d} \right) \\
&= \bar{L}_0 \sum_{i=1}^d \beta_i M^j \exp\left(\frac{-\alpha_i}{4(R_{\mathcal{X}} + \|\alpha\|_\infty)} M \right) \exp\left(\frac{-\alpha_i}{4(R_{\mathcal{X}} + \|\alpha\|_\infty)} M \right) e^{\alpha_i \sqrt{d}} \\
&\leqslant \bar{L}_0 e^{-j} \sum_{i=1}^d \left(\frac{4j(R_{\mathcal{X}} + \|\alpha\|_\infty)}{\alpha_i} \right)^j \beta_i \exp\left(\frac{-\alpha_i}{4(R_{\mathcal{X}} + \|\alpha\|_\infty)} M \right) e^{\alpha_i \sqrt{d}} \\
&\leqslant \bar{L}_0 e^{-3} \sum_{i=1}^d \left(\frac{12(R_{\mathcal{X}} + \|\alpha\|_\infty)}{\alpha_i} \right)^3 \beta_i \exp\left(\frac{-\alpha_i}{4(R_{\mathcal{X}} + \|\alpha\|_\infty)} M \right) e^{\alpha_i \sqrt{d}} \leqslant \delta
\end{aligned}$$

if for each $i = 1, \dots, d$

$$M \geqslant 4(R_{\mathcal{X}} + \|\alpha\|_\infty) \left(\sqrt{d} + \max_i \frac{1}{\alpha_i} \log\left(\frac{\bar{L}_0 d\beta_i}{\delta e^3} \left(\frac{12(R_{\mathcal{X}} + \|\alpha\|_\infty)}{\alpha_i} \right)^3 \right) \right).$$

Therefore, similar to the Gaussian case, the conclusion follows for $\bar{L}_0 = \left(1 + \frac{R_{\mathcal{X}}}{\min_i \alpha_i} \right)^d$, and for $j = 1, 2, 3$,

$$\bar{L}_j \propto \bar{L}_0 (R_{\mathcal{X}} + \|\alpha\|_\infty)^j \left(\sqrt{d} + \max_i \frac{1}{\alpha_i} \log\left(\frac{d\beta_i \bar{L}_0 (R_{\mathcal{X}} + \|\alpha\|_\infty)}{\delta \alpha_i} \right) \right)^j.$$

□

F.3.2 Admissiblity of the kernel

Metric variation We have the following lemma on the variation of the Fisher metric:

Lemma F.5. *Suppose that $d_{\mathbf{H}}(x,x') \leqslant c$, then $\left\|\mathrm{Id} - \mathbf{H}_x^{1/2}\mathbf{H}_{x'}\right\| \leqslant (1+ce^c)d_{\mathbf{H}}(x,x')$.*

Proof. Note that $|1-|(x_i+\alpha_i)/(x_i'+\alpha_i)|| \leqslant \max\{e^{d_\kappa(x_i,x_i')}-1, 1-e^{-d_\kappa(x_i,x_i')}\} \leqslant d_\kappa(x_i,x_i')(1+ce^c)$ for all $d_\kappa(x_i,x_i') \leqslant c$. Therefore,

$$\|\mathrm{Id} - \mathbf{H}_x\mathbf{H}_{x'}\|^2 = \sum_i |1-|(x_i+\alpha_i)/(x_i'+\alpha_i)||^2 \leqslant (1+ce^c)d_{\mathbf{H}}(x,x')$$

provided that $d_{\mathbf{H}}(x,x') \leqslant c$.

□

Admissiblity of the kernel The following theorem provides bounds for K and its normalised derivatives.

Theorem F.4. *1.* $|K(x,x')| \leqslant \min\{2^d e^{-\frac{1}{2}d_{\mathbf{H}}(x,x')}, \frac{8}{8+d_{\mathbf{H}}(x,x')^2}\}$.

2. $\left\|K^{(10)}(x,x')\right\| \leqslant \min\{2\sqrt{d}\,|K|\,, \sqrt{2}\}$.

