Revisiting RIP guarantees for sketching operators on mixture models

Ayoub Belhadji, Rémi Gribonval

Revisiting RIP guarantees for sketching operators on mixture models

Ayoub Belhadji Rémi Gribonval
Univ Lyon, ENS de Lyon
Inria, CNRS, UCBL
LIP UMR 5668, Lyon, France

November 25, 2022

Abstract

In the context of sketching for compressive mixture modeling, we revisit existing proofs of the Restricted Isometry Property of sketching operators with respect to certain mixtures models. After examining the shortcomings of existing guarantees, we propose an alternative analysis that circumvents the need to assume importance sampling when drawing random Fourier features to build random sketching operators. Our analysis is based on new deterministic bounds on the restricted isometry constant that depend solely on the set of frequencies used to define the sketching operator; then we leverage these bounds to establish concentration inequalities for random sketching operators that lead to the desired RIP guarantees. Our analysis also opens the door to theoretical guarantees for structured sketching with frequencies associated to fast random linear operators.

Contents

1 Introduction **2**

2 Main tools **5**
2.1 Sketching operator . . . 5
2.2 Separated mixture model, normalized secant set, and dipoles . . . 6
2.3 Kernel coherence . . . 7
2.4 Location-based families and shift-invariant kernels . . . 8
2.5 Random Fourier features . . . 9
2.6 Existing results and their limitations . . . 10

3 Main results **13**
3.1 On the necessity of conditions (43) . . . 13
3.2 Sharp deterministic bounds on $\delta(\mathcal{S}_k|\mathcal{A})$ . . . 14
3.2.1 From the normalized secant set to normalized monopoles and balanced dipoles . . . 14
3.2.2 Expression using the supremum of certain empirical processes . . . 16
3.2.3 Lipschitz property and covering numbers . . . 17
3.3 Results for random sketching . . . 18
3.3.1 Towards bounds for structured random sketching . . . 24
3.4 Lower bounds . . . 25
3.4.1 Lower bounds on variance terms . . . 26

4 Conclusion **29**

A Proofs **32**
A.1 Proof of Lemma 4 . . . 32
A.2 Proof of Theorem 6 . . . 32
A.3 Proof of Proposition 2 . . . 35
A.4 Proof of Proposition 3 . . . 36
A.5 Proof of Proposition 4 . . . 37
A.5.1 Proof of Proposition 9 . . . 40
A.5.2 Proof of Proposition 10 . . . 41
A.6 Proof of Proposition 5 . . . 42
A.7 Proof of Theorem 3 . . . 44
A.7.1 Construction of z_t, $1 \leq t \leq T$ . . . 45
A.7.2 Construction of $f_{\ell,t}$ satisfying (132) . . . 45
A.7.3 Proof of the bound (133) . . . 47
A.7.4 Proof of Proposition 6 . . . 49
A.8 Proof of Theorem 4 . . . 50
A.9 Proof of Theorem 5 and its corollaries . . . 51
A.9.1 Proof of Lemma 2 . . . 52
A.9.2 Some properties of sub-exponential random variables . . . 52
A.9.3 Proof of Theorem 5 . . . 54
A.9.4 Proofs of Corollary 1 and Corollary 2 . . . 55
A.9.5 Some helpful results . . . 58
A.9.6 Proof of Lemma 7 . . . 60

1 Introduction

Building up linear operators that preserve the distances between two sets is at heart of many problems in the field of inverse problems. The fetch for such linear operators gave birth to a rich literature at the intersection of signal processing and machine learning [Ach01, BM01, Sar06, AC06, MM09, MM12, Can08, FR13]. Recently, a new family of inverse problems emerged in the contexte of compressive learning, also called sketched learning [KBGP18, GBKT21a, GBKT21b]. These inverse problems are tailored to be used in the field of mixture modeling. In a nutshell, sketched learning is a paradigm aiming to scale up these learning tasks by conducting the learning task on a low dimensional vector, also called a *sketch*, that contains a "gist" of the initial dataset and that is suited for a specific learning task: *sketching* is the procedure that outputs the sketch for a given dataset. In practice, sketching boils down to embed a probability distribution π typically on $\mathcal{X} = \mathbb{R}^d$ into $\mathbb{C}^m$ by considering a *sketching operator* $\mathcal{A}$ such that[1]

$$\mathcal{A}\pi := \int_{\mathcal{X}} \Phi(x) \mathrm{d}\pi(x) \in \mathbb{C}^m \tag{1}$$

where Φ is a $\mathbb{C}^m$-valued function defined on $\mathcal{X}$ called the *feature map*. As shown in [GBKT21a], building up the linear operator $\mathcal{A}$ is indirectly constrained by the targeted learning task (e.g.: k-means clustering, or Gaussian Mixture Modeling). This constraint can be expressed using the Maximum Mean Discrepancy (MMD) [GBR+12] defined as follows: considering a positive definite kernel[2] [BTA11] κ :

[1] Integrability is treated informally in this introduction and will be more formally discussed in Section 2.1.
[2] In the rest of this paper when we write *kernel* we implicitly assume a positive definite kernel.

$\mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the MMD with respect to κ between two probability distributions π and π' on $\mathcal{X}$ is defined using the norm $\|\cdot\|_\kappa$ on the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated to κ by

$$\|\pi - \pi'\|_\kappa := \sup_{f \in \mathcal{H}, \|f\|_\kappa \leq 1} \left| \int_{\mathcal{X}} f(x) \mathrm{d}\pi(x) - \int_{\mathcal{X}} f(x) \mathrm{d}\pi'(x) \right|. \tag{2}$$

Equiped with the MMD, the general theory of [GBKT21a] suggests to look for sketching operators that satisfy

$$\forall \pi, \pi' \in \mathfrak{G}, \;\; (1-\delta)\|\pi - \pi'\|_\kappa^2 \leq \|\mathcal{A}\pi - \mathcal{A}\pi'\|_2^2 \leq (1+\delta)\|\pi - \pi'\|_\kappa^2, \tag{3}$$

where $\mathfrak{G}$ is a particular set of probability measures on $\mathcal{X}$ and $\delta \in [0, 1)$. When the kernel κ is shift-invariant, it is possible to build up a sketching operator satisfying (3) by considering random Fourier features [ZM15, KBGP18, GBKT21b]. Initially, random Fourier features were introduced to scale up kernel methods [RR07]. This family of approximations is suitable for a shift-invariant kernel κ for which Bochner's theorem [Wen04, Rud17] holds:

$$\forall x, y \in \mathcal{X} = \mathbb{R}^d, \;\; \kappa(x, y) = \int_{\mathbb{R}^d} e^{2\pi \imath \omega (x-y)} \hat{\kappa}(\omega) \mathrm{d}\omega, \tag{4}$$

with $\hat{\kappa}$ a non-negative function. Although the framework of [GBKT21a] holds for more general kernels, the focus of this paper is indeed on shift-invariant kernels, for which we have $\kappa(x, x) = \kappa(0, 0) = \int_{\mathbb{R}^d} \hat{\kappa}(\omega) \mathrm{d}\omega$ for every $x \in \mathcal{X} = \mathbb{R}^d$. Moreover, we often simplify the analysis by assuming a normalized kernel, *i.e.*, $\kappa(0, 0) = 1$. The results are easily extended to the non-normalized case. With the normalization assumption, the function $\hat{\kappa}$ satisfies $\int_{\mathbb{R}^d} \hat{\kappa}(\omega) \mathrm{d}\omega = 1$ and can be interpreted as a probability density function on the *frequency vector* $\omega \in \mathbb{R}^d$. Based on the identity (4), the *random Fourier feature map* is constructed as follows: let $\omega_1, \ldots, \omega_m \in \mathbb{R}^d$ to be i.i.d. random variables with probability density $\hat{\kappa}(\omega)$ with respect to the Lebesgue measure on $\mathbb{R}^d$, and define the random feature map

$$\Phi(x) := \frac{1}{\sqrt{m}} (\phi_{\omega_i}(x))_{i \in [m]} \in \mathbb{C}^m, \tag{5}$$

where here $\phi_\omega(x) = e^{2\pi \imath \omega^\top x} \in \mathbb{C}$ for each $\omega \in \mathbb{R}^d$, and for any integer $n \in \mathbb{N}$ we denote $[n] := \{1, \ldots, n\}$. With this design, the (random) empirical kernel $\kappa_\Phi(x, y) := \langle \Phi(x), \Phi(y) \rangle$ (where for any $u, v \in \mathbb{C}^m$ $\langle u, v \rangle := u^\top \overline{v}$, with $\overline{v}$ the complex conjugate of v) satisfies for every pair of vector $x, y \in \mathbb{R}^d$

$$\mathbb{E}\kappa_\Phi(x, y) = \kappa(x, y). \tag{6}$$

The study of the approximation $\kappa_\Phi(x, y) \approx \kappa(x, y)$ is a well established area of research. Indeed, since the publication of [RR07], many works followed tackling various aspects of this class of approximations. In particular, sharp uniform error bounds on compact sets for RFF were derived in [SS15b, SS15a], and connexions with kernel-based quadrature were established in [Bac17]. Moreover, the quest to various designs of the frequencies were proposed: frequencies based on quasi-Monte Carlo sequences [YSAM14], frequencies based on structured matrices [LSS13, CS16, CRS+18]. We refer the reader to [LHCS21] for an exhaustive review on this topic.

As it was shown in [SGF+10], the identity (6) somehow extends to pairs of probability distributions π, π' on $\mathcal{X}$

$$\mathbb{E}\|\pi - \pi'\|^2_{\kappa_\Phi} = \|\pi - \pi'\|^2_\kappa. \tag{7}$$

This formula was crucial in the study of *characteristic* kernels carried in [SGF+10]. The study of the fluctuations of $\|\pi - \pi'\|^2_{\kappa_\Phi}$ around its expected value $\|\pi - \pi'\|^2_\kappa$ was carried in [ZM15] and [SS15b]. Nevertheless, these results do not imply (3): the established guarantees have an additive form $|\|\pi - \pi'\|^2_{\kappa_\Phi} - \|\pi - \pi'\|^2_\kappa| \leq t$ for a given pair (π, π') of probability distributions, while (3) have a multiplicative form and holds uniformly on $\mathfrak{G} \times \mathfrak{G}$. A study of conditions under which (3) holds was undertaken in [GBKT21b] with a focus on the case where $\mathfrak{G} = \mathfrak{G}_k$ is a set of mixtures with k components that depend on parameters that belong to $\mathbb{R}^d$. In this context, it was shown that (3) holds with high probability provided that:

1. the sketch dimension, m, is large enough: a sufficient sketch size was proved to satisfy $m = \mathcal{O}(k^2 d)$ (up to logarithmic factors); and
2. a variant of the RFF (5) is used, with an appropriate *importance sampling scheme*.

The dependency of this provably good sketch size in k and d does not match the empirical simulations carried out (*without* importance sampling) in [KTTG17], which suggest that (3) can hold with high probability for some $m = \mathcal{O}(kd)$ (again, possibly up to logarithmic factors).

To bridge the gap between theory and practice, one would ideally like to both get rid of the importance sampling assumption, which does not seem needed in practice, and to achieve guarantees for sketch sizes with a better dependency in k. In this work, we revisit the analysis of [GBKT21b] with two main contributions:

- we prove that $m = \mathcal{O}(k^2 d)$ remains a provably good size *even without importance sampling*;
- we explain why the high level structure of the proof technique of [GBKT21b], as well as the structure of our new approach, prevents them from achieving better dependencies in k of provably good sketch sizes.

As we shall see later, our analysis is based on tools inspired from the literature of sparse recovery with incoherent dictionaries ; see Chapter 5 in [FR13]. This quadratic dependency on number k of mixture components, which plays the role of a sparsity level, is thus not surprising. Whether the "right" dependency in k of provably good sketch sizes remain an open question.

These contributions are obtained by introducing several technical ingredients. First, we provide deterministic sufficient conditions on the sketching operator $\mathcal{A}$ so that the Restricted Isometry Property (RIP) (3) holds for mixture models using weighted Fourier features under some conditions. This is achieved thanks to a parametrization of the so-called *normalized secant set* of the set $\mathfrak{G}_k$ with respect to the MMD. This parametrization uses the notion of *dipoles* defined and used in [GBKT21b]. We extend the use of these tools and show how the proof of (3) boils down to the study of suprema of functions defined on $\mathbb{R}^d$. The benefit of this approach is to reduce the study from the normalized secant set, which is a set of signed measures with a geometry that is hard to grasp, to the much easier study of a few explicit functions defined on subsets of $\mathbb{R}^d$. Second, we leverage these deterministic sufficient conditions to establish the RIP when the sketching operator is random. This result

is instantiated to carry out the proof of (3) in the case of a sketching operator built using i.i.d. frequencies. To constrast our result, which *does not* require the use of importance sampling with the analysis given in [GBKT21b], we establish that the high level structure of the latter *requires importance sampling*. The technique we propose is thus both more general and much closer to practice. Moreover, we establish lower bounds showing the impossibility to achieve sharper estimates of provably good sketch sizes sufficient for a sketching operator based on Fourier feature maps: these impossibility results hold both for our analysis and existing one [GBKT21b]. Finally, we discuss the few steps that remain open to exploit our analysis to prove the RIP even when the frequencies are not necessarily independent, e.g. in the context of structured random Fourier features [LSS13, CS16, CRS+18].

This article is structured as follows. In Section 2 we recall some notions and results from [GBKT21a, GBKT21b] that are relevant to our study, as well as their main limitations which motivate this work. In Section 3 we present our results. We conclude and discuss some perspectives in Section 4.

2 Main tools

We recall some results and definitions relevant to position our contributions.

2.1 Sketching operator

The (random) feature maps that we will consider will bear special relations with the considered kernel κ, this will soon be discussed. For the moment we observe that

- for *any* bounded vector-valued function $\Phi : \mathcal{X} \to \mathbb{C}^m$, one can define a *sketching operator*
$$\mathcal{A} : \left\{ \begin{array}{ll} \mathcal{P}(\mathcal{X}) & \to \mathbb{C}^m \\ \pi & \mapsto \int_{\mathcal{X}} \Phi(x) \mathrm{d}\pi(x), \end{array} \right. \tag{8}$$
with $\mathcal{P}(\mathcal{X})$ the set of probability distributions on $\mathcal{X}$. Jordan decomposition [Hal50] allows to extend $\mathcal{A}$ to the set $\mathcal{M}(\mathcal{X})$ of finite signed measures, see [GBKT21a, Appendix A.2.].
- for *any* bounded kernel, one can define for every probability distributions $\pi, \pi' \in \mathcal{P}(\mathcal{X})$
$$\langle \pi, \pi' \rangle_\kappa := \mathbb{E}_{X \sim \pi} \mathbb{E}_{Y \sim \pi'} \kappa(X, Y). \tag{9}$$
and extend this "inner product", as well as the definition (2) of the MMD, to all finite signed measures in $\mathcal{M}(\mathcal{X})$. Thanks to a polarization identity, $\langle \nu, \nu' \rangle_\kappa = \frac{1}{4} \left(\|\nu + \nu'\|_\kappa^2 - \|\nu - \nu'\|_\kappa^2 \right)$, these can be manipulated as usual inner products and norms[3].

We will write $\langle f, \pi \rangle := \mathbb{E}_{X \sim \pi} f(X)$, with implicit integrability assumption of function f with respect to the probability distribution π. This is extended by a Jordan decomposition to $\langle f, \pi \rangle$ with $\nu \in \mathcal{M}(\mathcal{X})$. It should always be clear whether the bracket notation $\langle ., . \rangle$ stands for this shorthand or for the classical Hermitian inner product between vectors in $\mathbb{C}^m$.

[3]Note that $\mathcal{M}(\mathcal{X})$ equipped with $\|.\|_\kappa$ is not necessarily a Hilbert space, since $\mathcal{M}(\mathcal{X})$ is not necessarily complete with respect to $\|.\|_\kappa$. See [SZ21, Theorem 3.1] for details.

2.2 Separated mixture model, normalized secant set, and dipoles

We focus our analysis on mixture modeling with a location-based family [GBKT21b, Definition 6.1]: given a base probability distribution π_0 on $\mathbb{R}^d$ (for example π_0 may be the Dirac at zero, or a centered Gaussian) and a family $\Theta \subseteq \mathbb{R}^d$ of translation parameters, we consider $(\pi_\theta)_{\theta\in\Theta}$ where π_θ is the distribution of $X+\theta$ where $X \sim \pi_0$ and observe that the map $\mathcal{I} : \theta \mapsto \mathcal{I}(\theta) = \pi_\theta$ is injective. Given a translation invariant metric ρ on $\Theta \subseteq \mathbb{R}^d$ (for example, $\rho(\theta,\theta')$ may be the Euclidean distance between θ and θ', or $\|\theta'-\theta\|$ with any norm $\|\cdot\|$ on $\mathbb{R}^d$), we denote $\mathcal{T} := (\Theta,\rho,\mathcal{I})$ and consider the set of 2-*separated* k-*mixtures* [GBKT21b] defined as

$$\mathfrak{G}_k = \left\{ \sum_{i=1}^k u_i \pi_{\theta_i}; u_i \geq 0, \sum_{i\in[k]} u_i = 1, \theta_i \in \Theta, \forall i \neq i' \in [k], \rho(\theta_i,\theta_{i'}) \geq 2 \right\}. \tag{10}$$

More general separated mixture models of the form (10) can be defined [GBKT21b, Section 5.2] with $\mathcal{T} := (\Theta,\rho,\mathcal{I})$ for any metric space (Θ,ρ) and injective map $\mathcal{I} : \Theta \to \mathcal{P}(\mathcal{X})$, in which case we also denote $\pi_\theta := \mathcal{I}(\theta)$. The study of (3) for general separated mixture models motivates the introduction of the normalized secant set $\mathcal{S}_k$ defined as follows

$$\mathcal{S}_k := \left\{ \frac{\nu-\nu'}{\|\nu-\nu'\|_\kappa}; \nu,\nu' \in \mathfrak{G}_k, \|\nu-\nu'\|_\kappa > 0 \right\}. \tag{11}$$

Indeed, (3) is equivalent to $\sup_{\nu\in\mathcal{S}_k} \left| \|\mathcal{A}\nu\|_2^2 - 1 \right| \leq \delta$. In the following, for every set $\mathfrak{T} \subset \{\frac{\nu}{\|\nu\|_\kappa} : \nu \in \mathcal{M}(\mathcal{X}),\ \|\nu\|_\kappa > 0\}$ of normalized finite signed measures and every sketching operator $\mathcal{A}$ we denote

$$\delta(\mathfrak{T}|\mathcal{A}) := \sup_{\nu\in\mathfrak{T}} \left| \|\mathcal{A}\nu\|_2^2 - 1 \right|. \tag{12}$$

As we shall see, the elements of the normalized secant set may be approximated as a mixture of elementary measures called normalized dipoles.

Definition 1 (Dipoles [GBKT21b, Definitions 5.3, 5.6])**.** *A finite signed measure $\iota \in \mathcal{M}(\mathcal{X})$ is a* dipole *w.r.t. $\mathcal{T} = (\Theta,\rho,\mathcal{I})$ if $\iota = \alpha_1\pi_{\theta_1} - \alpha_2\pi_{\theta_2}$, where $\theta_1,\theta_2 \in \Theta$, $\rho(\theta_1,\theta_2) \geq 1$ and $\alpha_1,\alpha_2 \geq 0$. Two dipoles ι,ι', are* 1-separated *if $\iota = \alpha_1\pi_{\theta_1} - \alpha_2\pi_{\theta_2}$, $\iota' = \alpha'_1\pi_{\theta'_1} - \alpha'_2\pi_{\theta'_2}$, where $\rho(\theta_i,\theta'_j) \geq 1$ for $i,j \in \{1,2\}$. The set of* normalized dipoles *(with respect to kernel κ) is denoted*

$$\mathfrak{D} = \mathfrak{D}(\mathcal{T}) := \left\{ \iota = \tilde{\iota}/\|\tilde{\iota}\|_\kappa, \tilde{\iota} \text{ is a dipole such that } \|\tilde{\iota}\|_\kappa > 0 \right\}, \tag{13}$$

and $\mathfrak{D}^2_{\neq} \subseteq \mathfrak{D}\times\mathfrak{D}$ denotes the set of pairs of 1-separated normalized dipoles.

Dipoles offer a convenient parametrization of the (un-normalized) secant set.

Lemma 1 ([GBKT21b, Lemma 5.4])**.** *Let $\pi,\pi' \in \mathfrak{G}_k$. There exist $\ell \leq 2k$ nonzero dipoles $(\iota_l)_{l\in[\ell]}$ that are pairwise 1-separated and satisfy*

$$\pi - \pi' = \sum_{i=1}^\ell \iota_i. \tag{14}$$

In other words, every element of the (unnormalized) secant set $\pi-\pi'$ is the sum of at most $2k$ dipoles. The decomposition (14) is convenient when calculating the squared MMD norm $\|\pi-\pi'\|_\kappa^2 = \sum_{i=1}^\ell \|\iota_i\|_\kappa^2 + \sum_{i\neq i'} \langle \iota_i,\iota_{i'}\rangle_\kappa$. In particular, under some additional assumptions on the kernel that we now discuss, the cross scalar products $\langle \iota_i,\iota_{i'}\rangle_\kappa$ are close to 0 so that $\|\pi-\pi'\|_\kappa^2$ can be approximated by $\sum_{i=1}^\ell \|\iota_i\|_\kappa^2$.

2.3 Kernel coherence

To conduct our analysis, we require further assumptions on the compatibility of the positive definite kernel κ with the parameterized family of distributions $\mathcal{T} := (\Theta, \rho, \mathcal{I})$.

Definition 2. *Given a family $\mathcal{T} := (\Theta, \rho, \mathcal{I})$, a kernel κ is said to be*

1. non-degenerate *with respect to $\mathcal{T}$ if $\|\pi_\theta\|_\kappa > 0$ for every $\theta \in \Theta$. This allows to define the $\mathcal{T}$-*normalized kernel $\bar{\kappa}$ *as*

$$\forall \theta, \theta' \in \Theta, \ \ \bar{\kappa}(\theta, \theta') := \frac{\langle \pi_\theta, \pi_{\theta'} \rangle_\kappa}{\|\pi_\theta\|_\kappa \|\pi_{\theta'}\|_\kappa}. \tag{15}$$

2. locally characteristic *with respect to $\mathcal{T}$ [GBKT21b, Definition 5.5] if it is non-degenerate with respect to $\mathcal{T}$ and $|\bar{\kappa}(\theta, \theta')| < 1$ for every $\theta, \theta' \in \Theta$ such that $0 < \rho(\theta, \theta') \leq 1$. This ensures that $\|\pi_\theta - \alpha \pi_{\theta'}\|_\kappa > 0$ for every $\alpha \in \mathbb{R}$ whenever $0 < \rho(\theta, \theta') \leq 1$.*

Definition 3 (Coherence [GBKT21b, Definition 5.7])**.** *Given an integer $\ell \geq 1$ the ℓ-*coherence of κ with respect to $\mathcal{T} = (\Theta, \rho, \mathcal{I})$*, denoted $c_\ell = c_\ell(\kappa)$ is the smallest $c \geq 0$ such that, for any pairwise 1-separated dipoles $(\iota_i)_{i \in [\ell]}$ such that $\sum_{i=1}^\ell \|\iota_i\|_\kappa^2 > 0$, we have*

$$1 - c \leq \frac{\| \sum_{i=1}^\ell \iota_i \|_\kappa^2}{\sum_{i=1}^\ell \|\iota_i\|_\kappa^2} \leq 1 + c. \tag{16}$$

The kernel κ has mutual coherence μ with respect to $\mathcal{T}$ *if it is locally characteristic wrt $\mathcal{T}$ and*

$$\mu = \mu(\mathfrak{D}^2_{\neq} | \kappa) := \sup_{(\iota, \iota') \in \mathfrak{D}^2_{\neq}} |\langle \iota, \iota' \rangle_\kappa| \tag{17}$$

By analogy with the kernel coherence we define the coherence of a sketching operator:

Definition 4 (Operator Coherence)**.** *For any sketching operator $\mathcal{A}$ and any set $\mathfrak{T} \subseteq \mathfrak{D}^2_{\neq}$*

$$\mu(\mathfrak{T} | \mathcal{A}) := \sup_{(\iota, \iota') \in \mathfrak{T}} |\mathfrak{Re} \langle \mathcal{A}\iota, \mathcal{A}\iota' \rangle|. \tag{18}$$

As shown in [GBKT21b, Lemma 5.8], if the kernel κ has mutual coherence μ with respect to $\mathcal{T}$ then κ has ℓ-coherence bounded by $c' := \mu(\ell - 1)$. This is reminiscent (and indeed inspired by) classical results on incoherent dictionaries in sparse recovery, see e.g. [FR13, Chapter 5]. In particular, if κ has mutual coherence bounded by $\mu < 1/(2k-1)$ then the quasi-Pythagorean property (16) holds for $\ell = 2k$ with

$$c = c_{2k} \leq (2k-1)\mu < 1. \tag{19}$$

This implies that the normalized secant set $\mathcal{S}_k$ is made of "nice" mixtures of separated dipoles.

Proposition 1 ([GBKT21b, Lemma B.1])**.** *Let $k \geq 1$ be an integer, and denote by $c = c_{2k}$ the $2k$-coherence of the kernel κ with respect to $\mathcal{T}$. Under the assumption that $c < 1$, we have*

$$\mathcal{S}_k \subset \left\{ \sum_{i=1}^{2k} \alpha_i \iota_i : (1+c)^{-1} \leq \sum_{i=1}^{2k} \alpha_i^2 \leq (1-c)^{-1}, \alpha_i \geq 0, (\iota_i, \iota_j) \in \mathfrak{D}^2_{\neq}, 1 \leq i \neq j \leq 2k \right\}. \tag{20}$$

In other words, the normalized secant set $\mathcal{S}_k$ is made of mixtures of $2k$ normalized dipoles with weights of controlled ℓ^2 norm. This decomposition comes in handy when looking for an upper bound of high order moments as we will study soon for measure concentration arguments.

2.4 Location-based families and shift-invariant kernels

In most of this paper we focus on shift-invariant kernels, generally assumed to be normalized ($\kappa(0,0)=1$). As often we use the abuse of notation $\kappa(x,y)=\kappa(x-y)$. When the family $(\pi_\theta)_{\theta\in\Theta}$ used to define the mixture model (10) is location-based, we have [GBKT21b, Proposition 6.2]

$$\forall \theta\in\Theta,\quad \|\pi_\theta\|_\kappa = \|\pi_0\|_\kappa, \tag{21}$$

hence κ is non-degenerate with respect to $\mathcal{T}$ if, and only if, $\|\pi_0\|_\kappa>0$. Moreover, the $\mathcal{T}$-normalized kernel $\bar{\kappa}$ itself is also shift-invariant. We also abuse notations and denote $\bar{\kappa}(\theta-\theta')=\bar{\kappa}(\theta,\theta')=\frac{1}{\|\pi_0\|_\kappa^2}\kappa(\pi_\theta,\pi_{\theta'})$. The low-coherence property is satisfied by classical shift-invariant kernels and location-based families $\mathcal{T}$ [GBKT21b].

Example 1 (Mixtures of Diracs and the Gaussian kernel [GBKT21b, Definition 6.9])**.** *In this case, π_0 is the Dirac distribution at 0, $\rho=\|.\|_2/\epsilon$ where $\epsilon>0$, and κ is the Gaussian kernel:*

$$\kappa(x,x') := \exp\Big(-\frac{\|x-x'\|_2^2}{2s^2}\Big), \tag{22}$$

with $s>0$ a scale parameter. The normalized kernel writes [GBKT21b, Section 6.3.1]

$$\bar{\kappa}(\theta-\theta') = \exp\Big(-\frac{\|\theta-\theta'\|_2^2}{2s^2}\Big). \tag{23}$$

By [GBKT21b, Theorem 5.16, Lemma 6.10, Theorem 6.11], κ is locally characteristic and its mutual coherence with respect to $\mathcal{T}$ is smaller than or equal to $12/(16(2k-1))$ as soon as

$$\epsilon \geq s/s_k^\star \quad \text{with } s_k^* := (4\sqrt{\log(ek)})^{-1}. \tag{24}$$

Example 2 (Mixtures of Gaussians and the Gaussian kernel [GBKT21b, Definition 6.9])**.** *In this case, we consider the Mahalanobis norm $\|.\|_{\mathbf{\Sigma}}$, defined by $\|x\|_{\mathbf{\Sigma}} := \sqrt{\langle x,\mathbf{\Sigma}^{-1}x\rangle} = \|\mathbf{\Sigma}^{-1/2}x\|_2$ for $x\in\mathbb{R}^d$, where $\mathbf{\Sigma}\in\mathbb{R}^d$ is a positive definite matrix, $\rho=\|.\|_{\mathbf{\Sigma}}/\epsilon$ where $\epsilon>0$, and $\pi_0=\mathcal{N}(0,\mathbf{\Sigma})$. Finally κ is the Gaussian kernel:*

$$\kappa(x,x') := \exp\Big(-\frac{\|x-x'\|_{\mathbf{\Sigma}}^2}{2s^2}\Big), \tag{25}$$

with $s>0$ a scale parameter. The normalized kernel writes [GBKT21b, Section 6.3.1]

$$\bar{\kappa}(\theta-\theta') = \exp\Big(-\frac{\|\theta-\theta'\|_{\mathbf{\Sigma}}^2}{2(2+s^2)}\Big). \tag{26}$$

By [GBKT21b, Theorem 5.16, Lemma 6.10, Theorem 6.11], κ is locally characteristic and its mutual coherence with respect to $\mathcal{T}$ is smaller than or equal to $12/(16(2k-1))$ as soon as

$$\epsilon \geq \frac{\sqrt{s^2+2}}{s_k^\star} \quad \text{with } s_k^* := (4\sqrt{\log(ek)})^{-1}. \tag{27}$$

2.5 Random Fourier features

The analysis in [KBGP18, GBKT21b] is conducted using a variant of the random Fourier feature map described in the introduction, using importance sampling. It also uses the more general notion of a κ-compatible random sketching operator.

Definition 5. *Consider a kernel κ on $\mathcal{X} \times \mathcal{X}$ and a random feature map Φ defined as in (5) from a parametric family $\{\phi_\omega : \mathcal{X} \to \mathbb{C}\}_{\omega \in \Omega}$ and random parameters $(\omega_1, \dots, \omega_m)$, i.i.d. or not. The feature map (and by extension the resulting sketching operator $\mathcal{A}$) is said to be κ-compatible if the expected value (with respect to the draw of frequencies) of the hermitian inner-product $\langle \Phi(x), \Phi(y) \rangle$ is exactly $\kappa(x, y)$, cf (6).*

Definition 6. *Given a weight function $w : \mathbb{R}^d \to (0, +\infty)$, a sketching operator $\mathcal{A}$ is a w-weighted Fourier feature (w-FF) sketching operator if it is built from a feature map Φ as in (5) with some frequency vectors $\omega_1, \dots, \omega_m$ and individual components defined as*

$$\phi_\omega = \phi_\omega^w : x \mapsto e^{2\pi \imath \omega^\top x} / w(\omega). \tag{28}$$

If the frequency components are jointly drawn (i.i.d. or not) from some probability distribution then $\mathcal{A}$ is called a w-weighted random *Fourier feature (w-RFF) sketching operator.*

We will often drop the the dependency of ϕ_ω in w for brevity of notation, and call $\mathcal{A}$ a WFF (or RFF) sketching operator when there is no need to specify the corresponding w.

Definition 7. *Consider a normalized shift-invariant kernel κ on $\mathcal{X} = \mathbb{R}^d$. A weight function $w : \mathbb{R}^d \to (0, \infty)$ is said to be κ-compatible if*

$$\int_{\mathbb{R}^d} w^2(\omega) \hat{\kappa}(\omega) \mathrm{d}\omega = 1. \tag{29}$$

Example 3. *Given a normalized shift-invariant kernel κ on $\mathcal{X} = \mathbb{R}^d$, if w is κ-compatible then*

$$\Lambda(\omega) := w^2(\omega) \hat{\kappa}(\omega) \tag{30}$$

defines a probability density function. For frequency vectors drawn (i.i.d. or not) with common marginal probability density function Λ and any x, y we have

$$\mathbb{E}_{\omega \sim \Lambda} \phi_\omega^w(x) \overline{\phi_\omega^w(y)} = \int w^{-2}(\omega) e^{2\pi \imath \omega^\top (x-y)} \Lambda(\omega) \mathrm{d}\omega = \int e^{2\pi \imath \omega^\top (x-y)} \hat{\kappa}(\omega) d\omega = \kappa(x - y), \tag{31}$$

hence the random feature map Φ is κ-compatible (Definition 5)

$$\mathbb{E}_{\omega_1, \dots, \omega_m} \langle \Phi(x), \Phi(y) \rangle = \frac{1}{m} \sum_{j=1}^m \mathbb{E}_{\omega_1, \dots, \omega_m} \phi_{\omega_j}^w(x) \overline{\phi_{\omega_j}^w(y)} = \frac{1}{m} \sum_{j=1}^m \mathbb{E}_{\omega_j \sim \Lambda} \phi_{\omega_j}^w(x) \overline{\phi_{\omega_j}^w(y)} = \kappa(x, y),$$

moreover,

$$\forall \nu, \nu' \in \mathcal{S}_k, \ \ \mathbb{E}_{\omega \sim \Lambda} \langle \phi_\omega^w, \nu \rangle \overline{\langle \phi_\omega^w, \nu \rangle} = \langle \nu, \nu' \rangle_\kappa. \tag{32}$$

Specializing this to $\nu = \nu'$ yields $\mathbb{E}_{\omega \sim \Lambda} |\langle \phi_\omega^w, \nu \rangle|^2 = \|\nu\|_\kappa^2$.

With possibly distinct marginal densities $\omega_j \sim \Lambda_j$, we similarly get that the expectation of $\langle \Phi(x), \Phi(y) \rangle$ is $\int w^{-2}(\omega) e^{2\pi \imath \omega^\top (x-y)} \frac{1}{m} (\sum_{j=1}^m \Lambda_j(\omega)) d\omega$ hence the same

conclusion holds if, and only if, the "average marginal density" satisfies almost everywhere

$$\frac{1}{m}(\sum_{j=1}^{m}\Lambda_j(\omega)) = w^2(\omega)\hat{\kappa}(\omega). \tag{33}$$

It can also occur that the frequencies $\omega_1, \ldots, \omega_m$ are not independent, for example using *structured random features* [LSS13, CS16, CRS+18]. Assuming here for simplicity that m is a multiple of d, the construction of such frequencies is such that the matrix $\boldsymbol{\Omega} \in \mathbb{R}^{d\times m}$ with columns ω_j, $1 \leq j \leq m$ is defined as the concatenation of m/d i.i.d. random matrices $\boldsymbol{B}_i \in \mathbb{R}^{d\times d}$, $1 \leq i \leq m/d$. This is advantageous when each $\boldsymbol{B}_i$ is structured in such a way that the product $\boldsymbol{B}z$ for $z \in \mathbb{R}^d$ costs $\mathcal{O}(d\log(d))$ instead of $\mathcal{O}(d^2)$, e.g. when d is a power of two and $\boldsymbol{B} = \boldsymbol{D}_1\boldsymbol{H}\boldsymbol{D}_2\boldsymbol{H}\boldsymbol{D}_3\boldsymbol{H}$ with $\boldsymbol{H}$ the matrix associated to the (fast) Hadamard transform, and $\boldsymbol{D}_\ell$, $1 \leq \ell \leq 3$ appropriate random diagonal matrices. When d is not a power of 2 and/or m is not a multiple of d the construction can be adapted using padding techniques. We refer the reader to [Cha20, Chapter 5], where an overview of such techniques is summarized. It can also be proved that under appropriate conditions the resulting random feature map Φ is still κ-compatible (Definition 5), see e.g. [LSS13, Lemma 7] and [Cha20, Lemma 5.6] for results when κ is a Gaussian kernel.

2.6 Existing results and their limitations

To establish a bound of the type (3) with a general mixture model, or equivalently to bound the constant $\delta(\mathcal{S}_k|\mathcal{A})$ of (12) the strategy deployed in [GBKT21b] exploits covering numbers, pointwise concentration, and a deterministic bound on a certain Lipschitz constant. Indeed, if a family $(\nu_i)_{i\in[\mathcal{N}]}$ of elements of the secant set $\mathcal{S}_k$ satisfies

$$\forall \nu \in \mathcal{S}_k,\ \ \exists i \in [\mathcal{N}],\ \ \big|\|\mathcal{A}\nu\|_2^2 - \|\mathcal{A}\nu_i\|_2^2\big| \leq \frac{\tau}{2}, \tag{34}$$

then

$$\sup_{\nu\in\mathcal{S}_k}\big|\|\mathcal{A}\nu\|_2^2 - 1\big| \leq \sup_{i\in[\mathcal{N}]}\big|\|\mathcal{A}\nu_i\|_2^2 - 1\big| + \frac{\tau}{2}. \tag{35}$$

hence proving that $\delta(\{\nu_1, \ldots, \nu_\mathcal{N}\}|\mathcal{A}) \leq \tau/2$ holds is sufficient to deduce that $\delta(\mathcal{S}_k|\mathcal{A}) \leq \tau$. Moreover, assuming there is $v > 0$ such that for every sketch size $m \geq 1$ the corresponding RFF sketching operator $\mathcal{A}$ satisfies a punctual concentration estimate of the form

$$\sup_{\nu\in\mathcal{S}_k}\mathbb{P}\big(\big|\|\mathcal{A}\nu\|_2^2 - 1\big| > \frac{\tau}{2}\big) \leq 2\exp\left(-\frac{m}{v}\right), \tag{36}$$

a union bound allows to deduce that $\delta(\{\nu_1, \ldots, \nu_\mathcal{N}\}|\mathcal{A}) \leq \tau/2$ holds with probability at least $1 - 2\mathcal{N}\cdot\exp(-m/v)$ on the draw of $\mathcal{A}$. These arguments show that, for any $0 < \eta < 1$, if the sketch size satisfies $m \geq v\log 2\mathcal{N}/\eta$ then $\delta(\mathcal{S}_k|\mathcal{A}) \leq \tau$ with probility at least $1 - \eta$.