3. $\left\|K^{(11)}\right\| \leqslant \min\{9d\,|K|\,, 8\}$

4. $\left\|K^{(20)}\right\| \leqslant \min\{10d\,|K|\,, 8\}$ *and* $\lambda_{\min}(-K^{(20)}) \geqslant \left(2 - 12d_{\mathbf{H}}(x,x')^2\right)K$.

5. $\left\|K^{(12)}\right\| \leqslant \min\{66\,|K|\,d^{3/2}, 16\sqrt{d}+49\}$ *and* $\left\|K^{(12)}(x,x')\right\| \leqslant 34$ *if* $d_{\mathbf{H}}(x,x') \leqslant 1$.

6. $\left\|K^{(22)}\right\| \leqslant 16d+9$.

In particular, for $d_{\mathbf{H}}(x,x') \geqslant 2d\log(2) + 2\log\left(\frac{52d^{3/2}s_{\max}}{h}\right)$, *we have* $\left\|K^{(ij)}(x,x')\right\| \leqslant \frac{h}{s_{\max}}$.

To prove this result, we first present some bounds for the univariate Laplace kernel in Section F.3.3 before applying these bounds in Section F.3.4.

F.3.3 1D Laplace kernel

In the following $\kappa^{(ij)}(x,x') \overset{\text{def.}}{=} h_x^{-i/2}h_{x'}^{-j/2}\partial_x^i\partial_{x'}^j\kappa(x,x')$.

Lemma F.6. *We have*

(i) $\kappa(x,x') = \operatorname{sech}\left(\frac{d_\kappa(x,x')}{2}\right) \leqslant 2e^{-\frac{1}{2}d_\kappa(x,x')}$,

(ii) $\left|\kappa^{(10)}(x,x')\right| = 2\left|\tanh\left(\frac{d_\kappa(x,x')}{2}\right)\kappa(x,x')\right|$, *and* $\left|\kappa^{(10)}\right| \leqslant 2\,|\kappa|$.

(iii) $\left|\kappa^{(11)}\right| \leqslant 4\,|\kappa|^3 + 4\,|\kappa|$

(iv) $\left|\kappa^{(20)}\right| \leqslant 6\,|\kappa|$ *and* $-\kappa^{(20)} \geqslant 2\kappa(x,x')\left(1 - 2\tanh\left(\frac{d_\kappa(x,x')}{2}\right)\right)$.

(v) $\left|\kappa^{(12)}\right| \leqslant 49\,|\kappa|$.

(vi) $\kappa^{(22)}(x,x) = 9$ *for all* x.

Proof. We first state the partial derivatives of κ:

$$\begin{aligned}
\kappa(x,x') &= \frac{2\sqrt{xx'}}{x+x'},\\
\partial_x\kappa(x,x') &= \frac{x'(x'-x)}{\sqrt{xx'}(x+x')^2}\\
\partial_x\partial_{x'}\kappa(x,x') &= \frac{-x^2+6xx'-(x')^2}{2\sqrt{xx'}(x+x')^3}\\
\partial_x^2\kappa(x,x') &= -\frac{(x')^2\left((x+x')^2+4x(x'-x)\right)}{2\left(xx'\right)^{3/2}(x+x')^3}\\
&= -\frac{(x')^2}{2\left(xx'\right)^{3/2}(x+x')} - \frac{2x'(x'-x)}{\left(xx'\right)^{1/2}(x+x')^3}\\
\partial_x\partial_{x'}^2\kappa(x,x') &= \frac{x^3+13x^2x'-33x(x')^2+3(x')^3)}{4x'(xx')^{1/2}(x+x')^4}\\
\partial_x^2\partial_{x'}^2\kappa(x,x') &= -\frac{3x^4+60x^3x'-270x^2(x')^2+60x(x')^3+3(x')^4}{8xx'(xx')^{1/2}(x+x')^5}
\end{aligned}$$

(i)

$$\kappa(x,x') = 2\left(\sqrt{\frac{x}{x'}}+\sqrt{\frac{x'}{x}}\right)^{-1} = \frac{2}{e^{-\frac{d_\kappa(x,x')}{2}}+e^{\frac{d_\kappa(x,x')}{2}}} = \frac{1}{\cosh(\frac{d_\kappa(x,x')}{2})} \leqslant 2e^{-\frac{1}{2}d_\kappa(x,x')},$$