The smallest size of a family satisfying (34) is a covering number of $X = \mathcal{S}_k$ with the pseudo-metric $d(\nu, \nu') := |\|\mathcal{A}\nu\|_2^2 - \|\mathcal{A}\nu'\|_2^2|$ (see e.g. [CS01] for the well-known definition of coverings in a pseudo-metric space (X, d), and covering numbers, denoted $\mathcal{N}(X, d, \epsilon)$ at scale $\epsilon > 0$). However, in the case of random sketching, this pseudo-metric *depends on the random feature map* Φ (or equivalently on the random sketching operator $\mathcal{A}$). To circumvent this difficulty the approach of [GBKT21b,

Proof of Lemma B.4] is to observe that

$$\forall \nu, \nu' \in \mathcal{S}_k, \quad \big| \|\mathcal{A}\nu\|_2^2 - \|\mathcal{A}\nu'\|_2^2 \big| \leq M \|\nu - \nu'\|_{\mathcal{F}}, \qquad \text{with } M := 2 \sup_{\nu \in \mathcal{S}_k} \|\nu\|_{\mathcal{F}}, \tag{37}$$

where $\|\nu\|_{\mathcal{F}} := \sup_{\omega \in \mathbb{R}^d} |\langle \phi_\omega, \nu \rangle|$ defines a *deterministic* pseudo-norm on finite signed measures. Therefore, a covering (ν_i) of $\mathcal{S}_k$ with respect to $d'(\nu, \nu') := \|\nu - \nu'\|_{\mathcal{F}}$ at scale $\tau/2M$ satisfies (34), with $\mathcal{N} = \mathcal{N}(\mathcal{S}_k, d', \|\cdot\|_{\mathcal{F}}, \tau/2M)$. Inequality (37) means that the (random) function $\nu \mapsto \|\mathcal{A}\nu\|_2^2$ is Lipschitz with respect to the metric d', with (deterministic) Lipschitz constant M.

For RFF sketching with location-based families, we will see that getting a finite M highly constrains the function w (cf (28)). *A primary objective of this paper is to relax this constraint.*

For the concentration estimate (36), the approach of [GBKT21b] is generic for general mixture models and general sketching operators defined with a feature map as in (5) using a parameterized family ϕ_ω where $\omega \sim \Lambda$ are i.i.d. parameters. It combines a Bernstein inequality with an assumption on higher order moments on normalized dipoles. Indeed, a consequence of Proposition 1 is that if the $2k$-coherence of κ is bounded by c then[4]

$$\forall q \geq 2, \quad \sup_{\nu \in \mathcal{S}_k} \mathbb{E}_{\omega \sim \Lambda} |\langle \phi_\omega, \nu \rangle|^{2q} \leq \Big(\frac{2k}{1-c} \Big)^{2q} \sup_{\iota \in \mathfrak{D}} \mathbb{E}_{\omega \sim \Lambda} |\langle \phi_\omega, \iota \rangle|^{2q}, \tag{38}$$

i.e., controlling the moments on normalized dipoles is enough to control them on the normalized secant. Assuming a κ-compatible random feature map, this was shown to imply a concentration of the squared norm of the sketch $\|\mathcal{A}\nu\|_2^2$ around its expected value [GBKT21b, Theorem 5.11].

The following is the specialization of [GBKT21b, Theorem 5.11] to the case of a κ-compatible RFF sketching operator with i.i.d. frequencies $\omega_j \sim \Lambda$, $1 \leq j \leq m$.

Theorem 1 ([GBKT21b, Theorem 5.11])**.** *Let $\mathcal{T} := (\Theta, \rho, \mathcal{I})$ be a location-based family with base distribution π_0 on $\mathbb{R}^d$, with $\Theta \subseteq \mathbb{R}^d$ a bounded subset, $\rho(\cdot, \cdot) := \|\cdot - \cdot\|$ where $\|\cdot\|$ is some norm on $\mathbb{R}^d$. Consider a normalized shift-invariant kernel κ with an integer $k \geq 1$ such that κ has its $2k$-coherence with respect to the location-based family $\mathcal{T}$ bounded by $0 \leq c \leq 3/4$*

Consider w a κ-compatible weight function, $\Lambda = w^2 \hat{\kappa}$, and assume that there $\beta_1 > 0$ and $\beta_2 \geq 1$ such that

$$\forall q \geq 2, \quad \sup_{\iota \in \mathfrak{D}} \mathbb{E}_{\omega \sim \Lambda} |\langle \phi_\omega^w, \iota \rangle|^{2q} \leq \frac{q!}{2} \beta_1 \beta_2^{q-1}. \tag{39}$$

Then for every $m \geq 1$ the RFF sketching operator $\mathcal{A}$ built with i.i.d. random frequencies $\omega_j \sim \Lambda$ is κ-compatible and

$$\forall \tau > 0, \quad \sup_{\nu \in \mathcal{S}_k} \mathbb{P}\big(\big| \|\mathcal{A}\nu\|_2^2 - 1 \big| > \tau \big) \leq 2 \exp\left(- \frac{m}{v} \right), \tag{40}$$

where $v = v(k, \tau) := 2 v_k (1 + \tau/3)/\tau^2$, with $v_k = 16 e k \beta_2 \log^2(4ek\beta_1)$.

The quantity $v(k, \tau) > 0$ is reminiscent of a variance and depends on k, τ (as displayed by the notation) but also more implicitly on $\sup_{\iota \in \mathfrak{D}} \mathbb{E}_{\omega \sim \Lambda} |\langle \phi_\omega^w, \iota \rangle|^{2q}$, $q \geq 2$

[4] See before Eq. (109) in the proof of Theorem 5.11 in [GBKT21b, Section B.1]

via the constants β_1, β_2. As will shortly see, the dependence in k is something we pay when estimating the sketch size, and it is natural to wonder whether this is due to the analysis or intrinsic. *A side contribution of this paper is to show that this is somehow inevitable for this type of result.*

The above arguments from [GBKT21b] lead to the following result.

Theorem 2. *Consider $\mathcal{T}$ κ, w, Λ, $\beta_1 > 0, \beta_2 \geq 1$ as in Theorem 1. Assume that (39) holds, and that the constant M defined in (37) is finite. Then, with $v(k,\cdot)$ as in Theorem 1, we have*

$$\forall \tau > 0,\ \mathbb{P}\big(\delta(\mathcal{S}_k|\mathcal{A}) > \tau\big) \leq 2\mathcal{N}\Big(\mathcal{S}_k, \|.\|_{\mathcal{F}}, \frac{\tau}{2M}\Big) \exp\left(-\frac{m}{v(k,\tau/2)}\right). \tag{41}$$

In particular, under the assumptions of Theorem 2, the property $\delta(\mathcal{S}_k|\mathcal{A}) \leq \tau$ holds with probability at least $1-\eta$ where $0 < \eta < 1$ as soon as the sketch size satisfies

$$m \geq v(k,\tau/2) \log\Big(2\mathcal{N}\Big(\mathcal{S}_k, \|.\|_{\mathcal{F}}, \frac{\tau}{2M}\Big)/\eta\Big). \tag{42}$$

However, this estimate on a sufficient sketch size is only relevant when the constant M defined in (37) is finite and $\mathcal{N}(\mathcal{S}_k, \|.\|_{\mathcal{F}}, \tau/(2M)) < +\infty$. By [GBKT21b, Theorem 5.12, Theorem 5.15, Lemma 6.4, Lemma 6.7], this is the case when κ is *strongly locally characteristic* with respect to $\mathcal{T}$ (a property satisfied in Examples 1 and 2, cf [GBKT21b, Theorem 6.11]) under the assumption that

$$\sup_{\omega \in \mathbb{R}^d} |\langle \phi_\omega^w, \pi_0 \rangle| \cdot \max(1, \|\omega\|_\star, \|\omega\|_\star^2) < +\infty \tag{43}$$

with $\|.\|_\star$ the dual norm of $\|.\|$ defined by $\|u\|_\star := \sup_{\|v\|\leq 1} u^\top v$. Then, by [GBKT21b, Theorem 5.12, Lemma 6.7], we have for fixed $\epsilon > 0$, $\log \mathcal{N}(\mathcal{S}_k, \|.\|_{\mathcal{F}}, \epsilon) = \mathcal{O}(kd)$ up to logarithmic factors in $k, d, 1/\epsilon$, so that a sufficient sketch size to have the desired result with high probability can be shown to satisfy $m = \mathcal{O}(k^2 d)$ up to logarithmic factors[5] in $k, d, M/\tau$.

Condition (43) constrains the choice of weight function w for models such as mixtures of Diracs. Indeed, for this family of mixtures $|\langle \phi_\omega^w, \pi_0 \rangle| = 1/w(\omega)$ so that (43) implies $w(\omega) \gtrsim \max(1, \|\omega\|_\star, \|\omega\|_\star^2)$ (which imposes constraints on the behavior of w both around $\omega \to 0$ and $\|\omega\|_\star \to +\infty$). This is the main reason why the analysis in [GBKT21b] is limited to random Fourier features *with importance sampling*, while plain sampling seems enough in practical experiments. *The main contribution of this paper is to establish results valid without assuming* (43).

In summary, besides a (strongly locally characteristic) kernel with bounded $2k$-coherence and a κ-compatible weight function w, existing results are built on the following assumptions:

1. moment conditions (39) on $\Lambda = w^2\hat{\kappa}$, to establish punctual concentration;
2. growth conditions (43) on w, to control the Lipschitz constant M from (37) and the associated covering numbers;
3. i.i.d. frequencies $\omega_j \sim \Lambda$, $1 \leq j \leq m$.

Under these assumptions, a sketch size of the order of $k^2 d$ (up to logarithmic factors) is proved to be sufficient. As experiments conducted in [KTTG17] suggested that the same result should hold with a smaller sufficient sketch size $m = \mathcal{O}(kd)$, this raises the

[5] If the sup in (43) is not only finite but at most polynomial in k, d then $\log M/\tau$ is also logarithmic in k, d.

question whether the theoretical bound of Theorem 2 can be refined. An approach which would provide an easy fix would be if we could simply remove a k factor in the punctual concentration estimate of Theorem 1 via a more subtle analysis. This would indeed naturally insert in the analysis of either [GBKT21b] or of this work to yield the desired order of magnitude of m. A second contribution of this work is to show that such a uniform improvement of concentration estimates is simply not possible.

3 Main results

This work aims to overcome some shortcomings of the theoretical analysis given in [GBKT21b]. First, we show in Section 3.1 that some of the growth conditions (43) are necessary to exploit the analysis given in [GBKT21b]. Then, in Section 3.3 we give our main contribution: we provide an alternative analysis that allows to completely relax the growth conditions (43) on w. This yields results (still with a sketch size m of the order of k^2d) for a much less constrained family of importance sampling schemes, including plain sampling $w \equiv 1$. This is primarily achieved by circumventing the deterministic control of the Lipschitz constant M from (37): instead, we provide a stochastic control of a "typical" Lipschitz constant, thanks to a reduction of the stochastic control of $\delta(\mathcal{S}_k|\mathcal{A})$ to a stochastic control of its equivalent on dipoles, $\delta(\mathfrak{D}|\mathcal{A})$, and of the coherence of the sketching operator, $\mu(\mathfrak{D}^2_{\neq}|\mathcal{A})$. This yields a substantial streamlining of the analysis which is then further reduced to the equivalent quantities for so-called normalized monopoles and balanced normalized dipoles. This is achieved thanks to deterministic bounds on $\delta(\mathcal{S}_k|\mathcal{A})$ established in Section 3.2. Finally, we show in Section 3.4 that both the analysis given in [GBKT21b] and the one given in this work cannot be fixed by a simple improvement of concentration estimates to to close the gap between sufficient sketch sizes endowed with theoretical guarantees, which scale essentially as $O(k^2d)$, and practically observed sketch sizes, which scale as $O(kd)$ [KTTG17].

3.1 On the necessity of conditions (43)

As mentioned in Section 2.6, the analysis of [GBKT21b] assumes that condition (43) holds to obtain that $M := \sup_{\nu\in\mathcal{S}_k} \|\nu\|_{\mathcal{F}}$ and the covering numbers $\mathcal{N}(\mathcal{S}_k, \|.\|_{\mathcal{F}}, \epsilon)$ are finite. Here we establish a partial converse. The following Proposition is proved in Appendix A.3.

Proposition 2. *Consider a normalized shift-invariant kernel κ. Consider a location-based family $\mathcal{T} = (\Theta, \varrho, \mathcal{I})$ with base distribution π_0 where Θ contains a neighborhood of zero and $\varrho(\theta, \theta') := \|\theta - \theta'\|$ for some arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$. Assume that $\bar{\kappa}$, as defined in (15), is C^2 at zero, and assume that $\nabla^2\bar{\kappa}(0) \in \mathbb{R}^{d\times d}$, the Hessian matric of $\bar{\kappa}$ at 0, is non-zero. Then, for weighted Fourier features, we have for every integer $k \geq 1$, and separated k-mixture model $\mathfrak{G}_k$ from (10)*

$$\sup_{\nu\in\mathcal{S}_k} \|\nu\|_{\mathcal{F}} \geq \sup_{\omega\in\mathbb{R}^d} \frac{|\langle \phi^w_\omega, \pi_0\rangle|}{\|\pi_0\|_\kappa} \max\left(1, \frac{2\pi}{\sqrt{\|\nabla^2\bar{\kappa}(0)\|_{\mathrm{op}}}}\|\omega\|_\star\right).$$

A direct consequence of Proposition 2 is that if $\sup_{\nu\in\mathcal{S}_k} \|\nu\|_{\mathcal{F}} < +\infty$ then

$$\sup_{\omega\in\mathbb{R}^d} |\langle \phi^w_\omega, \pi_0\rangle| \max(1, \|\omega\|_\star) < +\infty$$

which is reminiscent of (43) and plays the role of a partial converse.

In the setting of mixtures of Diracs defined in Example 1, we have $|\langle \phi_\omega^w, \pi_0 \rangle| = 1/w(\omega)$ (cf (28)) and $\bar{\kappa}$ is C^2 at 0 with $\nabla^2 \bar{\kappa}(0) \neq 0$, thus we can apply Proposition 2, and we get that there exists a constant $C > 0$ such that

$$\forall \omega \in \mathbb{R}^d, \;\; w(\omega) \geq C \max\left(1, \|\omega\|_\star\right).$$

Thus, the proof technique of [GBKT21b], which is summarized in Theorem 2, indeed *requires* the weight functions $w(\omega)$ to grow with $\|\omega\|_\star$ to provide non-trivial results. It is in particular inapplicable to the "flat" weight function $w(\omega) = 1$. In contrast, this weight function is covered by our Corollary 2. It is an interesting challenge left to future work to determine if (43) is in fact fully necessary to have both $\sup_{\nu \in \mathcal{S}_\kappa} \|\nu\|_{\mathcal{F}} < +\infty$ *and* $\mathcal{N}(\mathcal{S}_k, \|.\|_{\mathcal{F}}, \epsilon) < \infty$ for some $\epsilon > 0$.

3.2 Sharp deterministic bounds on $\delta(\mathcal{S}_k|\mathcal{A})$

In light of Proposition 2, we propose an alternative analysis to control $\delta(\mathcal{S}_k|\mathcal{A})$ that does not require condition (43). This subsection focuses on the deterministic part of this analysis: first, we upper bound $\delta(\mathcal{S}_k|\mathcal{A})$ (which is defined as a supremum over the normalized secant set) using quantities defined on simpler sets made of dipoles (Proposition 3); then the latter are themselves controlled in terms of even simpler, quantities defined in terms of monopoles and balanced dipoles (Proposition 4); finally all considered quantities are explicitly written as suprema of empirical averages over frequency vectors ω_j (Proposition 5). This will be used in the next subsection to control all quantities in the context of a random sketching operator $\mathcal{A}$. As we will see, the main price to pay for this alternative analysis is that (unlike in Theorem 2, and more generally in results of the same flavor inspired by compressive sensing) $\delta(\mathcal{S}_k|\mathcal{A})$ is no longer proved to be *arbitrarily small* with high probability when the sketch size m is large enough, but only *arbitrarily close to a quantity (smaller than one) depending on the* $2k$*-coherence* of the kernel κ.

3.2.1 From the normalized secant set to normalized monopoles and balanced dipoles

As a first step we bound the targeted quantity, which is defined as a supremum on the normalized secant set, in terms of two suprema defined on simpler sets of normalized dipoles.

Proposition 3 (From the secant set to normalized dipoles)**.** *Consider a kernel* κ*, a family* $\mathcal{T}$*, and an integer* $k \geq 1$ *such that* κ *has its* $2k$*-coherence with respect to* $\mathcal{T}$ *bounded by* $0 \leq c < 1$*. Consider a sketching operator* $\mathcal{A}$ *defined via* (1) *with any feature map* $\Phi(\cdot)$ *such that* $\mathcal{A}\pi_\theta$ *is well-defined for every probability distribution in the family* $\mathcal{T}$*. We have*

$$\delta(\mathcal{S}_k|\mathcal{A}) \leq \frac{1}{1-c}\Big(c + \delta(\mathfrak{D}|\mathcal{A}) + (2k-1)\mu(\mathfrak{D}_{\neq}^2|\mathcal{A})\Big). \tag{44}$$

The proof, which is given in Appendix A.4, is essentially an adaptation of a bound of the restricted isometry constant for incoherent dictionaries in sparse recovery, see e.g. [FR13, Chapter 5]. The minor technicality is to take into account deviations to the normalization of dictionary columns, which is captured by the term $\delta(\mathfrak{D}|\mathcal{A})$.

The upper bound (44) reduces the study of $\delta(\mathcal{S}_k|\mathcal{A})$ to that of $\delta(\mathfrak{D}|\mathcal{A})$ and $\mu(\mathfrak{D}^2_{\neq}|\mathcal{A})$. Note that this bound is only relevant if we can ensure that $\delta(\mathcal{S}_k|\mathcal{A}) < 1$ when $\delta(\mathfrak{D}|\mathcal{A})$ and $\mu(\mathfrak{D}^2_{\neq}|\mathcal{A})$ are sufficiently small, i.e., if $c/(1-c) < 1$, which is possible to achieve in practice by a proper selection of the parameters of the kernel (see Example 2).

We now push the analysis further to scrutinize $\delta(\mathfrak{D}|\mathcal{A})$ and $\mu(\mathfrak{D}^2_{\neq}|\mathcal{A})$. As these two quantities are defined as suprema of a function defined on $\mathfrak{D}$ and $\mathfrak{D}^2_{\neq}$ respectively, which are abstract sets of measures for which the topology is hard to grasp intuitively, we show that both $\delta(\mathfrak{D}|\mathcal{A})$ and $\mu(\mathfrak{D}^2_{\neq}|\mathcal{A})$ boil down to suprema of functions defined on subsets of $\mathbb{R}^d$.

From now on we specialize to a location-based family $\mathcal{T}$ and a shift-invariant kernel κ that is locally characteristic with respect to $\mathcal{T}$. In this setting we have the following property of normalized dipoles [GBKT21b, Lemma C.1]

$$\mathfrak{D} = \left\{ \frac{\nu}{\|\nu\|_\kappa},\ \nu = \|\pi_0\|_\kappa^{-1} s(\pi_{\theta'} - \alpha\pi_\theta); s \in \{-1,1\}, 0 \le \alpha \le 1, 0 < \varrho(\theta,\theta') \le 1 \right\}, \tag{45}$$

where since κ is locally characteristic we have $\|\nu\|_\kappa > 0$. In other words, for such κ and $\mathcal{T}$, a normalized dipole is characterized by a sign $s \in \{-1,1\}$, the two nodes $\theta, \theta' \in \Theta$ that satisfies $0 < \varrho(\theta,\theta') \le 1$ and a parameter $\alpha \in [0,1]$ that characterizes how balanced is the normalized dipole. This suggests the following definitions.

Definition 8. *Given a location-based family $\mathcal{T}$ and a shift-invariant kernel κ that is locally characteristic with respect to $\mathcal{T}$, the set of normalized monopoles is defined by*

$$\mathfrak{M} = \left\{ \frac{\nu}{\|\nu\|_\kappa},\ \nu = \|\pi_0\|_\kappa^{-1} s\pi_\theta; s \in \{-1,1\}, \theta \in \Theta \right\}. \tag{46}$$

The set of balanced normalized dipoles is defined by

$$\hat{\mathfrak{D}} = \left\{ \frac{\nu}{\|\nu\|_\kappa},\ \nu = \|\pi_0\|_\kappa^{-1} s(\pi_{\theta'} - \pi_\theta); s \in \{-1,1\}, 0 < \varrho(\theta,\theta') \le 1 \right\}. \tag{47}$$

In a nutshell, normalized dipoles correspond to $\alpha = 0$, while normalized balanced dipoles correspond to $\alpha = 1$. Moreover, with a slight abuse of notation we define shorthands to denote the sets of all elements $(\iota, \iota') \in \mathfrak{D}^2_{\neq}$ (i.e., of pairs of separated normalized dipoles) where each elements is resticted to be either a monopole or a balanced dipole:

$$\mathfrak{M}^2_{\neq} := \mathfrak{M}^2 \cap \mathfrak{D}^2_{\neq},\ \ \mathfrak{M} \times \hat{\mathfrak{D}}_{\neq} := (\mathfrak{M} \times \hat{\mathfrak{D}}) \cap \mathfrak{D}^2_{\neq},\ \ \hat{\mathfrak{D}}^2_{\neq} := \hat{\mathfrak{D}}^2 \cap \mathfrak{D}^2_{\neq}. \tag{48}$$

Now, we are ready to state the following result which is proved in Appendix A.5.

Proposition 4 (From normalized dipoles to normalized monopoles and balanced dipoles)**.** *Consider $\mathcal{T} = (\Theta, \rho, \mathcal{I})$ a location-based family with base distribution π_0 where $\rho(\cdot,\cdot) = \|\cdot - \cdot\|$ for some norm $\|\cdot\|$, and κ a normalized shift-invariant kernel that is locally characteristic with respect to $\mathcal{T}$. Considering the sets of (normalized) monopoles and dipoles associated to $\mathcal{T}$, and $\mathcal{A}$ a WFF sketching operator (cf Definition 6) with arbitrary frequencies $\omega_1, \dots, \omega_m$, we have*

$$\delta(\mathfrak{D}|\mathcal{A}) = \max\left(\delta(\mathfrak{M}|\mathcal{A}), \delta(\hat{\mathfrak{D}}|\mathcal{A})\right). \tag{49}$$

If in addition $\kappa \geq 0$ we also have

$$1 \leq \frac{\mu(\mathfrak{D}^2_{\neq}|\mathcal{A})}{\max(\mu(\mathfrak{M}^2_{\neq}|\mathcal{A}), \mu(\hat{\mathfrak{D}}^2_{\neq}|\mathcal{A}), \mu(\mathfrak{M} \times \hat{\mathfrak{D}}_{\neq}|\mathcal{A}))} \leq 3. \tag{50}$$

The lower bound holds regardless of the assumption $\kappa \geq 0$.

Inspecting the proof shows that $\delta(\mathfrak{D}|\mathcal{A}) \geq \max\left(\delta(\mathfrak{M}|\mathcal{A}), \delta(\hat{\mathfrak{D}}|\mathcal{A})\right)$ is valid for any sketching operator such that $\mathcal{A}\pi_\theta$ is well defined for any distribution in the family $\mathcal{T}$. Similarly the lower bound in (50) holds under this relaxed assumption. It remains open whether the converse bounds extend (possibly with weaker constants) beyond the case of WFF operators and location-based families. It also remains open whether the upper bound in (50) (or a qualitatively equivalent but larger bound) still holds without the assumption that $\kappa \geq 0$. This is left to future work, as this assumption is satisfied by all concrete kernels we will work with.

3.2.2 Expression using the supremum of certain empirical processes

The main overall consequence of Proposition 3 and Proposition 4 is that under the appropriate assumptions we have

$$\delta(\mathcal{S}_k|\mathcal{A}) \leq \frac{1}{1-c}\Big(c + \max\left(\delta(\mathfrak{M}|\mathcal{A}), \delta(\hat{\mathfrak{D}}|\mathcal{A})\right) + (6k-3)\max\left(\mu(\mathfrak{M}^2_{\neq}|\mathcal{A}), \mu(\hat{\mathfrak{D}}^2_{\neq}|\mathcal{A}), \mu(\mathfrak{M}\times\hat{\mathfrak{D}}_{\neq}|\mathcal{A}))\right)\Big) \tag{51}$$

As we now show, the advantage behind this dissection is that the study of the quantities $\delta(\hat{\mathfrak{D}}|\mathcal{A}), \mu(\mathfrak{M}^2_{\neq}|\mathcal{A}), \mu(\mathfrak{M} \times \hat{\mathfrak{D}}|\mathcal{A}), \mu(\hat{\mathfrak{D}}^2_{\neq}|\mathcal{A})$ boils down to the study of suprema of the absolute value of auxiliary functions defined as empirical means. We prove in Appendix A.6 the following result.

Proposition 5. *Consider κ a normalized shift-invariant kernel, $\mathcal{T} = (\Theta, \rho, \mathcal{I})$ a location-based family with base distribution π_0 where $\rho(\cdot,\cdot) = \|\cdot - \cdot\|$ for some norm $\|\cdot\|$, and assume that κ is locally characteristic with respect to $\mathcal{T}$. Let $\mathbf{\Omega} \in \mathbb{R}^{d\times m}$ be a matrix with arbitrary columns $\omega_1, \ldots, \omega_m$ and $\mathcal{A} = \mathcal{A}_{\mathbf{\Omega}}$ be a WFF sketching operator (cf Definition 6) with frequencies $\omega_1, \ldots, \omega_m$. With ϕ_ω defined as in (28), define for $\omega \in \mathbb{R}^d$, $x, x' \in \mathbb{R}^d$ such that $\bar{\kappa}(x) < 1$ and $y \in \mathbb{R}^d$*

$$\psi(\omega) := \frac{|\langle \pi_0, \phi_\omega\rangle|^2}{\|\pi_0\|^2_\kappa} \tag{52}$$

$$f_{\mathrm{d}}(x|\omega) := \frac{2\sin^2\left(\langle \omega, x\rangle/2\right)}{1-\bar{\kappa}(x)} \tag{53}$$

$$f_{\mathrm{mm}}(y|\omega) := \cos\left(\langle \omega, y\rangle\right) \tag{54}$$

$$f_{\mathrm{md}}(x, y|\omega) := 2\frac{\sin\left(\langle \omega, x\rangle/2\right)\sin\left(\langle \omega, y + x/2\rangle\right)}{\sqrt{2(1-\bar{\kappa}(x))}} \tag{55}$$

$$f_{\mathrm{dd}}(x_1, x_2, y|\omega) := 4\frac{\sin\left(\langle \omega, x_1/2\rangle\right)\sin\left(\langle \omega, x_2/2\rangle\right)\cos\left(\langle \omega, y + x_2/2 - x_1/2\rangle\right)}{\sqrt{2(1-\bar{\kappa}(x_1))}\sqrt{2(1-\bar{\kappa}(x_2))}}. \tag{56}$$

Denote $\mathbf{\Omega} \in \mathbb{R}^{d\times m}$ the matrix with columns ω_j, $1 \leq j \leq m$ and $\Psi_{\mathrm{m}}(\mathbf{\Omega}) := \frac{1}{m}\sum_{j=1}^{m}\psi(\omega_j)$ and for $\ell \in \{\mathrm{m}, \mathrm{d}, \mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$

$$\Psi_\ell(\cdot|\mathbf{\Omega}) := \frac{1}{m}\sum_{j=1}^{m}\psi(\omega_j)f_\ell(\cdot|\omega_j). \tag{57}$$

Denote $\Theta - \Theta := \{x - x' : (x, x') \in \Theta^2\}$. *With the sets of (normalized) monopoles and dipoles associated to* $\mathcal{T}$ *as defined in* (46), (47) *and* (48), *we have*

$$\delta(\mathfrak{M}|\mathcal{A}) = \big|1 - \Psi_{\mathrm{m}}(\boldsymbol{\Omega})\big|, \tag{58}$$

$$\text{with } \Theta_{\mathrm{d}} := \Big\{x \in \Theta - \Theta,\ 0 < \|x\| \leq 1\Big\}, \quad \text{we have } \delta(\hat{\mathfrak{D}}|\mathcal{A}) = \sup_{x \in \Theta_{\mathrm{d}}} \big|1 - \Psi_{\mathrm{d}}(x|\boldsymbol{\Omega})\big|, \tag{59}$$

$$\text{with } \Theta_{\mathrm{mm}} := \Big\{y \in \Theta - \Theta,\ 1 \leq \|y\|\Big\}, \quad \text{we have } \mu(\mathfrak{M}^2_{\neq}|\mathcal{A}) = \sup_{y \in \Theta_{\mathrm{mm}}} \big|\Psi_{\mathrm{mm}}(y|\boldsymbol{\Omega})\big|, \tag{60}$$

$$\text{there exists a set } \Theta_{\mathrm{md}} \subset \Theta_{\mathrm{d}} \times \Theta_{\mathrm{mm}},\ \text{s.t. } \mu(\mathfrak{M} \times \hat{\mathfrak{D}}_{\neq}|\mathcal{A}) = \sup_{(x,y) \in \Theta_{\mathrm{md}}} \big|\Psi_{\mathrm{md}}(x, y|\boldsymbol{\Omega})\big|, \tag{61}$$

$$\text{there exists a set } \Theta_{\mathrm{dd}} \subset \Theta_{\mathrm{d}} \times \Theta_{\mathrm{d}} \times \Theta_{\mathrm{mm}},\ \text{s.t. } \mu(\hat{\mathfrak{D}}^2_{\neq}|\mathcal{A}) = \sup_{(x_1,x_2,y) \in \Theta_{\mathrm{dd}}} \big|\Psi_{\mathrm{dd}}(x_1, x_2, y|\boldsymbol{\Omega})\big|. \tag{62}$$

NB: Since κ is locally characteristic, $\bar{\kappa}(x) < 1$ for $x \in \Theta_{\mathrm{d}}$ hence all of the above functions are well defined.

3.2.3 Lipschitz property and covering numbers

The study of the suprema of functions $\Psi_\ell(z|\boldsymbol{\Omega})$ (as defined in (57)) for random i.i.d. frequencies ω_j is classical and fits within the general theory of empirical processes. It typically relies on establishing pointwise concentration inequalities for $\Psi_\ell(z|\boldsymbol{\Omega})$ and showing that with high probability on the draw of frequencies $\boldsymbol{\Omega}$ the function $\Psi_\ell(\cdot|\boldsymbol{\Omega})$ is Lipschitz with respect to a metric Δ_ℓ on Θ_ℓ such that the covering numbers of Θ_ℓ with respect to Δ_ℓ are well controlled.

The following result establishing a Lipschitz bound is proved in Appendix A.7.

Theorem 3 (Lipschitz bound)**.** *Let* $\mathcal{T} := (\Theta, \rho, \mathcal{I})$ *be a location-based family with base distribution* π_0 *on* $\mathbb{R}^d$, *with* $\Theta \subseteq \mathbb{R}^d$ *a bounded subset,* $\rho(\cdot, \cdot) := \|\cdot - \cdot\|$ *where* $\|\cdot\|$ *is some norm on* $\mathbb{R}^d$. *Let* κ *be a non-degenerate normalized shift-invariant kernel on* $\mathbb{R}^d$. *Assume that there is some norm* $\|\cdot\|_a$ *on* $\mathbb{R}^d$ *and a function* $\tilde{\kappa} : [0, +\infty) \to \mathbb{R}$ *such that, with* $R := \sup_{x \in \Theta_{\mathrm{d}}} \|x\|_a$, *the normalized kernel* $\bar{\kappa}(x)$ *defined in* (15) *satisfies for every* $x \in \mathbb{R}^d$ *such that* $\|x\|_a \leq R$

$$\bar{\kappa}(x) = \tilde{\kappa}(\|x\|_a)\,. \tag{63}$$

Assume that the following function is of class $\mathcal{C}^1$ *on* $(0, R)$

$$\alpha : r > 0 \mapsto \alpha(r) := \frac{r}{\sqrt{1 - \tilde{\kappa}(r)}} \tag{64}$$

and that

$$C_\kappa := \sup_{0 < r \leq R} \max(1, \alpha^2(r), |\alpha'(r)|^2) < \infty. \tag{65}$$

Consider $\psi(\cdot)$ *defined as in* (52) *with* ϕ_ω *defined as in* (28), $\boldsymbol{\Omega} \in \mathbb{R}^{d \times m}$ *with arbitrary columns* $\omega_1, \ldots, \omega_m$, *and* $\Psi_\ell(\cdot|\boldsymbol{\Omega}), \Theta_\ell$ *defined as in Proposition 5 for* $\ell \in \{\mathrm{d}, \mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$. *Then, we have for each* $\ell \in \{\mathrm{d}, \mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$

$$\forall z, z' \in \Theta_\ell,\ |\Psi_\ell(z|\boldsymbol{\Omega}) - \Psi_\ell(z'|\boldsymbol{\Omega})| \leq 6\Psi_0(\boldsymbol{\Omega}) \cdot C_\kappa \cdot\ \Delta_\ell(z, z'), \tag{66}$$

where

$$\Psi_0(\boldsymbol{\Omega}) := \frac{1}{m}\sum_{j=1}^{m} \psi(\omega_j) f_0(\omega_j), \quad f_0(\omega) := \sum_{i=1}^{3} \|\omega\|_{a,\star}^{i}. \tag{67}$$

and the metrics Δ_ℓ *on the domains* Θ_ℓ, $\ell \in \{\mathrm{d}, \mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$ *are defined as*

$$\Delta_{\mathrm{d}}(x, x') := |\|x\|_a - \|x'\|_a| + \left\| \frac{x}{\|x\|_a} - \frac{x'}{\|x'\|_a} \right\|_a, \tag{68}$$

$$\Delta_{\mathrm{mm}}(y, y') := \|y - y'\|_a, \tag{69}$$

$$\Delta_{\mathrm{md}}((x, y), (x', y')) := \Delta_{\mathrm{d}}(x, x') + \Delta_{\mathrm{mm}}(y, y'), \tag{70}$$

$$\Delta_{\mathrm{dd}}((x_1, x_2, y), (x_1', x_2', y')) := \Delta_{\mathrm{d}}(x_1, x_1') + \Delta_{\mathrm{d}}(x_2, x_2') + \Delta_{\mathrm{mm}}(y, y'), \tag{71}$$

Covering numbers are controlled using the following result established in Appendix A.7.4.

Proposition 6 (Covering numbers)**.** *Define* $D := \mathrm{diam}_a(\Theta) = \sup_{x \in \Theta - \Theta} \|x\|_a$. *For every* $\tau > 0$, *and for each* $\ell \in \{\mathrm{d}, \mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$, *we have*

$$\mathcal{N}_\ell(\tau) := \mathcal{N}(\Theta_\ell, \Delta_\ell, \tau) \leq \big(1 + 64(D+1)/\tau\big)^{3d+2}. \tag{72}$$

3.3 Results for random sketching

In this section, we leverage Section 3.2 to establish RIP results for random sketching. We first establish a generic result before exploiting it for specific examples and showing that it allows to improve upon and to extend existing work from the literature.