(ii) We have, assuming that $x > x'$,

$$\begin{aligned}
\kappa^{(10)}(x,x') &= 2x\partial_x\kappa(x,x') = 2\frac{x'-x}{x+x'}\kappa(x,x')\\
&= 2\left(\frac{1}{\frac{x}{x'}+1}-\frac{1}{1+\frac{x'}{x}}\right)\kappa(x,x')\\
&= 2\left(\frac{1}{1+\exp(d_\kappa(x,x'))}-\frac{1}{1+\exp(-d_\kappa(x,x'))}\right)\\
&= 2\left(\frac{\exp(-d_\kappa(x,x'))-\exp(d_\kappa(x,x'))}{2+\exp(d_\kappa(x,x'))+\exp(d_\kappa(x,x'))}\right)\\
&= \frac{-2\sinh(d_\kappa(x,x'))}{1+\cosh(d_\kappa(x,x'))}\kappa(x,x')\\
&= -2\tanh(d_\kappa(x,x')/2)\kappa(x,x'),
\end{aligned}$$

(iii)

$$\begin{aligned}
\kappa^{(11)} &= 4xx'\partial_{x'}\partial_x\kappa(x,x') = 4xx'\frac{4xx'-(x-x')^2}{2\sqrt{xx'}(x+x')^3}\\
&= 4\kappa(x,x')^3 - \frac{4(x-x')^2}{(x+x')^2}\kappa(x,x')\\
&= \kappa(x,x')\left(4\kappa(x,x')^2-4\tanh^2(d_\kappa(x,x')/2)\right)
\end{aligned}$$

so $\left|\kappa^{(11)}\right| \leqslant 4\left|\kappa\right|^3 + 4\left|\kappa\right|$.

(iv)

$$\begin{aligned}\kappa^{(20)} &= 4x^2\partial_x^2\kappa(x,x') = -\frac{4\left(xx'\right)^{1/2}\left((x+x')^2+4x(x'-x)\right)}{2(x+x')^3}\\ &= -2\kappa(x,x')\left(1+\frac{2x(x'-x)}{(x+x')^2}\right)\end{aligned}$$

so $\left|\kappa^{20}\right| \leqslant 6\left|\kappa\right|$. Also,

$$-\kappa^{(20)} \geqslant 2\kappa(x,x')\left(1-2\tanh(d_\kappa(x,x')/2)\right)$$

(v)

$$\begin{aligned}\kappa^{(12)} &= 2x(2x')^2\partial_x\partial_{x'}^2\kappa(x,x')\\ &= \kappa(x,x')\left(1+\frac{2v(5u^2-18uv+v^2)}{(u+v)^3}\right)\end{aligned}$$

so $\left|\kappa^{(12)}\right| \leqslant 49\left|\kappa\right|$.

(vi)

$$\begin{aligned}\kappa^{(22)} &= 16(xx')^2\partial_x^2\partial_{x'}^2\kappa(x,x')\\ &= -3-\frac{48xx'(x^2-6xx'+(x')^2)}{(x+x')^4}\end{aligned}$$

and $\kappa^{(22)}(x,x) = 9$. □

F.3.4 Proof of Theorem F.4

Let $\mathrm{d}_\ell \overset{\text{def.}}{=} d_\kappa(x_\ell+\alpha_\ell, x'_\ell+\alpha_\ell)$ and note that $d_{\mathbf{H}}(x,x') = \sqrt{\sum_\ell \mathrm{d}_\ell^2}$. Define $g = \left(2\tanh(\frac{\mathrm{d}_\ell}{2})\right)_{\ell=1}^d$. We first prove that

(i) $|K(x,x')| \leqslant \prod_{\ell=1}^d \operatorname{sech}(\mathrm{d}_\ell/2) \leqslant \prod_{\ell=1}^d \frac{1}{1+\mathrm{d}_\ell^2/8} \leqslant \frac{1}{1+\frac{1}{8}d_{\mathbf{H}}(x,x')^2}$.

(ii) $\left\|K^{(10)}(x,x')\right\| \leqslant \|g\|_2 \left|K\right|$.