Theorem 4. *Consider* $\mathcal{T} = (\Theta, \rho, \mathcal{I})$, $\kappa \geq 0$ *and* $\|\cdot\|_a$ *satisfying the assumptions of Theorem 3, and* C_κ *as in (65). Assume that* κ *has its mutual coherence with respect to* $\mathcal{T}$ *bounded by* $0 < \mu < 1/10$. *Let* $k \geq 1$ *be an integer such that* $1 \leq k < \frac{1}{10\mu}$, *and denote* $c := (2k-1)\mu$. *Let* w *be a* κ*-compatible weight function (cf Definition 7). Consider an integer* $m \geq 1$ *and* $\boldsymbol{\Omega} \in \mathbb{R}^{d \times m}$ *a random matrix (possibly with non i.i.d. columns* $\omega_1, \ldots, \omega_m$*), such that the average marginal density of the* ω_j*'s satisfies (33). Denote* $\Psi_\ell(\cdot|\boldsymbol{\Omega}), \Theta_\ell$ *as in Proposition 5 for* $\ell \in \{\mathrm{d}, \mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$, *and* $\Psi_0(\boldsymbol{\Omega})$ *as in Theorem 3.*

Given any $M > 0$, $0 < \tau < 1 - 5c$, $v > 0$, *if the following inequalities hold*

$$\mathbb{P}\Big(\Psi_0(\boldsymbol{\Omega}) > M\Big) \leq 2\exp\Big(-\frac{m}{v}\Big), \tag{73}$$

$$\mathbb{P}\Big(|\Psi_{\mathrm{m}}(\boldsymbol{\Omega}) - \mathbb{E}\Psi_{\mathrm{m}}(\boldsymbol{\Omega})| > \frac{\tau}{4}\Big) \leq 2\exp\Big(-\frac{m}{v}\Big), \tag{74}$$

$$\forall z \in \Theta_{\mathrm{d}}, \ \mathbb{P}\Big(|\Psi_{\mathrm{d}}(z|\boldsymbol{\Omega}) - \mathbb{E}\Psi_{\mathrm{d}}(z|\boldsymbol{\Omega})| > \frac{\tau}{8}\Big) \leq 2\exp\Big(-\frac{m}{v}\Big), \tag{75}$$

$$\forall \ell \in \{\mathrm{mm}, \mathrm{md}, \mathrm{dd}\}, \ \forall z \in \Theta_\ell, \ \mathbb{P}\Big(|\Psi_\ell(z|\boldsymbol{\Omega}) - \mathbb{E}\Psi_\ell(z|\boldsymbol{\Omega})| > \frac{\tau}{16k}\Big) \leq 2\exp\Big(-\frac{m}{v}\Big), \tag{76}$$

then the w-FF sketching operator $\mathcal{A}$ *(cf Definition 6) with frequencies* $\omega_1, \ldots, \omega_m$ *satisfies*

$$\mathbb{P}\left(\delta(\mathcal{S}_k|\mathcal{A}) > \frac{4c+\tau}{1-c}\right) \leq 12 \cdot \exp\Big(-\frac{m}{v}\Big)(1 + C/\tau)^{3d+2}, \tag{77}$$

where

$$C := 6144 C_\kappa M \cdot k(1 + \mathrm{diam}_a(\Theta)). \tag{78}$$

Under the assumption of the theorem, a consequence is that for any $0 < \eta < 1$

$$m \geq v(3d+2)\log(1 + C/\tau) + \log(10/\eta) \implies \mathbb{P}\left(\delta(\mathcal{S}_k|\mathcal{A}) > \frac{4c+\tau}{1-c}\right) \leq \eta. \tag{79}$$

hence estimating the order of magnitude of v, C and τ satisfying the assumptions of the theorem is key to estimate a sufficient sketch size. The proof of Theorem 4 is given in Appendix A.8. Our next result establishes the concentration inequalities (74)-(75)-(76) under a sub-exponentiality assumption on functions associated to the random frequencies ω_j, $1 \leq j \leq m$.

Definition 9. *A real-valued random variable is sub-exp(ν, β), where $\nu, \beta \geq 0$, if*

$$\mathbb{E}e^{\lambda(X - \mathbb{E}X)} \leq e^{\frac{\lambda^2\nu^2}{2}}, \ \ \forall |\lambda| \leq \frac{1}{\beta}. \tag{80}$$

The case $\beta = 0$ corresponds to a sub-Gaussian variable.

If X is sub-exp(ν, β) then by the standard Cramér-Chernoff method[6]

$$\forall t > 0, \ \ \mathbb{P}\Big(\big|X - \mathbb{E}X\big| > t\Big) \leq 2\max\left(e^{-\frac{t^2}{2\nu^2}}, e^{-\frac{t}{\beta}}\right) \leq 2\exp\left(-\frac{t^2}{2\nu^2 + \beta t}\right). \tag{81}$$

We establish the inequalities (74)-(75)-(76) by showing that $\Psi_{\mathrm{m}}(\boldsymbol{\Omega})$ and $\Psi_\ell(z|\boldsymbol{\Omega})$, $\ell \in \{\mathrm{d}, \mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$, $z \in \Theta_\ell$, are sub-exponential with controlled expectations. =A well-known property of sub-exp random variables is that if $X_1, \ldots, X_m$ are independent sub-exp(ν, β) then $\frac{1}{m}\sum_{j=1}^m X_j$ is sub-exp$(\nu/\sqrt{m}, \beta/m)$. Thus, when the frequencies $\omega_j \sim \Lambda$, $1 \leq j \leq m$ are i.i.d. random variables, in order to prove that $\Psi_m(\boldsymbol{\Omega})$ is sub-exp$(\nu/\sqrt{m}, \beta/m)$, it is enough to prove that the random variables $\psi(\omega_j)$, $1 \leq j \leq m$, are sub-exp(ν, β). Similarly, for $\ell \in \{\mathrm{d}, \mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$ and $z \in \Theta_\ell$, in order to prove that $\Psi_\ell(z|\boldsymbol{\Omega})$ is sub-exp$(\nu/\sqrt{m}, \beta/m)$ it is enough to prove that the random variables $\psi(\omega)$ and $\psi(\omega) f_\ell(z|\omega)$, $\omega \sim \Lambda$, are sub-exp(ν, β). For this purpose, the following lemma (proved in Appendix A.9.1) will be crucial.

Lemma 2. *Under the assumptions of Theorem 5 (see below), for each $\ell \in \{\mathrm{d}, \mathrm{mm}, \mathrm{md}\}$ and $z \in \Theta_\ell$, there exists $x_0' \in \mathbb{R}^d$ satisfying $\|x_0'\|_a = 1$ such that*

$$|f_\ell(z|\omega)| \leq (\sqrt{C_\kappa}|\langle \omega, x_0'\rangle|)^{p_\ell}, \quad \forall \omega \in \mathbb{R}^d, \quad \textit{with } p_{\mathrm{d}} = 2, \ p_{\mathrm{mm}} = 0, \ p_{\mathrm{md}} = 1. \tag{82}$$

Moreover, for each $z \in \Theta_{\mathrm{dd}}$, there are $x_i' \in \mathbb{R}^d$ such that $\|x_i'\|_a = 1$, $i = 1, 2$ and

$$|f_{\mathrm{dd}}(z|\omega)| \leq \frac{C_\kappa}{4}\Big(\langle \omega, x_1'\rangle^2 + \langle \omega, x_2'\rangle^2\Big), \quad \forall \omega \in \mathbb{R}^d.$$

Using Lemma 2, we obtain that the random variables $\psi(\omega) f_\ell(z|\omega)$ are (almost surely) bounded by random variables of the form $\psi(\omega)(\sqrt{C_\kappa}|\langle \omega, x\rangle|)^p$, with $x \in \mathbb{R}^d$ and $p \in \{0, 1, 2\}$, allowing to leverage the following lemma proved in Appendix A.9.2.

Lemma 3. *Consider real-valued random variables X, Y where Y is sub-exp(ν, β) and $|X| \leq Y$ almost surely. Then X is sub-exp(ν', β), where*

$$\nu' := \sqrt{2\nu^2 + 16(\mathbb{E}(Y))^2}. \tag{83}$$

[6] See e.g. [BLM13a, Section 2.2] and the proof of Theorem 2.8.1. in [Ver18].

The following theorem considers a slightly generalized case with block-i.i.d. variables, covering structured random features.

Theorem 5. *Consider $\mathcal{T} := (\Theta, \rho, \mathcal{I})$ a location-based family with base distribution π_0 on $\mathbb{R}^d$, with $\Theta \subseteq \mathbb{R}^d$ a bounded subset, $\rho(\cdot,\cdot) := \|\cdot - \cdot\|$ where $\|\cdot\|$ is some norm on $\mathbb{R}^d$. Let $\kappa \geq 0$ be a non-degenerate normalized shift-invariant kernel on $\mathbb{R}^d$, and assume that there is some norm $\|\cdot\|_a$ on $\mathbb{R}^d$ and a function $\tilde{\kappa} : [0, +\infty) \to \mathbb{R}$ such that, with $R := \sup_{x \in \Theta_d} \|x\|_a$, the normalized kernel $\bar{\kappa}(x)$ defined in (15) satisfies $\bar{\kappa}(x) = \tilde{\kappa}(\|x\|_a)$ for every $x \in \mathbb{R}^d$ such that $\|x\|_a \leq R$. Assume that the function α defined in (64) is of class $\mathcal{C}^1$ on $(0, R)$ and the constant C_κ defined in (65) is finite. Moreover, assume that κ has its mutual coherence with respect to $\mathcal{T}$ bounded by μ where $0 < \mu < \frac{1}{10}$. Let $1 \leq k < \frac{1}{10\mu}$ and define $c := (2k-1)\mu$.*

Let w be a κ-compatible weight function and m be an integer multiple of $b \in \mathbb{N}^*$, and consider $\mathcal{A}$ a w-FF sketching operator (Definition 6) associated to the frequencies $(\omega_1, \ldots, \omega_m)$ that are block-i.i.d. corresponding to m/b i.i.d. $d \times b$ random matrices $\boldsymbol{B}_i$, $1 \leq i \leq m/b$ such that (33) holds. Let $\tau \in (0, 1 - 5c)$, and assume that

1. *there exists $\nu, B > 0$ and $\beta \geq 0$ such that for each $x \in \mathbb{R}^d$ satisfying $\|x\|_a = 1$, the following random variables are sub-exp(ν, β) with $|\mathbb{E}(Z_p)| \leq B$*

$$Z_p := \frac{1}{b} \sum_{j=1}^{b} \psi(\omega_j)(\sqrt{C_\kappa}|\langle \omega_j, x \rangle|)^p, \quad p \in \{0, 1, 2\}. \tag{84}$$

2. *there exists $M > 0$ such that*

$$\mathbb{P}\Big(\Psi_0(\boldsymbol{\Omega}) > M\Big) \leq 2 \exp\Big(-\frac{m}{v}\Big). \tag{85}$$

$$v := \frac{256k^2 b(2\nu'^2 + \beta\tau)}{\tau^2}, \quad \text{where } \nu' := \sqrt{2}\sqrt{\nu^2 + 8B^2}. \tag{86}$$

Then $\mathcal{A}$ satisfies

$$\mathbb{P}\left(\delta(\mathcal{S}_k|\mathcal{A}) > \frac{4c + \tau}{1 - c}\right) \leq 12 \cdot \exp\Big(-\frac{m}{v}\Big)(1 + C/\tau)^{3d+2}, \tag{87}$$

where

$$C := 6144 C_\kappa M \cdot k(1 + \mathrm{diam}_a(\Theta)). \tag{88}$$

When $B_\psi := \sup_{\omega \in \mathbb{R}^d} \psi(\omega) < +\infty$, (87) also holds with v from (86) by replaced with

$$v' := \frac{256k^2 b(2B_\psi^2 \nu'^2 + B_\psi \beta\tau)}{\tau^2} \tag{89}$$

if we assume (85) with v' instead of v and replace Item 1 by the same assumption on the random variables

$$Z'_p := \frac{1}{b} \sum_{j=1}^{b} (\sqrt{C_\kappa}|\langle \omega_j, x \rangle|)^p, \quad p \in \{0, 1, 2\}, \tag{90}$$

Theorem 5 is obtained by applying Theorem 4, see Appendix A.9.3.

Next we give two examples: for mixtures of Gaussians, Item 2 is established using sub-exponentiality; for mixtures of Diracs, Item 2 requires a bit more work.

The case of a mixture of Gaussians. We consider the kernel κ and the overall setting of Example 2, and a w-RFF sketching operator $\mathcal{A}$ with "flat" κ-compatible weight function $w \equiv 1$ and i.i.d. frequencies $\omega \sim \mathcal{N}(0, \mathbf{\Sigma}^{-1}/s^2)$. In this setting, the function ψ defined in (52) satisfies[7]

$$\forall \omega \in \mathbb{R}^d, \;\; \psi(\omega) = \frac{|\langle \pi_0, \phi_\omega \rangle|^2}{\|\pi_0\|_\kappa^2} = \frac{e^{-\omega^{\mathrm{T}} \mathbf{\Sigma} \omega}}{(1+2s^{-2})^{-\frac{d}{2}}}. \tag{91}$$

Given the definition (67) of $f_0(\cdot)$, we deduce that $\psi(\omega)$ and $\psi(\omega) f_0(\omega)$ are bounded, so that Hoeffding's inequality yields Item 2 for an explicit M independent of m and for any $v > 0$, while the variant of Item 1 with Z'_p follows from the sub-exponentiality of $|\langle \omega, x \rangle|^2$, $s \in \{0, 1, 2\}$, since $\omega \sim \mathcal{N}(0, \mathbf{\Sigma}^{-1}/s^2)$. Details are given in Appendix A.9.4, including sub-exponentiality constants and a proof that $\mathcal{T}$ and κ satisfy the assumptions of Theorem 5. When all is said and done, we obtain the following result.

Corollary 1. *Consider $\Theta \subseteq \mathbb{R}^d$, an integer $k \geq 1$, a scale $s > 0$, and*

$$\epsilon \geq \sqrt{2+s^2}(4\sqrt{\log(5ek)}). \tag{92}$$

With $\mathcal{T}$, κ, $\mathbf{\Sigma}$ as in Example 2 and $\mathcal{A}$ the w-RFF sketching operator with "flat" κ-compatible weight function $w \equiv 1$ and m i.i.d. frequencies $\omega_j \sim \Lambda := \mathcal{N}(0, \mathbf{\Sigma}^{-1}/s^2)$, the mutual coherence of κ with respect to $\mathcal{T}$ is bounded by μ where $0 < \mu < \frac{1}{10k}$. Moreover, for each $0 < \tau < 1 - 5c$, where $c := (2k-1)\mu$, we have

$$\mathbb{P}\left(\delta(\mathcal{S}_k | \mathcal{A}) > \frac{4c+\tau}{1-c} \right) \leq 12 \exp\left(-\frac{m}{v} \right) (1 + C/\tau)^{3d+2}, \tag{93}$$

where $v = v_k(\tau)$ with

$$v_k(\tau) := 512k^2 \left((C_0/\tau)^2 + \frac{1}{3}(C_0/\tau) \right), \;\; \text{with } C_0 \leq 7\sqrt{3}\epsilon^2 s^{-2}(1+2s^{-2})^{d/2}, \tag{94}$$

$$C \leq \left(43000\epsilon^2 (1+2s^{-2})^{d/2} \right) \cdot k(1 + \mathrm{diam}_a(\Theta)). \tag{95}$$

In contrast to [GBKT21b, Theorem 6.11], the RIP constant here is not guaranteed to be (with high probability) arbitrarily close to zero, but only smaller than the quantity $(4c+\tau)/(1-c)$, which can be made arbitrarily close to $4c/(1-c) < 1$ (since $c = (2k-1)\mu < (2k-1)/10k < 1/5$). The assumption (92) relating the separation parameter ϵ and the scale parameter s is essential to guarantee that $10k\mu < 1$. This technical condition is important since our bounds are only valid under the assumption that $5c = 5(2k-1)\mu < 1$. In particular, we deduce that the probability that the event $\{\delta(\mathcal{S}_k|\mathcal{A}) \leq (4c+\tau)/(1-c)\}$ holds is larger than $1-\eta$, with $\eta \in]0, 1]$, whenever

$$m \geq (3d+2) v_k(\tau) \log\left(1 + C/\tau\right) + \log(12/\eta).$$

In other words, a sufficient sketch size m scales as $\mathcal{O}(d v_k(\tau) \log(C/\tau))$. Typically, we seek to determine the dependency of the sketch size in terms of the sparsity k and the dimension d. Considering a near minimum separation parameter ϵ according to (92), a fixed diameter, and $1 \leq \log k = \mathcal{O}(d)$, let us highlight two regimes:

[7] See [GBKT21b, Section 6.3.1].

1. the regime $\sqrt{d} \lesssim s = \mathcal{O}(\texttt{poly}(d))$: then $(1+2s^{-2})^{d/2} = \mathcal{O}(1)$ so that $C_0 = \mathcal{O}(\log k)$, $v_k(\tau) = \mathcal{O}((\tau^{-1}k\log k)^2)$ and $\log(C) = \mathcal{O}(\log(kd))$ and the the sufficient sketch size scales as $\mathcal{O}((\tau^{-1}k\log k)^2 \log kd)$.
2. the regime where s is of the order of one: then $(1+2s^{-2})^{d/2} = \mathcal{O}(e^{cd/2})$, with $c = \log(1+2s^{-2})$ so that $v_k(\tau) = \mathcal{O}((\tau^{-1}k\log k)^2 d e^{cd})$ and $\log(C) = \mathcal{O}(d)$ so that the sufficient sketch size scales as $\mathcal{O}((\tau^{-1}k\log k)^2 d^2 e^{cd})$;

In both regimes, we obtain similar results as in [GBKT21b, Table 1]. There exists an intermediate regime, $c_1 d^{1/4} \leq s^2 \leq c_2\sqrt{d}$, for large d, where we expect that Theorem 5 can be leveraged to obtain better sketch size estimates, that would not be achievable with the techniques of [GBKT21b]. A closer inspection of our proof techniques indeed suggests that better dependencies can be obtained by relying on Item 1 of Theorem 5 (with Z_p) rather than on its variant with Z'_p. Concretely, this means obtaining better sub-exponentiality constants for the random variables $\psi(\omega)|\langle \omega, x\rangle|^p$.

As an example, for $p = 0$, the proof given in Appendix A.9.4 only relies on the crude uniform deterministic bound $\psi(\omega) \leq B_\psi := \sup_{\omega\in\mathbb{R}^d} \psi(\omega) = (1+2s^{-2})^{d/2}$, hence it cannot yield a better result than Hoeffding's bound [Hoe94]

$$\forall \epsilon > 0,\ \ \mathbb{P}\Big(\psi(\omega) - \mathbb{E}\psi(\omega) > \epsilon\Big) \leq \exp\Big(-\frac{2\epsilon^2}{B_\psi^2}\Big).$$

It is well known that exploiting the variance of $\psi(\omega)$, $V_\psi := \mathbb{V}\psi(\omega)$ can lead to better results using Bernstein's concentration inequality [BLM13a, Theorem 2.10]

$$\forall \epsilon > 0,\ \ \mathbb{P}\Big(\psi(\omega) - \mathbb{E}\psi(\omega) > \epsilon\Big) \leq \exp\Big(-\frac{\epsilon^2}{2(V_\psi + B_\psi\epsilon)}\Big).$$

In the setting of Corollary 1, since $\omega \sim \Lambda$ is Gaussian, we can compute explicitly $\mathbb{E}_{\omega\sim\Lambda}\psi(\omega) = 1$ and $V_\psi := \mathbb{V}\psi(\omega) \leq \mathbb{E}_{\omega\sim\Lambda}\psi^2(\omega) = (1+4s^{-4}(1+4s^{-2})^{-1})^{d/2}$. Thus $\log(V_\psi) = d\log(1+4s^{-4}(1+4s^{-2})^{-1})/2$, while $\log(B_\psi) = d\log(1+2s^{-2})/2$, hence in the regime $c_1 d^{1/4} \leq s^2 \leq c_2\sqrt{d}$ we have $V_\psi = \mathcal{O}(1)$ while $B_\psi \geq e^{c_3\sqrt{d}}$. Empirical experiments further suggest that even this Bernstein bound remains crude, and even essentially vacuous (on the order of one): this is illustrated on Figure 1, as well the behavior of the following *conjectured* upper bound

$$\mathbb{P}\Big(\psi(\omega) - \mathbb{E}\psi(\omega) > \epsilon\Big) \leq \exp\Big(-\frac{\epsilon^2}{2(V_\psi + \sqrt{V_\psi}\epsilon)}\Big). \qquad \text{(Conjecture)}$$

which remains to be proved for a range of $0 < \epsilon < \epsilon_{\max}$ to be determined. Proving this conjecture and similar ones for $\langle \omega, x\rangle^p \psi(\omega)$, $p = 1, 2$ would lead to definite improvements to the bounds derived in Corollary 1 and consecrate the advantage of the analysis developed in this work over the analysis given in [GBKT21a], justifying the supplementary assumption on the RIP constant (to be larger than $4c/(1-c)$) that our analysis requires. This is however left to future work.