(iii) $\left\|K^{(11)}\right\| \leqslant \left|K\right|\left(\|g\|_2^2+5\right)$

(iv) $\left\|K^{(20)}\right\| \leqslant \left|K\right|\left(\|g\|_2^2+6\right)$ and $\lambda_{\min}\left(K^{(20)}\right) \geqslant K\left(2-3\|g\|_2^2\right)$.

(v) $\left\|K^{(12)}\right\| \leqslant \left|K\right|\left(\|g\|_2^3+16\|g\|_2+49\right)$

(vi) $\left\|K^{(22)}\right\| \leqslant 16d+9$.

The result would then follow because

- $\operatorname{sech}(x) \leqslant 2e^{-x}$ and $\operatorname{sech}(x) \leqslant (1+x^2/2)^{-1}$.
- $|\tanh(x)| \leqslant \min\{x,1\}$, so $\|g\| \leqslant \min\{d_{\mathbf{H}}(x,x'), 2\sqrt{d}\}$,

For example, $\left\|K^{(12)}\right\| \leqslant \frac{1}{1+\frac{1}{8}d_{\mathbf{H}}(x,x')^2}\left(d_{\mathbf{H}}(x,x')^3+16d_{\mathbf{H}}(x,x')+24\right) \leqslant 8d_{\mathbf{H}}(x,x')+\frac{\sqrt{8}}{2}+24 \leqslant 34$ when $d_{\mathbf{H}}(x,x') \leqslant 1$.

In the following, we write $\kappa_\ell^{(ij)} \overset{\text{def.}}{=} \kappa^{(ij)}(x_\ell+\alpha_\ell, x'_\ell+\alpha_\ell)$ and $\kappa_\ell \overset{\text{def.}}{=} \kappa_\ell^{(00)}$ and $K_i \overset{\text{def.}}{=} \prod_{j\neq i}\kappa_j$. Moreover, we will make use of the inequalities for $\kappa^{(ij)}$ derived in Lemma F.6.

(i) We have

$$|K(x,x')| \leqslant \prod_{\ell=1}^{d} \operatorname{sech}(\mathrm{d}_\ell) \leqslant \prod_{\ell=1}^{d} \left(1+\frac{\mathrm{d}_\ell^2}{2}\right)^{-1} \leqslant \frac{1}{1+d_{\mathbf{H}}(x,x')^2}.$$

(ii)

$$K^{(10)}(x,x') = \left(\kappa_\ell^{(10)} K_\ell\right)_{\ell=1}^{d} \implies \left\|K^{(10)}(x,x')\right\| \leqslant \|g\|_2 \, |K| \,.$$

(iii) For $i \neq j$

$$\left|K_{ij}^{(11)}\right| = \left|\kappa_i^{(10)} \kappa_j^{(01)} K_{ij}\right| \leqslant 4 \tanh\left(\frac{\mathrm{d}_i}{2}\right) \tanh\left(\frac{\mathrm{d}_j}{2}\right) |K| \,,$$

and $\left|K_{ii}^{(11)}\right| = \left|\kappa_i^{(11)} K_i\right| \leqslant 5\,|K|$. So, given $p \in \mathbb{R}^d$ of unit norm,

$$\begin{aligned}
\langle K^{(11)} p,\, p\rangle &= \sum_{i=1}^{d} \sum_{j\neq i} \kappa_i^{(10)} \kappa_j^{(01)} K_{ij} p_i p_j + \sum_{i=1}^{d} p_i^2 \kappa_i^{(11)} K_i \\
&\leqslant |K| \left(\sum_{i=1}^{d} \sum_{j\neq i} 4 \tanh(\mathrm{d}_i/2) \tanh(\mathrm{d}_j/2) p_i p_j + 5 \sum_{i=1}^{d} p_i^2 \right) \\
&\leqslant |K| \left(\|g\|_2^2 + 5 \right)
\end{aligned}$$

(iv) For $i \neq j$, $K_{ij}^{(20)} = \kappa_i^{(10)} \kappa_j^{(10)} K_{ij}$, and $\left|K_{ii}^{(20)}\right| = \left|\kappa_i^{(20)} K_i\right| \leqslant 6\,|K|$ and $-K_{ii}^{(20)} \geqslant 2K\left(1 - 2\tanh\left(\frac{\mathrm{d}_i}{2}\right)\right)$.