The case of a mixture of Diracs We now consider the setting of Example 1, and a w-RFF sketching operator $\mathcal{A}$ corresponding to the "flat" κ-compatible weight function $w \equiv 1$ and i.i.d. frequencies $\omega \sim \mathcal{N}(0, s^{-2}\mathbb{I}_d)$. In this setting, the function ψ defined in (52) satisfies[8]

$$\forall \omega \in \mathbb{R}^d,\ \ \psi(\omega) = \frac{|\langle \pi_0, \phi_\omega\rangle|^2}{\|\pi_0\|_\kappa^2} = 1. \tag{96}$$

[8] See again [GBKT21b, Section 6.3.1].

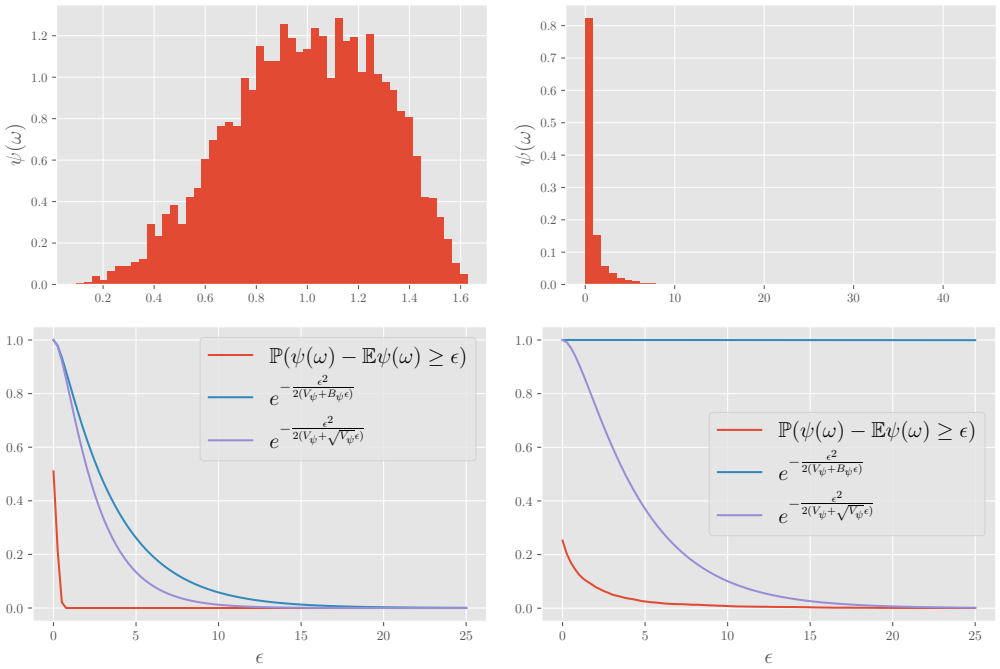

Figure 1: (top) Histogram of $\psi(\omega)$ when $\omega \sim \mathcal{N}(0, s^{-2}\mathbb{I}_d)$; (bottom) Empirical graph of $\mathbb{P}(\psi(\omega) - \mathbb{E}\psi(\omega) \geq \epsilon)$ and two candidate analytic bounds for $s = 3$ and $d = 5$ (left), $d = 100$ (right).

This is a setting where the analysis adopted in [GBKT21b] cannot be applied, since the condition (43) does not hold. Indeed, by (96) and by the fact that $\|\pi_0\|_\kappa = 1$ we have $\sup_{\omega \in \mathbb{R}^d} |\langle \phi_\omega^w, \pi_0 \rangle| \cdot \max(1, \|\omega\|_\star, \|\omega\|_\star^2) = +\infty$.

Reasoning as in the case of a mixture of Gaussians allows to establish the variant of Item 1 with Z'_p required in Theorem 5. However, in this setting, $\psi(\omega) f_0(\omega)$ is no longer sub-exponential hence establishing Item 2 in Theorem 5 requires more work and a choice of $M > 0$ that depends at most polynomially on the sketch size m. As detailed in Appendix A.9.4, we get the following result as a corollary of Theorem 5.

Corollary 2. *Consider $\Theta \subseteq \mathbb{R}^d$, an integer $k \geq 1$, a scale $s > 0$, and*

$$\epsilon \geq s(4\sqrt{\log(5ek)}). \tag{97}$$

With $\mathcal{T}$, κ, $\mathbf{\Sigma}$ as in Example 1 and $\mathcal{A}$ the w-RFF sketching operator with "flat" κ-compatible weight function $w \equiv 1$ and m i.i.d. frequencies $\omega_j \sim \Lambda := \mathcal{N}(0, s^{-2}\mathbb{I}_d)$, where $\mathbb{I}_d$ is the identity matrix of dimension d, the mutual coherence of κ with respect to $\mathcal{T}$ is bounded by μ where $0 < \mu < \frac{1}{10k}$.
Moreover, for each $0 < \tau < 1 - 5c$, where $c := (2k-1)\mu$, we have

$$\mathbb{P}\left(\delta(\mathcal{S}_k|\mathcal{A}) > \frac{4c + \tau}{1 - c}\right) \leq 12 \exp\left(-\frac{m}{v}\right)(1 + C(\tau, m)/\tau)^{3d+2}, \tag{98}$$

where

$$v = v_k(\tau) := 512k^2 \left((C_0/\tau)^2 + \frac{1}{3}(C_0/\tau) \right), \ \textit{with } C_0 \leq 7s^{-2} \max(1, \sqrt{3}\epsilon^2), \tag{99}$$

$$C(\tau, m) := \left(6144(1 + 2s^{-1})^3 \max(1, \sqrt{3}\epsilon^2)(2d^{3/2} + \sqrt{m}\tau^{3/2}) \right) \cdot k(1 + \mathrm{diam}_a(\Theta)). \tag{100}$$

As in the case of Gaussian mixtures, and in contrast to [GBKT21b, Theorem 6.11], the RIP constant is not guaranteed to be (with high probability) arbitrarily close to zero: it can only be made arbitrarily close to $4c/(1-c) < 1$. This is the price we pay for being able to handle plain importance sampling with $w \equiv 1$. Observe that $1 + C(\tau, m)/\tau \leq \sqrt{m}(1 + C(\tau, 1)/\tau) \leq m(1 + C(\tau, 1)/\tau)$ hence the r.h.s. of (98) is upper bounded by

$$12 \exp\left(-\frac{m}{v}\left(1 - (3d+2)v\frac{\log(m)}{m}\right) + (3d+2)\log\left(1 + C(\tau,1)/\tau\right)\right). \tag{101}$$

We deduce that for $\eta \in (0, 1]$, and for the $w - RFF$ sketching operator described in Corollary 2, the probability that the event $\left\{\delta(\mathcal{S}_k|\mathcal{A}) \leq (4c + \tau)/(1 - c)\right\}$ holds is larger than $1 - \eta$, as soon as

$$\frac{m}{\log m} \gtrsim (3d+2)v \log\left(1 + C(\tau,1)/\tau\right) + \log(12/\eta) = \Omega(k^2 d).$$

In other words, a sufficient sketch size is $\mathcal{O}(k^2 d)$: our analysis allows to obtain the same dependencies on k and d as the analysis developed in [GBKT21b] but *without* assuming the conditions (43) that imposes constraints on the importance sampling weight w. It would be tempting to think that a judicious choice of the weight function w would allow to further improve the dependency on k of the sketch size from $\mathcal{O}(k^2 d)$ to $\mathcal{O}(kd)$. Unfortunately, as we shall see in Section 3.4, our analysis does not allow us to make such an improvement. The investigation of the role of the weight function is deferred for future work. Finally, observe that the constant $\log(1 + C(\tau, 1)/\tau)$ depends logarithmically on D (the diameter of Θ), on s^{-1}, and on $d \log k$. Note that the logarithmic dependency on D is well known in the Fourier features literature [SS15a], while the dependency on s^{-1} has empirical implications: the parameter s should be choosen small enough so that the mutual coherence of κ with respect to $\mathcal{T}$ is bounded by μ, yet not too small. Interestingly, this phenomenon was documented in several empirical investigations [KTTG17, Cha20].

3.3.1 Towards bounds for structured random sketching

A benefit of the theoretical analysis presented in this work is to pave the way to theoretical guarantees on the RIP of sketching operators based on structured features. Indeed, as evoked in Section 2.5, there are now many constructions of structured random Fourier features where m is a multiple of d and the frequencies $\omega_1, \ldots, \omega_m$ are block-i.i.d. with blocks of size $b = d$: the matrix $\boldsymbol{\Omega} \in \mathbb{R}^{d \times m}$ with columns ω_j, $1 \leq j \leq m$ is the concatenation of i.i.d. random matrices $\boldsymbol{B}_i \in \mathbb{R}^{d \times d}$, $1 \leq i \leq m/d$.

Such constructions can be designed to lead to (non independent but) identically distributed Gaussian frequencies $\omega_j \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$, so that the average marginal density satisfies (33) with a Gaussian kernel and the simplest κ-compatible weight function $w \equiv 1$, see e.g. [Cha20, Chapter 5]. This allows to reuse the proof techniques

used to establish Corollary 1-Corollary 2, showing as an intermediate step that each $X_j := \psi(\omega_j)(\sqrt{C_\kappa}|\langle\omega_j, x\rangle|)^p$ (resp. $X_j := (\sqrt{C_\kappa}|\langle\omega_j, x\rangle|)^p$, when $B_\psi < +\infty$), with arbitrary $x \in \mathbb{R}^p$ satisifying $\|x\|_a = 1$ and $p \in \{0, 1, 2\}$, is sub-exp(ν, β) with $|\mathbb{E}X| \leq B$.

If the variables X_j were independent, we would deduce that their blockwise-empirical average Z_p defined in (84) (resp. Z'_p from (90)), which satisfies $|\mathbb{E}Z_p| \leq B$, is sub-exp$(\nu/\sqrt{d}, \beta/d)$. Even though the X_i's are not independent, we can still prove that Z_p (resp. Z'_p) satisfy the first assumption (Item 1) of Theorem 5. Indeed, by Lemma 9, the random variable $d \times Z_p = \sum_{j=1}^d \psi(\omega_j)(\sqrt{C_\kappa}|\langle\omega_j, x\rangle|)^p$ is sub-exp$(d\nu, d\beta)$, since it is a sum of d random variables that are sub-exp(ν, β), thus Z_p (and similarly Z'_p) is sub-exp(ν, β).

In the case of mixtures of Gaussians, the variables $\psi(\omega_j)f_0(\omega_j)$ are bounded and the very same reasoning as in the proof of Corollary 1, yields that (85) holds true for the same constant M and any $v > 0$. Gathering all of the above shows that for Gaussian mixture models, sketching with the considered structured random Fourier features satisfies the RIP: (93) holds with $v = dv_k(\tau)$, where $v_k(\tau)$ given in (94). The price we pay for such a quick analysis is an additional d factor in the sufficient sketch size m.

While this result does not assume the independence of the random variables X_j, it comes with a cost: settling for the use of Lemma 9 worsens the 'variance' term (86) by a factor d compared to the fully i.i.d. setting. Thus, a refined analysis of the expectation and of the sub-exponentiality constants of the random variables Z_p is required in order to prove competitive bounds on sketch sizes for structured sketching. Ideally, we may hope to prove that Z_p and Z'_p are sub-exp(ν, β) with $\nu = \mathcal{O}(1/\sqrt{d})$ and $\beta = \mathcal{O}(1/d)$ hold for some families of structured random matrices. Note however that this would require to slightly sharpen the bound obtained in Theorem 5. Indeed, the main bottleneck in the sketch size would be the constant ν' defined in (86) that involves the term B which is a constant that does not depend on d. Improving Theorem 5 would either require to refine Lemma 3 in order to circumvent the presence of this constant in ν', or to more directly rely on Theorem 4 and in establishing autonomous concentration bounds. This is left to future work. Finally, to handle the case of mixtures of Diracs, one would need to revisit Proposition 11 to control the behavior of $\Psi_0(\mathbf{\Omega})$ in the structured setting.

3.4 Lower bounds

To conclude, we provide several lower bounds that complete the picture established in this section.

First, we show that condition (43), which is known to be sufficient to control the covering numbers appearing in Theorem 2, is indeed close to necessary for these covering numbers to be well-defined and finite. This shows that existing theory (such as [GBKT21b, Theorem 5.13]) is simply too restrictive to provide guarantees for perhaps the most natural setting where there is "no" importance sampling, i.e., $w \equiv 1$, which is in contrast covered by our new results.

Second, we investigate the gap between sufficient sketch sizes endowed with theoretical guarantees, which scale as $O(k^2 d)$, and practically observed sketch sizes, which scale as $O(kd)$. We demonstrate that a proof route which could seem natural to bridge this gap is in fact a dead-end, leaving possible improvements to further work.

3.4.1 Lower bounds on variance terms

The empirical investigations in [KTTG17] showed that a practically sufficient sketch size scales as $\mathcal{O}(dk)$ compared to the theoretically sufficient sketch size $\mathcal{O}(dk^2)$ obtained by the analysis given in [GCK+20] and the analysis given in this work. This suggest that there is still room for improvement on the theoretical bounds of sketching. We investigate below theoretical approaches that may seem natural ways to improve the proof techniques respectively introduced in [GCK+20] and in this work. Our main conclusion is that these approaches *cannot* lead to the desired explanation of the empirical findings of [KTTG17].

Limits of the proof technique of [GCK+20] After a careful examination of the proof given in [GCK+20], it may be tempting to improve the concentration inequality (40) and target one of the form

$$\forall \tau > 0, \quad \sup_{\nu \in \mathcal{S}_k} \mathbb{P}\Big(\big| \|\mathcal{A}\nu\|_2^2 - 1 \big| > \frac{\tau}{2} \Big) \leq 2 \exp\Big(- \frac{m}{v_0(\tau)} \Big),$$

with $v_0(\tau)$ *independent of* k (under appropriate incoherence assumptions on κ, that depend of k). This would indeed easily provide the desired result by combining the technical ingredients as in the proof of Theorem 2, however under standard assumptions[9] on the growth of $v_0(\tau)$ when $\tau \to \infty$, it is a classical exercice[10] to show that this implies bounded moments $\mathbb{E}\|\mathcal{A}\nu\|_2^{2q}$, for $q \geq 2$, depending only on $v_0(\tau)$ and m, in particular this would also imply the existence of a constant $C > 0$, *independent of* k, such that

$$\forall \nu \in \mathcal{S}_k, \ \ \mathbb{V}\|\mathcal{A}\nu\|_2^2 \leq \frac{C}{m}.$$

where $\mathbb{V}(\cdot)$ denotes the variance of a scalar random variable. However, as we now show, under typical assumptions on the $2k$-coherence of the kernel, this variance grows linearly with k.

We begin with a technical lemma proved in Appendix A.1.

Lemma 4. *Consider a normalized shift-invariant kernel κ and $\pi_0 \in \mathcal{P}(\mathbb{R}^d)$. If $w : \mathbb{R}^d \to (0, +\infty)$ is κ-compatible and satisfies*

$$\int \big| \langle \phi_\omega^1, \pi_0 \rangle \big|^4 w^{-2}(\omega) \hat{\kappa}(\omega) \mathrm{d}\omega < +\infty, \tag{102}$$

then the following shift-invariant kernel is well-defined

$$\kappa_w(\theta, \theta') := \int_{\mathbb{R}^d} \big| \langle \phi_\omega^1, \pi_0 \rangle \big|^4 w^{-2}(\omega) \hat{\kappa}(\omega) e^{2\pi i \omega^\top (\theta - \theta')} \mathrm{d}\omega, \quad \theta, \theta' \in \mathbb{R}^d \tag{103}$$

and satisfies $\kappa_w(0,0) \geq \|\pi_0\|_\kappa^4$. The following weight function is κ-compatible and satisfies (102)

$$w_0(\omega) := \|\pi_0\|_\kappa^{-1} \cdot |\langle \phi_\omega^1, \pi_0 \rangle|. \tag{104}$$

Moreover $\kappa_{w_0}(0,0) = \|\pi_0\|_\kappa^4$ and more generally

$$\kappa_{w_0}(\theta, \theta') = \|\pi_0\|_\kappa^2 \langle \pi_\theta, \pi_{\theta'} \rangle_\kappa, \quad \theta, \theta' \in \mathbb{R}^d. \tag{105}$$

[9] A subgaussian tail or a sub-exponential tail.
[10] See Theorem 2.3 in [BLM13b] and Lemma 1 in [BKZ20].

In the special case of a Dirac base distribution π_0, (104) simply defines a "flat" weight function $w_0 \equiv 1$ since $\|\pi_0\|_\kappa^2 = \langle \pi_0, \pi_0 \rangle_\kappa = \kappa(0,0) = 1 = |\langle \phi_\omega^1, \pi_0 \rangle|$.

Theorem 6. *Consider a normalized shift-invariant kernel κ. Consider a location-based family $\mathcal{T}$ with base distribution π_0, an integer $k \geq 1$, and the separated k-mixture model $\mathfrak{G}_k$ from (10) where $\varrho(\theta, \theta') := \|\theta - \theta'\|$ for some arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$. Consider a vector $\theta^* \in \mathbb{R}^d$ such that $\|\theta^*\| \geq 1$, and observe that the following two k-mixtures are 1-separated with respect to ϱ, i.e., $\nu_{k,1}, \nu_{k,2} \in \mathfrak{G}_k$, so that $\nu_k := (\nu_{k,1} - \nu_{k,2})/\|\nu_{k,1} - \nu_{k,2}\|_\kappa \in \mathcal{S}_k$:*

$$\nu_{k,1} = \sum_{i=1}^{k} \frac{1}{k} \pi_{(2i-2)\theta^*}, \quad \nu_{k,2} = \sum_{i=1}^{k} \frac{1}{k} \pi_{(2i-1)\theta^*}. \tag{106}$$

1. *If the $2k$-coherence of κ with respect to $\mathcal{T}$ is bounded by $0 \leq c < 1$ then for any κ-compatible weight function w that satisfies (102), the w-RFF sketching operator $\mathcal{A}$ with m i.i.d frequencies $\omega_j \sim \Lambda := w^2 \hat{\kappa}$ satisfies*

$$\mathbb{V}\|\mathcal{A}\nu_k\|^2 \geq \frac{1}{m}\Big(C_w k - 1\Big), \tag{107}$$

where $$C_w := \frac{\kappa_w(0,0)}{\|\pi_0\|_\kappa^4} \cdot \frac{4/3 - 2c_w}{(1+c)^2} \qquad \text{with } c_w := 2k \sup_{\theta,\theta' : \|\theta - \theta'\| \geq 1} \frac{|\kappa_w(\theta, \theta')|}{\kappa_w(0,0)}. \tag{108}$$

2. *If the mutual coherence of κ is bounded by $c/2k$ we have $c_{w_0} \leq c$ with w_0 as in (104).*

The proof of Theorem 6 is given in Appendix A.2. For a kernel with mutual coherence bounded by $c/2k$ with $c \leq 1/2$, we obtain $\frac{4/3 - 2c_{w_0}}{(1+c)^2} \geq (4/3 - 1)/(3/2)^2 = 4/27 \geq 1/7$. Since $\kappa_{w_0}(0,0)/\|\pi_0\|_\kappa^4 = 1$ (by Lemma 4) we get $C_{w_0} \geq 1/7$. Finally, since the mutual coherence of κ is bounded by $c/2k$, its $2k$-coherence is bounded by c (cf (19)), and (107) implies that the w_0-RFF sketching operator $\mathcal{A}$ with m i.i.d frequencies $\omega_j \sim \Lambda_0 := w_0^2 \hat{\kappa}$ satisfies

$$\mathbb{V}\|\mathcal{A}\nu_k\|^2 \geq \frac{k/7 - 1}{m}. \tag{109}$$

In the special case of a mixture of Diracs, since $w_0 \equiv 1$, it is not difficult to check that $\kappa_{w_0} = \kappa$. For a non-negative kernel $\kappa \geq 0$, the fact that $\kappa_{w_0} \geq 0$ allows to improve an intermediate bound (in Equation (118) in the proof), leading to the same result where $4/3 - 2c_{w_0}$ is replaced with $4/3 - c_{w_0}$ in the definition of C_{w_0}. This shows that (109) is then valid even with a mutual coherence bounded by $c/2k$ with $c < 1$.

Under the assumptions of Theorem 6, with $w = w_0$, we have $\mathbb{V}\|\mathcal{A}\nu_k\|^2 = \Omega(k/m)$. This implies that with this weight function, and even with the usual incoherence assumption, the term $v(k, \tau)$ in (40) *cannot* be bounded from above by a universal constant that is independent of k. This result highlights the difference between classical compressed sensing and sketching. Indeed, if we consider a random Gaussian matrix $A \in \mathbb{R}^{m \times d}$ with i.i.d. entries $\mathcal{N}(0, 1/m)$, it is well known that for every normalized vector $x \in \mathbb{R}^d$ (such that $\|x\|_2 = 1$) the variance of $\|Ax\|_2^2$ *does not depend on the sparsity* k of the vector x; see e.g. [FR13, Lemma 9.8].

Figure 2 illustrates this claim: we compare the variance of $\|Ax_k\|_2^2$ where $x_k \in \mathbb{R}^d$ is a normalized vector of sparsity k to the variance of $\|\mathcal{A}\nu_k\|_2^2$ where $\mathcal{A}$ and ν_k are defined in Theorem 6, with π_0 the Dirac distribution, $w = w_0 \equiv 1$, and the $2k$-coherence of κ is smaller than $1/2$. We observe that the variance of $\|Ax_k\|_2^2$ is practically flat as a function of k, while the variance of $\|\mathcal{A}\nu_k\|_2^2$ is linear in k. This observation shows that the study of the RIP in the set of mixtures of Diracs $\mathfrak{G}_k$ is not a mere extension of the existing RIP literature in Euclidean spaces.

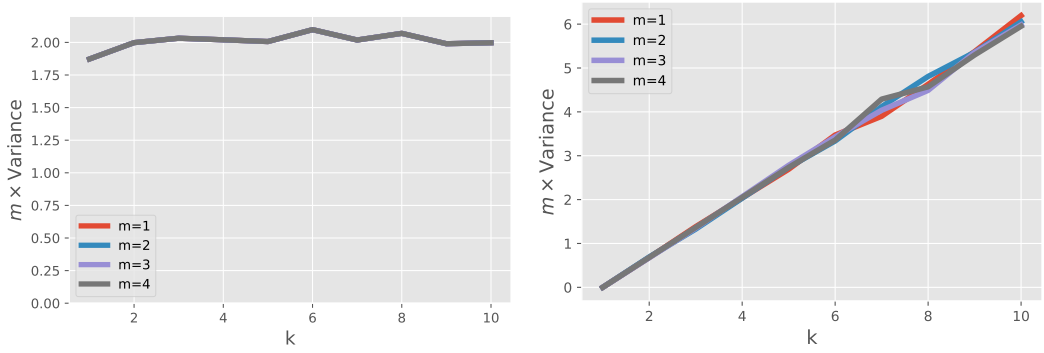

Figure 2: The term $m \times \mathbb{V}\|Ax_k\|^2$ (left) compared to the term $m \times \mathbb{V}\|\mathcal{A}\nu_k\|^2$ (right).

Limits of the proof technique proposed in this paper. Now, a careful examination of the analysis leading to Corollary 1 and Corollary 2 suggests that the unwanted $O(k^2 d)$ instead of $O(kd)$ behaviour of the sufficient sketch size results from the requirement of concentration inequality (76) in Theorem 4. In particular, in order to improve the theoretical guarantees of sketching using i.i.d. random frequencies, it would be tempting to seek a weight function w such that the random variable

$$\psi_\ell(z|\omega) := \psi(\omega) f_\ell(z|\omega)$$

is sub-exp$(\nu_\ell/, b_\ell)$ with $\nu_\ell = \mathcal{O}(1/\sqrt{k})$, for each $\ell \in \{\mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$ and $z \in \Theta_\ell$.

The reader can check that such an approach would indeed allow to establish guarantees with a sketch size $\mathcal{O}(kd)$. However, this would also imply that for $\ell \in \{\mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$ and $z \in \Theta_\ell$, the variance of $\psi_\ell(z|\omega)$ would satisfy $\mathbb{V}\psi_\ell(z|\omega) = \mathcal{O}(1/k)$. Now, when κ has mutual coherence bounded by $\mu < 1/(2k-1)$ (this is a natural assumption in our context), the expectation of this variable satisfies $\mathbb{E}\psi_\ell(z|\omega) = \mathcal{O}(1/k)$, hence we would obtain that

$$\mathbb{E}[\psi_\ell^2(z|\omega)] = \mathcal{O}(1/k). \tag{110}$$

The following result shows that (110) *cannot hold* in the specific setting of mixture of Diracs.

Proposition 7. *Consider $\mathcal{T}$ to be a location-based family with the dirac in 0 as a base distribution, and consider κ to be a normalized shift-invariant kernel such that $\kappa \geq 0$. Consider ϕ_ω as defined in (28), then for any κ-compatible weight function w and for $\omega \sim \Lambda := w^2\hat{\kappa}$, we have*

$$\forall y \in \Theta_{\mathrm{mm}}, \quad \mathbb{E}[\psi_{\mathrm{mm}}^2(y|\omega)] \geq \frac{1}{4}. \tag{111}$$

This lower bound holds *irrespective of how small the mutual coherence of κ may be.*

Proof. Let $y \in \Theta_{\mathrm{mm}}$. Since $\|\pi_0\|_\kappa^2 = \kappa(0,0) = 1$ and $|\langle \pi_0, \phi_\omega \rangle| = 1/w(\omega)$, by (52)-(54) we have

$$\psi_{\mathrm{mm}}(y|\omega) := \psi(\omega) f_{\mathrm{mm}}(y|\omega) = \frac{|\langle \pi_0, \phi_\omega \rangle|^2}{\|\pi_0\|_\kappa^2} \cos(\langle \omega, y \rangle) = \frac{\cos(\langle \omega, y \rangle)}{w^2(\omega)}. \tag{112}$$

As w is κ-compatible, we have $\int_{\mathbb{R}^d} w^2(\omega)\hat{\kappa}(\omega)\mathrm{d}\omega = 1$, thus by Cauchy-Schwarz inequality we get

$$\begin{aligned}
\mathbb{E}_{\omega\sim\Lambda}[\psi_{\mathrm{mm}}^2(y|\omega)] = \int_{\mathbb{R}^d} \frac{\cos^2(\langle \omega, y \rangle)}{w^4(\omega)} w^2(\omega)\hat{\kappa}(\omega)\mathrm{d}\omega &= \int_{\mathbb{R}^d} w^2(\omega)\hat{\kappa}(\omega)\mathrm{d}\omega \int_{\mathbb{R}^d} \frac{\cos^2(\langle \omega, y \rangle)}{w^2(\omega)} \hat{\kappa}(\omega)\mathrm{d}\omega \\
&\geq \left(\int_{\mathbb{R}^d} |\cos(\langle \omega, y \rangle)| \hat{\kappa}(\omega)\mathrm{d}\omega \right)^2 \\
&\geq \left(\int_{\mathbb{R}^d} \cos^2(\langle \omega, y \rangle) \hat{\kappa}(\omega)\mathrm{d}\omega \right)^2.
\end{aligned}$$

Finally, observe that $\cos(\langle \omega, y \rangle)^2 = (1 + \cos(2\langle \omega, y \rangle))/2$, so that we have, using that $\kappa \geq 0$,

$$\mathbb{E}_{\omega\sim\Lambda}[\psi_{\mathrm{mm}}^2(y|\omega)] \geq \left[\frac{1}{2} \left(\int_{\mathbb{R}^d} \hat{\kappa}(\omega)\mathrm{d}\omega + \int_{\mathbb{R}^d} \cos(2\langle \omega, y \rangle) \hat{\kappa}(\omega)\mathrm{d}\omega \right) \right]^2 = \left(\frac{1}{2}(1 + \kappa(2y, 0)) \right)^2 \geq \frac{1}{4}.$$

□

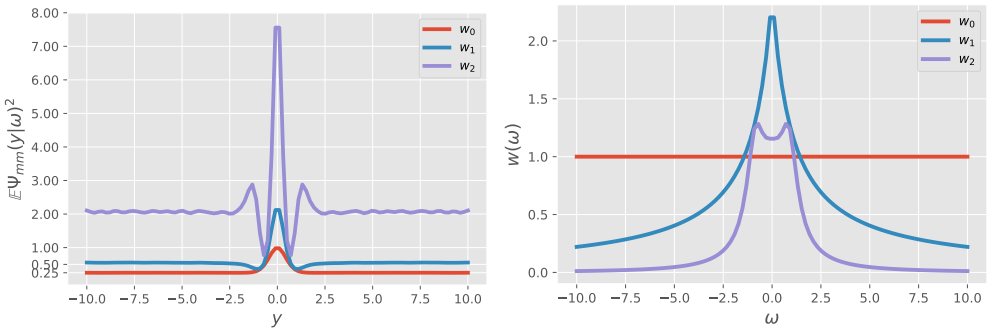

Figure 3: An illustration of the lower bound of Proposition 7 (left) for three choices of w (right): $w_1(\omega) = 1$, $w_2(\omega) = (1 + \|\omega\|)^{-1}$, $w_3(\omega) = (\|\omega\|^4 + 1)(\|\omega\|^6 + 1)^{-1}$.

Proposition 7 shows that improving the dependency of the sketch size on m cannot simply rely on improved concentration bounds: obtaining sharper bounds on sketch sizes that reflect the empirical findings of [KTTG17] seems to require a substantially subtler analysis which is beyond the scope of this paper.

4 Conclusion

In this work we revisited the theoretical analysis of the Restricted Isometry Property for sketching operators proposed in [GBKT21a, GBKT21b]. This property is

crucial in the field of compressive learning: it measures how the sketching operator preserves the MMD distance between measures belonging to a model set of measures. In particular, the sketching operators proposed in [GBKT21b] are suited for models of mixtures and are based on Fourier features. Nevertheless, the proposed theoretical analysis makes some additional assumptions that are summarized by the conditions (43). After investigating the partial necessity of the conditions (43) in the analysis of [GBKT21b], we proposed an alternative analysis based on deterministic bounds of $\delta(\mathcal{S}_k|\mathcal{A})$, then we showed how to leverage these deterministic bounds to establish the Restricted Isometry Property for stochastic sketching operators restricted to sets of mixtures based on location based measures. In particular, we showed that our revisited analysis allows to deal with realistic settings not covered by [GBKT21b].

Beyond these contributions, this work opens the door to further developments on the theoretical study of sketching operators used in the context of compressive learning. For instance, in the context of structured sketching introduced in [CGK18], the frequencies are rather block-i.i.d. samples but not i.i.d. samples (cf the end of Section 2.5). Theorem 4 remains valid in this context: indeed this result can be used without assuming that the frequencies $\omega_1, \ldots, \omega_m$ are i.i.d.. The main hindrance remaining on this direction is to check that the punctual concentration expressed by conditions (74), (75), (76) hold even when using block-i.i.d. frequencies. For this purpose, the existing results on the literature may help [LSS13, CS16, CRW17, MKBO18]. Another setting where our results may be useful is the study of deterministic sketching operators. Indeed, as shown in Section 3, the core of our analysis is based on deterministic bounds of $\delta(\mathcal{S}_k|\mathcal{A})$ presented in Section 3.2, and recent years have witnessed an increased interest into the theoretical study of *deterministic Fourier feature maps* [DDSR17, YSAM14]. Investigating whether deterministic sketching operators still satisfy the same guarantees as the stochastic ones is thus both a natural and challenging question.

As shown in Section 3.4, neither the analysis of [GBKT21b] nor our analysis unfortunately achieves to explain the empirical findings of [KTTG17], and there remains a gap between sufficient sketch sizes endowed with theoretical guarantees, which scale as $\mathcal{O}(k^2 d)$, and practically observed sketch sizes, which scale as $\mathcal{O}(kd)$. On the one hand, the quadratic theoretical dependency on the 'sparsity' k is not surprising given the known limits of sparse recovery guarantees exploiting dictionary coherence [FR13, Chapter 5]. Yet, the literature on compressive sensing manages to establish bounds essentially linear in the sparsity using random matrix techniques that do rely on mutual coherence [FR13, Chapter 9]. The proofs in this field exploit a fine study of the eigenvalues of random matrices, which was until now somehow overlooked in the community of compressive learning. Thus, an interesting direction of research is the study of the eigenvalues of the random matrices that appear in this context. Recent developments on the study of ridge kernel regression for random Fourier features may help [AKM+17, LTOS19]. In particular, in this line of research, the authors investigated the impact of the frequency distribution in the quality of the approximations based on random Fourier features. The techniques developed in these works may be helpful to understand the impact of the frequency distribution in the design of sketching operators. In the same vein, alternative frequency distributions, that define other kernels than the Gaussian kernel, have manifested better empirical performance when used in sketching-based learning tasks such as mixture learning; see Section 4.2 in [Cha20] for an example. This motivates to scrutinize the impact of the kernel on the design of the sketching operator.

Acknowledgement This project was supported by the AllegroAssai ANR project ANR-19-CHIA-0009. The authors would like to thank Titouan Vayer for his constructive feedback on an early version of this work and for sketching an early version of Lemma 7. Rémi Gribonval would like to thank Felix Krahmer for interesting discussions at the Oberwolfach Workshop 2148.

A Proofs

A.1 Proof of Lemma 4

By (102), the integral in (103) converges hence the kernel κ_w is well-defined and shift-invariant. Consider arbitrary $\theta, \theta' \in \mathbb{R}^d$ and denote $\pi_\theta, \pi_{\theta'}$ deduced from π_0 as in a location-based family. Recall that by definition, π_θ is the distribution of $X + \theta$ where $X \sim \pi_0$, and $\pi_{\theta'}$ is the distribution of $X + \theta'$. By (31), standard computations with the MMD[11] yield

$$\langle \pi_\theta, \pi_{\theta'} \rangle_\kappa = \int_{\mathbb{R}^d} \left| \langle \phi^1_\omega, \pi_0 \rangle \right|^2 \hat{\kappa}(\omega) e^{2\pi \imath \omega^\top (\theta - \theta')} \mathrm{d}\omega. \tag{113}$$

Specializing (113) to $\theta = \theta' = 0$ we get by Cauchy-Schwarz' inequality, since w is κ-compatible

$$\begin{aligned} \|\pi_0\|_\kappa^4 &= \left(\int_{\mathbb{R}^d} \left| \langle \phi^1_\omega, \pi_0 \rangle \right|^2 \hat{\kappa}(\omega) \mathrm{d}\omega \right)^2 = \left(\int_{\mathbb{R}^d} \left| \langle \phi^1_\omega, \pi_0 \rangle \right|^2 w^{-1}(\omega) \sqrt{\hat{\kappa}(\omega)} w(\omega) \sqrt{\hat{\kappa}(\omega)} \mathrm{d}\omega \right)^2 \\ &\leq \left(\int_{\mathbb{R}^d} \left| \langle \phi^1_\omega, \pi_0 \rangle \right|^4 w^{-2}(\omega) \hat{\kappa}(\omega) \mathrm{d}\omega \right)^2 \cdot \left(\int_{\mathbb{R}^d} w^2(\omega) \hat{\kappa}(\omega) \mathrm{d}\omega \right)^2 \overset{(103)\&(29)}{:=} \kappa_w(0,0)\,. \end{aligned}$$

The equality case of Cauchy-Schwarz is when $w(\omega) \propto \left| \langle \phi^1_\omega, \pi_0 \rangle \right|^2 w^{-1}(\omega)$, i.e., $w(\omega) \propto w_0(\omega)$ with w_0 defined in (104). We have

$$\int w_0^2(\omega) \hat{\kappa}(\omega) \mathrm{d}\omega \overset{(104)}{=} \|\pi_0\|_\kappa^{-2} \cdot \int |\langle \phi^1_\omega, \pi_0 \rangle|^2 \hat{\kappa}(\omega) \mathrm{d}\omega \overset{(113)}{=} \|\pi_0\|_\kappa^{-2} \cdot \langle \pi_0, \pi_0 \rangle_\kappa = 1\,,$$

hence w_0 is the only equality-case of Cauchy-Schwarz which is κ-compatible. The fact that $w_0(\omega)$ satisfies (102) follows from $w_0 \propto \left| \langle \phi^1_\omega, \pi_0 \rangle \right|^2 w_0^{-1}(\omega)$. Finally we write

$$\begin{aligned} \kappa_{w_0}(\theta, \theta') &\overset{(103)}{:=} \int_{\mathbb{R}^d} \left| \langle \phi^1_\omega, \pi_0 \rangle \right|^4 w_0^{-2}(\omega) \hat{\kappa}(\omega) e^{2\pi \imath \omega^\top (\theta - \theta')} \mathrm{d}\omega \\ &\overset{(104)}{=} \|\pi_0\|_\kappa^2 \cdot \int_{\mathbb{R}^d} \left| \langle \phi^1_\omega, \pi_0 \rangle \right|^2 \hat{\kappa}(\omega) e^{2\pi \imath \omega^\top (\theta - \theta')} \mathrm{d}\omega \overset{(113)}{=} \|\pi_0\|_\kappa^2 \cdot \langle \pi_\theta, \pi_{\theta'} \rangle_\kappa. \end{aligned}$$

A.2 Proof of Theorem 6

The proof relies on the following result which gives a closed formula of the variance of interest.