$$\begin{aligned}
\langle K^{(20)} p,\, p\rangle &= \sum_{i=1}^{d} \sum_{j\neq i} \kappa_i^{(10)} \kappa_j^{(10)} K_{ij} p_i p_j + \sum_{i=1}^{d} p_i^2 \kappa_i^{(20)} K_i \\
&\leqslant |K| \left(\sum_{i=1}^{d} \sum_{j\neq i} 4 \tanh(\mathrm{d}_i/2) \tanh(\mathrm{d}_j/2) p_i p_j + 6 \sum_{i=1}^{d} p_i^2 \right) \\
&\leqslant |K| \left(\|g\|_2^2 + 6 \right),
\end{aligned}$$

and

$$\langle -K^{(20)} p,\, p\rangle \geqslant K \left(2 - 2\,\|g\|_\infty - \|g\|_2^2\right)$$

(v) For i, j, ℓ all distinct,

$$K_{ij\ell}^{(12)} = \kappa_i^{(10)} \kappa_j^{(01)} \kappa_\ell^{(01)} K_{ij\ell} \leqslant 8 \tanh\left(\frac{\mathrm{d}_i}{2}\right) \tanh\left(\frac{\mathrm{d}_j}{2}\right) \tanh\left(\frac{\mathrm{d}_\ell}{2}\right) K,$$

for all i, ℓ,

$$K_{ii\ell}^{(12)} = 8 \kappa_i^{(11)} \kappa_\ell^{(01)} K_{i\ell} \leqslant 10 \tanh\left(\frac{\mathrm{d}_\ell}{2}\right) K$$

$$K_{iji}^{(12)} = \kappa_i^{(11)} \kappa_j^{(01)} K_{ij} \leqslant 10 \tanh\left(\frac{\mathrm{d}_j}{2}\right) K,$$

$K^{(12)}_{ijj} = \kappa^{(10)}_i \kappa^{(02)}_\ell K_{ij} \leqslant 12 \tanh\left(\frac{\mathrm{d}_i}{2}\right) K$, and $K^{(12)}_{iii} = \kappa^{(12)}_i K_i \leqslant 26K$. So, for $p, q \in \mathbb{R}^d$ of unit norm,

$$\begin{aligned}
&\sum_i \sum_j \sum_\ell K^{(12)}_{ij\ell} p_j p_\ell q_i = \sum_i \left(\sum_{j\neq i} \sum_\ell K^{(12)}_{ij\ell} p_j p_\ell q_i + \sum_\ell K^{(12)}_{ii\ell} p_i p_\ell q_i \right) \\
&= \sum_i \sum_{j\neq i} \left(\sum_{\ell \notin \{i,j\}} K^{(12)}_{ij\ell} p_j p_\ell q_i + K^{(12)}_{iji} p_j p_i q_i + K^{(12)}_{ijj} p_j^2 q_i \right) \\
&+ \sum_i \sum_{\ell \neq i} K^{(12)}_{ii\ell} p_i p_\ell q_i + \sum_i K^{(12)}_{iii} p_i^2 q_i \\
&\leqslant |K| \left(\|g\|_2^3 + 16 \|g\|_2 + 49 \right).
\end{aligned}$$

(vi)

$$\begin{aligned}
\left\| K^{(22)}(x,x) \right\| &= \sup_{\|p\|=1} \mathbb{E}[\langle \mathbf{H}_x^{-1/2} \nabla^2 \varphi_\omega(x) \mathbf{H}_x^{-1/2} p,\, \mathbf{H}_x^{-1/2} \nabla^2 \varphi_\omega(x) \mathbf{H}_x^{-1/2} p \rangle] \\
&\leqslant \sup_{\|p\|=1} \sum_i \sum_{k\neq i} \kappa^{(11)}_i \kappa^{(11)}_k p_i^2 + \sum_i \sum_{k\neq i} \kappa^{(12)}_i \kappa^{(10)}_k p_i p_k + \sum_i \sum_{k\neq i} \sum_{j\notin\{i,k\}} \kappa^{(11)}_i \kappa^{(10)}_k \kappa^{(01)}_j p_k p_j \\
&+ \sum_i \sum_{j\neq i} \kappa^{(21)}_i \kappa^{(01)}_j p_j p_i + \sum_i \kappa^{(22)}_i p_i^2 \\
&= \sup_{\|p\|=1} \sum_i \sum_{k\neq i} \kappa^{(11)}_i \kappa^{(11)}_k p_i^2 + \sum_i \kappa^{(22)}_i p_i^2 \\
&\leqslant d \left\| \kappa^{(11)} \right\|_\infty + \left\| \kappa^{(22)} \right\|_\infty \leqslant 16d + \left\| \kappa^{(22)} \right\|_\infty .
\end{aligned}$$

since $\kappa^{(10)}(x,x) = \kappa^{(01)}(x,x) = 0$, and $\kappa^{(11)}(x,x) = 4$ from the proof of (iii) in Lemma F.6.

G Tools

G.1 Probability tools

Lemma G.1 (Bernstein's inequality (Sridharan (2002), Thm. 6)). *Let $x_1, \ldots, x_n \in \mathbb{C}$ be i.i.d. bounded random variables such that $\mathbb{E}x_i = 0$, $|x_i| \leqslant M$ and $Var(x_i) \overset{\text{def.}}{=} \mathbb{E}[|x_i|^2] \leqslant \sigma^2$ for all i's.*

Then for all $t > 0$ we have

$$\mathcal{X}\left(\frac{1}{n} \sum_{i=1}^n x_i \geqslant t \right) \leqslant 4 \exp\left(- \frac{nt^2/4}{\sigma^2 + Mt/(3\sqrt{2})} \right). \tag{G.1}$$

Lemma G.2 (Matrix Bernstein (Tropp (2015), Theorem 6.1.1)). *Let $Y_1, ..., Y_m \in \mathbb{C}^{d_1, d_2}$ be complex random matrices with*

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

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

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

Then

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

Lemma G.3 (Vector Bernstein for complex vectors Minsker (2017)). *Let $Y_1, \ldots, Y_M \in \mathbb{C}^d$ be a sequence of*

independent random vectors such that $\mathbb{E}[Y_i] = 0$, $\|Y_i\|_2 \leqslant K$ *for* $i = 1, \dots, M$ *and set*

$$\sigma^2 \overset{\text{def.}}{=} \sum_{i=1}^{M} \mathbb{E}\,\|Y_i\|_2^2 .$$

Then, for all $t \geqslant (K + \sqrt{K^2 + 36\sigma^2})/M$,

$$\mathbb{P}\left(\left\| \frac{1}{M} \sum_{i=1}^{M} Y_i \right\|_2 \geqslant t \right) \leqslant 28 \exp\left(-\frac{Mt^2/2}{\sigma^2/M + tK/3} \right)$$

Lemma G.4 (Hoeffding's inequality ((Tang et al., 2013), Lemma G.1))**.** *Let the components of* $u \in C^k$ *be drawn i.i.d. from a symmetric distribution on the complex unit circle or* 0*, consider a vector* $w \in \mathbb{C}^k$*. Then, with probability at least* $1 - \rho$*, we have*

$$\mathbb{P}\left(|\langle u,\, w \rangle| \geqslant t \right) \leqslant 4e^{-\frac{t^2}{4\|w\|^2}} \tag{G.3}$$

Lemma G.5. *(Tropp, 2015, Theorem 4.1.1) Let the components of* $u \in \mathbb{R}^k$ *be a Rademacher sequence and let* $Y_1, \dots, Y_M \in \mathbb{C}^{d\times d}$ *be self-adjoint matrices. Set* $\sigma^2 \overset{\text{def.}}{=} \left\| \sum_{\ell=1}^{M} Y_\ell^2 \right\|$*. Then, for* $t > 0$*,*

$$\mathbb{P}\left(\left\| \sum_{\ell=1}^{M} u_\ell Y_\ell \right\| \geqslant t \right) \leqslant 2d \exp\left(-\frac{t^2}{2\sigma^2} \right). \tag{G.4}$$

We were only able to find a reference for this result in the case where u is a Rademacher sequence, however, by the contraction principle (see (Ledoux and Talagrand, 2013, Theorem 4.4)), a similar statement is true for Steinhaus sequences (we write only for the case of real symmetric matrices because this is all we require in this paper, but of course, the same argument extends to complex self-adjoint matrices):

Corollary G.1. *Let the components of* $u \in \mathbb{C}^k$ *i.i.d. from a symmetric distribution on the complex unit circle or* 0 *and let* $B_1, \dots, B_M \in \mathbb{R}^{d\times d}$ *be symmetric matrices. Set* $\sigma^2 \overset{\text{def.}}{=} \left\| \sum_{\ell=1}^{M} B_\ell^2 \right\|$*. Then, for* $t > 0$*,*

$$\mathbb{P}\left(\left\| \sum_{\ell=1}^{M} u_\ell B_\ell \right\| \geqslant t \right) \leqslant 4d \exp\left(-\frac{t^2}{4\sigma^2} \right). \tag{G.5}$$

Proof. By the union bound,

$$\mathbb{P}\left(\left\| \sum_{\ell=1}^{M} u_\ell B_\ell \right\| \geqslant t \right) \leqslant \mathbb{P}\left(\left\| \sum_{\ell=1}^{M} \mathrm{Re}\,(u_\ell) B_\ell \right\| \geqslant \frac{t}{\sqrt{2}} \right) + \mathbb{P}\left(\left\| \sum_{\ell=1}^{M} \mathrm{Im}\,(u_\ell) B_\ell \right\| \geqslant \frac{t}{\sqrt{2}} \right).$$

By the contraction principle (Ledoux and Talagrand, 2013, Theorem 4.4),

$$\mathbb{P}\left(\left\| \sum_{\ell=1}^{M} \mathrm{Re}\,(u_\ell) B_\ell \right\| \geqslant \frac{t}{\sqrt{2}} \right) \leqslant \mathbb{P}\left(\left\| \sum_{\ell=1}^{M} \xi_\ell B_\ell \right\| \geqslant \frac{t}{\sqrt{2}} \right)$$

where ξ is a Rademacher sequence, and the same argument applies to the case of $\mathrm{Im}\,(u_\ell)$. Therefore by Lemma G.5, we have $\mathbb{P}\left(\left\| \sum_{\ell=1}^{M} u_\ell B_\ell \right\| \geqslant t \right) \leqslant 4d \exp\left(-\frac{t^2}{4\sigma^2} \right)$. □

G.2 Linear algebra tools

The following simple lemma will be handy.

Lemma G.6. *For* $1 \leqslant i, j \leqslant s$*, take any scalars* $a_{ij} \in \mathbb{R}$*, vectors* $Q_{ij}, R_{ij} \in \mathbb{R}^d$ *and square matrices* $A_{ij} \in \mathbb{R}^{d\times d}$*.*

1. *Let $M \in \mathbb{R}^{sd\times sd}$ be a matrix formed by blocks :*

$$M = \begin{pmatrix} A_{11} & \dots & A_{1s} \\ \vdots & \ddots & \vdots \\ A_{s1} & \dots & A_{ss} \end{pmatrix}$$

Then we have

$$\|M\|_{\text{block}} = \sup_{\|x\|_{\text{block}}=1} \|Mx\|_{\text{block}} \leqslant \max_{1\leqslant i\leqslant s} \sum_{j=1}^{s} \|A_{ij}\| \tag{G.6}$$

Now, let $P \in \mathbb{R}^{sd\times s}$ be a rectangular matrix formed by stacking vectors $Q_{ij} \in \mathbb{R}^d$:

$$M = \begin{pmatrix} Q_{11} & \dots & Q_{1s} \\ \vdots & \ddots & \vdots \\ Q_{s1} & \dots & Q_{ss} \end{pmatrix}$$

Then,

$$\|M\|_{\infty\to\text{block}} \leqslant \max_{1\leqslant i\leqslant s} \sum_{j=1}^{s} \|Q_{ij}\|_2\,, \quad \left\|M^\top\right\|_{\text{block}\to\infty} \leqslant \max_{1\leqslant i\leqslant s} \sum_{j=1}^{s} \|Q_{ji}\|_2 \tag{G.7}$$

2. *Consider $A \in \mathbb{R}^{s(d+1)\times s(d+1)}$ decomposed as*

$$M = \begin{pmatrix} a_{11} & \dots & a_{1s} & Q_{11}^\top & \dots & Q_{1s}^\top \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{s1} & \dots & a_{ss} & Q_{s1}^\top & \dots & Q_{ss}^\top \\ R_{11} & \dots & R_{1s} & A_{11} & \dots & A_{1s} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ R_{s1} & \dots & R_{ss} & A_{s1} & \dots & A_{ss} \end{pmatrix}$$

Then,

$$\|M\| \leqslant \sqrt{\sum_{i,j} a_{ij}^2 + \|Q_{ij}\|^2 + \|R_{ij}\|^2 + \|A_{ij}\|^2},$$
$$\|M\|_{\text{Block}} \leqslant \max_i \{\sum_j |a_{ij}| + \|Q_{ij}\|,\ \sum_j \|R_{ij}\| + \|A_{ij}\|\}$$

Proof. The proof is simple linear algebra.

1. Let x be a vector with $\|x\|_{\text{block}} \leqslant 1$ decomposed into blocks $x = [x_1, \dots, x_s]$ with $x_i \in \mathbb{R}^d$, we have

$$\|Mx\|_{\text{block}}^2 = \max_{1\leqslant i\leqslant s} \left\| \sum_{j=1}^{s} A_{ij}x_j \right\| \leqslant \max_i \sum_j \|A_{ij}\|\,\|x_j\| \leqslant \max_i \sum_j \|A_{ij}\|$$

2. Similarly,

$$\left\|M^\top x\right\|_\infty = \max_{1\leqslant i\leqslant s} \left\| \sum_{j=1}^{s} Q_{ji}^\top x_j \right\| \leqslant \max_i \sum_j \|Q_{ji}\|\,\|x_j\| \leqslant \max_i \sum_j \|Q_{ji}\|$$

Then, taking $x \in \mathbb{R}^s$ such that $\|x\|_\infty \leqslant 1$, we have

$$\|Mx\|_{\text{block}} = \max_{1\leqslant i\leqslant s} \left\| \sum_{j=1}^{s} x_j Q_{ij} \right\| \leqslant \max_i \sum_j \|Q_{ij}\|$$

3. Taking $x = [x_1, \dots, x_s, X_1, \dots, X_s] \in \mathbb{R}^{s(d+1)}$ with $\|x\| = 1$, we have

$$\begin{aligned}
\|Mx\|^2 &= \sum_{i=1}^{s} \left(\sum_{j=1}^{s} a_{ij} x_j + Q_{ij}^\top X_j \right)^2 + \left\| \sum_{j=1}^{s} R_{ij} x_j + A_{ij} X_j \right\|^2 \\
&\leqslant \sum_{i=1}^{s} \left(\|x\| \sqrt{\sum_{j=1}^{s} a_{ij}^2 + \|Q_{ij}\|^2} \right)^2 + \left(\|x\| \sqrt{\sum_{j=1}^{s} \|R_{ij}\|^2 + \|A_{ij}\|^2} \right)^2 \\
&\leqslant \sum_{i,j} a_{ij}^2 + \|Q_{ij}\|^2 + \|R_{ij}\|^2 + \|A_{ij}\|^2
\end{aligned}$$

Now, if $\|x\|_{\text{Block}} = 1$, we have

$$\begin{aligned}
\|Mx\|_{\text{Block}} &= \max_i \left(\left| \sum_{j=1}^{s} a_{ij} x_j + Q_{ij}^\top X_j \right|, \left\| \sum_{j=1}^{s} R_{ij} x_j + A_{ij} X_j \right\| \right) \\
&\leqslant \max_i \left(\sum_{j=1}^{s} |a_{ij}| + \|Q_{ij}\|, \sum_{j=1}^{s} \|R_{ij} x_j + A_{ij} X_j\| \right)
\end{aligned}$$

□