Proposition 8. *Consider a normalized shift-invariant kernel κ, a κ-compatible positive weight function w satisfying (102), $\Lambda = w^2 \hat{\kappa}$ and $\mathcal{A}$ a κ-compatible random w-FF sketching operator as in Example 3 with m i.i.d. frequencies. Consider a location-based family $\mathcal{T}$ with base distribution π_0. Given any location parameters $\theta_1, \dots, \theta_{2k} \in \Theta$ and weights $u_1, \dots, u_{2k} \in \mathbb{R}$, we have*

$$\mathbb{V} \| \mathcal{A} \sum_{i \in [2k]} u_i \pi_{\theta_i} \|^2 = \frac{1}{m} \Big((\boldsymbol{u}^{\otimes 2})^\top \boldsymbol{K}_w(\theta) \boldsymbol{u}^{\otimes 2} - \big\| \sum_{i=1}^{2k} u_i \pi_{\theta_i} \big\|_\kappa^4 \Big), \tag{114}$$

[11] See [MFS+17, Section 2.1] and [GBKT21b, proof of Proposition 6.2].

where $\boldsymbol{u}^{\otimes 2} \in \mathbb{R}^{(2k)^2}$ *is the tensor product with itself of the vector* $\boldsymbol{u} \in \mathbb{R}^{2k}$ *containing the* u_i*, and*

$$\boldsymbol{K}_w(\theta) = [\kappa_w(\theta_{i_1} - \theta_{i_2}, \theta_{i_4} - \theta_{i_3})]_{(i_1,i_2),(i_3,i_4)\in[2k]^2} \in \mathbb{R}^{(2k)^2 \times (2k)^2} \tag{115}$$

with $[2k] := \{1, 2, \ldots, 2k\}$ *and* κ_w *is the shift-invariant kernel defined in* (103)*.*

Proof of Proposition 8. As $\|\mathcal{A}\sum_{i=1}^{2k} u_i \pi_{\theta_i}\|^2 = \frac{1}{m}\sum_{j\in[m]} \big|\langle \phi^w_{\omega_j}, \sum_{i=1}^{2k} u_i \pi_{\theta_i}\rangle\big|^2$ is the average of m i.i.d. variables, we have $\mathbb{V}\|\mathcal{A}\sum_{i=1}^{2k} u_i \pi_{\theta_i}\|^2 = \frac{1}{m}\mathbb{V}\big|\langle \phi^w_{\omega}, \sum_{i=1}^{2k} u_i \pi_{\theta_i}\rangle\big|^2$ so we now characterize

$$\begin{aligned}
\mathbb{V}|\langle \phi^w_\omega, \sum_{i=1}^{2k} u_i \pi_{\theta_i}\rangle|^2 &= \mathbb{E}_{\omega\sim\Lambda}\big|\langle \phi^w_\omega, \sum_{i=1}^{2k} u_i \pi_{\theta_i}\rangle\big|^4 - \Big(\mathbb{E}_{\omega\sim\Lambda}\big|\langle \phi^w_\omega, \sum_{i=1}^{2k} u_i \pi_{\theta_i}\rangle\big|^2\Big)^2 \\
&= \mathbb{E}_{\omega\sim\Lambda}\big|\langle \phi^w_\omega, \sum_{i=1}^{2k} u_i \pi_{\theta_i}\rangle\big|^4 - \big\|\sum_{i=1}^{2k} u_i \pi_{\theta_i}\big\|_\kappa^4
\end{aligned}$$

where we used (32) since w is κ-compatible and $\Lambda = w^2\hat{\kappa}$ (see Example 3). Given the expression (28) of $\phi_\omega := \phi^w_\omega = \phi^1_\omega / w(\omega)$ and since $\{\pi_\theta\}_{\theta\in\Theta}$ is location-based,

$$\langle \phi_\omega, \pi_\theta\rangle = \mathbb{E}_{X\sim\pi_\theta}\phi_\omega(X) = \mathbb{E}_{X'\sim\pi_0}\phi_\omega(X'+\theta) = e^{2\pi\imath\omega^\top\theta}\mathbb{E}_{X'\sim\pi_0}\phi_\omega(X') = e^{2\pi\imath\omega^\top\theta}\langle \phi_\omega, \pi_0\rangle,$$

so that we can develop

$$\begin{aligned}
\big|\langle \phi_\omega, \sum_{i=1}^{2k} u_i \pi_{\theta_i}\rangle\big|^4 &= \sum_{i_1=1}^{2k}\sum_{i_2=1}^{2k}\sum_{i_3=1}^{2k}\sum_{i_4=1}^{2k} u_{i_1}u_{i_2}u_{i_3}u_{i_4}\langle \phi_\omega, \pi_{\theta_{i_1}}\rangle\overline{\langle \phi_\omega, \pi_{\theta_{i_2}}\rangle}\langle \phi_\omega, \pi_{\theta_{i_3}}\rangle\overline{\langle \phi_\omega, \pi_{\theta_{i_4}}\rangle} \\
&= \sum_{i_1=1}^{2k}\sum_{i_2=1}^{2k}\sum_{i_3=1}^{2k}\sum_{i_4=1}^{2k} u_{i_1}u_{i_2}u_{i_3}u_{i_4}|\langle \phi_\omega, \pi_0\rangle|^4 e^{2\pi\imath\left(\omega^\top(\theta_{i_1}-\theta_{i_2}+\theta_{i_3}-\theta_{i_4})\right)}.
\end{aligned}$$

Moreover, given the expression (30) of the pdf $\Lambda(\omega)$ we have for every $i_1, i_2, i_3, i_4 \in [2k]$

$$\begin{aligned}
\mathbb{E}_{\omega\sim\Lambda}|\langle \phi_\omega, \pi_0\rangle|^4 e^{2\pi\imath\left(\omega(\theta_{i_1}-\theta_{i_2}+\theta_{i_3}-\theta_{i_4})\right)} &= \int_{\mathbb{R}^d} |\langle \phi_\omega, \pi_0\rangle|^4 e^{2\pi\imath\left(\omega^\top(\theta_{i_1}-\theta_{i_2}+\theta_{i_3}-\theta_{i_4})\right)} w^2(\omega)\hat{\kappa}(\omega)d\omega \\
&= \int_{\mathbb{R}^d} |\langle \phi^1_\omega, \pi_0\rangle|^4 w^{-4}(\omega) e^{2\pi\imath\left(\omega^\top(\theta_{i_1}-\theta_{i_2}+\theta_{i_3}-\theta_{i_4})\right)} w^2(\omega)\hat{\kappa}(\omega)d\omega \\
&= \kappa_w(\theta_{i_1} - \theta_{i_2}, \theta_{i_4} - \theta_{i_3}),
\end{aligned}$$

where κ_w is defined in (103). As a result, we have

$$\mathbb{E}\big|\langle \phi_\omega, \sum_{i=1}^{2k} u_i \pi_{\theta_i}\rangle\big|^4 = \sum_{i_1=1}^{2k}\sum_{i_2=1}^{2k}\sum_{i_3=1}^{2k}\sum_{i_4=1}^{2k} u_{i_1}u_{i_2}u_{i_3}u_{i_4}\kappa_w(\theta_{i_1}-\theta_{i_2}, \theta_{i_4}-\theta_{i_3}) = (\boldsymbol{u}^{\otimes 2})^\top \boldsymbol{K}_w(\theta)\boldsymbol{u}^{\otimes 2}$$

where, according to the notations of the proposition $\boldsymbol{u}^{\otimes 2} \in \mathbb{R}^{(2k)^2}$ is the vector with entries $u^{\otimes 2}_{(i_1,i_2)} = u_{i_1}u_{i_2}$ for each pair $(i_1, i_2) \in [2k] \times [2k]$, and $\boldsymbol{K}_w(\theta)$ is the square matrix of size $(2k)^2 \times (2k)^2$ defined in (115). Putting the pieces together yields (114) as claimed. □

By Proposition 8 with $\nu_k := \mu_{k,1} - \mu_{k,2} = \sum_{i=1}^{2k} u_i \pi_{\theta_i}$, where $u_i := \frac{(-1)^{i-1}}{k}$, $\theta_i := (i-1)\theta^*$,

$$\mathbb{V}\|\mathcal{A}\nu_k\|^2 = \frac{\mathbb{V}\|\mathcal{A}(\mu_{k,1}-\mu_{k,2})\|^2}{\|\mu_{k,1}-\mu_{k,2}\|_\kappa^4} = \frac{1}{m}\left(\frac{(\boldsymbol{u}^{\otimes 2})^\top \boldsymbol{K}_w(\theta)\boldsymbol{u}^{\otimes 2}}{\|\mu_{k,1}-\mu_{k,2}\|_\kappa^4} - 1\right). \tag{116}$$

We now bound $\|\mu_{k,1}-\mu_{k,2}\|_\kappa^4$ and $(\boldsymbol{u}^{\otimes 2})^\top \boldsymbol{K}_w(\theta)\boldsymbol{u}^{\otimes 2}$ to get the first part of the Theorem.

Bounding $\|\mu_{k,1}-\mu_{k,2}\|_\kappa^4$. Since κ is shift-invariant and $\mathcal{T}$ is a location-based family we have $\|\pi_{\theta_i}\|_\kappa^2 = \|\pi_0\|_\kappa^2$ for each $i \in [2k]$. Since $\varrho(\theta_i,\theta_j) = \|\theta_i - \theta_j\| = |i-j|\cdot\|\theta^*\| \geq 1$ for $1 \leq i \neq j \leq 2k$, the $2k$ (unnormalized) monopoles $\{u_i\pi_{\theta_i}\}_{i=1}^{2k}$ are pairwise 1-separated dipoles with respect to ρ. As κ has its $2k$-coherence with respect to $\mathcal{T}$ bounded by c (cf Definition 3-(16)), it follows that

$$\frac{2}{k}(1-c)\|\pi_0\|_\kappa^2 = (1-c)\sum_{i=1}^{2k}\|u_i\pi_{\theta_i}\|_\kappa^2 \leq \|\mu_{1,k}-\mu_{2,k}\|_\kappa^2 \leq (1+c)\sum_{i=1}^{2k}\|u_i\pi_{\theta_i}\|_\kappa^2 = \frac{2}{k}(1+c)\|\pi_0\|_\kappa^2,$$

where we used that $u_i^2 = 1/k^2$ for every i. Therefore

$$\frac{4\|\pi_0\|_\kappa^4(1-c)^2}{k^2} \leq \|\mu_{1,k}-\mu_{2,k}\|_\kappa^4 \leq \frac{4\|\pi_0\|_\kappa^4(1+c)^2}{k^2}. \tag{117}$$

Bounding $(\boldsymbol{u}^{\otimes 2})^\top \boldsymbol{K}_w(\theta)\boldsymbol{u}^{\otimes 2}$. Since $u_{i_1}u_{i_2} = (-1)^{i_1+i_2}/k^2$ for each $i_1, i_2 \in [2k]$, we write

$$(\boldsymbol{u}^{\otimes 2})^\top \boldsymbol{K}_w(\theta)\boldsymbol{u}^{\otimes 2} = \frac{1}{k^4}\sum_{i_1,i_2,i_3,i_4\in[2k]} (-1)^{i_1+i_2+i_3+i_4}\Big[\boldsymbol{K}_w(\theta)\Big]_{i_1,i_2,i_3,i_4}.$$

Consider the sets

$$\begin{aligned}
\mathcal{I}_= &:= \{(i_1,i_2,i_3,i_4)\in[2k]^4,\ i_1-i_2 = i_3-i_4\},\\
\mathcal{I}_+ &:= \{(i_1,i_2,i_3,i_4)\in[2k]^4,\ (-1)^{i_1+i_2+i_3+i_4} = 1\},\\
\mathcal{I}_- &:= \{(i_1,i_2,i_3,i_4)\in[2k]^4,\ (-1)^{i_1+i_2+i_3+i_4} = -1\},
\end{aligned}$$

and observe that $\mathcal{I}_= \subset \mathcal{I}_+$, $\mathcal{I}_+ \cup \mathcal{I}_- = [2k]^4$ and that the definition (115) implies

$$(-1)^{i_1+i_2+i_3+i_4}\Big[\boldsymbol{K}_w(\theta)\Big]_{i_1,i_2,i_3,i_4} \geq \begin{cases} \kappa_w(0,0), & \forall (i_1,i_2,i_3,i_4)\in\mathcal{I}_= \\ -|\kappa_w(\theta_{i_1}-\theta_{i_2},\theta_{i_4}-\theta_{i_3})| & \forall (i_1,i_2,i_3,i_4)\in\mathcal{I}_- \\ -|\kappa_w(\theta_{i_1}-\theta_{i_2},\theta_{i_4}-\theta_{i_3})| & \forall (i_1,i_2,i_3,i_4)\in\mathcal{I}_+\backslash\mathcal{I}_=. \end{cases} \tag{118}$$

Further observe that since $i_1-i_2-i_3+i_4 \equiv i_1+i_2+i_3+i_4[2]$, if $(i_1,i_2,i_3,i_4) \in \mathcal{I}_- \cup (\mathcal{I}_+\backslash\mathcal{I}_=)$, then either $i_1-i_2-i_3+i_4 \equiv 1[2]$ or $i_1-i_2-i_3+i_4 \neq 0$. In both cases, we have $|i_1-i_2-i_3+i_4| \geq 1$, hence $\|\theta_{i_1}-\theta_{i_2}-\theta_{i_3}+\theta_{i_4}\| = \|(i_1-i_2-i_3+i_4)\theta^*\| \geq 1$, and

$$|\kappa_w(\theta_{i_1}-\theta_{i_2},\theta_{i_3}-\theta_{i_4})| \leq \sup_{\theta,\theta':\|\theta-\theta'\|\geq 1}|\kappa_w(\theta,\theta')| \leq \frac{c_w}{2k}\kappa_w(0,0),$$

where we used definition (108). Observe moreover that since $\mathcal{I}_- = \{(i_1,i_2,i_3,i_4); (i_1,i_2,i_3)\in[2k]^3, i_4\in[2k], i_4 \equiv 1-(i_1+i_2+i_3)[2]\}$ we have $\sharp\mathcal{I}_- = (2k)^3\times k = 8k^4$. Similarly,

$\sharp(\mathcal{I}_+\backslash\mathcal{I}_=) = \sharp\mathcal{I}_+ - \sharp\mathcal{I}_= = 8k^4 - \sharp\mathcal{I}_=$. Finally, as we will show below $\sharp\mathcal{I}_= \geq 16k^3/3$ (the proof is postponed at the end of the section) we obtain

$$\begin{aligned}\sum_{(i_1,i_2,i_3,i_4)\in[2k]} (-1)^{i_1+i_2+i_3+i_4}\Big[\boldsymbol{K}_w(\theta)\Big]_{i_1,i_2,i_3,i_4} &\geq \Big(\sharp\mathcal{I}_= - \sharp(\mathcal{I}_+\backslash\mathcal{I}_=)\cdot\frac{c_w}{2k} - \sharp\mathcal{I}_-\cdot\frac{c_w}{2k}\Big)\cdot\kappa_w(0,0)\\ &= \Big(\sharp\mathcal{I}_=\Big(1+\frac{c_w}{2k}\Big) - 8c_wk^3\Big)\cdot\kappa_w(0,0)\\ &\geq (16/3 - 8c_w)k^3\cdot\kappa_w(0,0).\end{aligned}$$

Thus, $(\boldsymbol{u}^{\otimes 2})^\top \boldsymbol{K}_w(\theta)\boldsymbol{u}^{\otimes 2} \geq \frac{8}{k}\Big(\frac{2}{3} - c_w\Big)\cdot\kappa_w(0,0)$ and combining with (116) and (117) yields $\mathbb{V}\|\mathcal{A}\nu_k\|^2 \geq \frac{1}{m}(C_wk - 1)$ with

$$C_wk = \frac{8}{k}(2/3 - c_w)\kappa_w(0,0)\frac{k^2}{4\|\pi_0\|_\kappa^4(1+c)^2},$$

i.e., $C_w = 2(2/3 - c_w)\frac{\kappa_w(0,0)}{\|\pi_0\|_\kappa^4(1+c)^2}$ as defined in (108).

To conclude the proof of (107) we now establish the claimed lower bound on $\sharp\mathcal{I}_=$. Since $\mathcal{I}_=$ is the disjoint union of $\mathcal{I}_=^\ell := \{(i_1,i_2,i_3,i_4)\in[2k]^4, i_1 - i_2 = i_3 - i_4 = \ell\}$, $\ell\in\{-(2k-1),\dots,0,\dots,2k-1\}$, and $\sharp\mathcal{I}_=^\ell = \sharp\mathcal{I}_=^{-\ell} = (2k-\ell)^2$ for $0\leq\ell\leq 2k-1$, hence

$$\begin{aligned}\sharp\mathcal{I}_= &= \sharp\mathcal{I}_=^0 + 2\sum_{\ell=1}^{2k-1}\sharp\mathcal{I}_=^\ell = (2k)^2 + \sum_{\ell=1}^{2k-1}2(2k-\ell)^2 = \Big(2\sum_{\ell=0}^{2k-1}(2k-\ell)^2\Big) - (2k)^2\\ &= \Big(2\sum_{\ell'=1}^{2k}(\ell')^2\Big) - 4k^2 = \frac{1}{3}2k(2k+1)(4k+1) - \frac{12}{3}k^2 \geq \frac{16}{3}k^3. \qquad (119)\end{aligned}$$

where we used the well known fact that $\sum_{\ell'=1}^n(\ell')^2 = n(n+1)(2n+1)/6$ for every integer n.

We proceed to the second claim. For θ,θ' such that $\|\theta-\theta'\|\geq 1$, $\iota = \pi_\theta$ and $\iota' = \pi_{\theta'}$ are (non-normalized) 1-separated dipoles with respect to ϱ. The mutual coherence assumption and Lemma 4 yield

$$\frac{\kappa_{w_0}(\theta,\theta')}{\|\pi_0\|_\kappa^2} = \langle\pi_\theta,\pi_{\theta'}\rangle_\kappa \leq \frac{c}{2k}\|\pi_\theta\|_\kappa\|\pi_{\theta'}\|_\kappa = \frac{c}{2k}\|\pi_0\|_\kappa^2 = \frac{c}{2k}\langle\pi_0,\pi_0\rangle_\kappa = \frac{c}{2k}\frac{\kappa_{w_0}(0,0)}{\|\pi_0\|_\kappa^2}. \qquad (120)$$

This shows that c_{w_0} defined in (108) satisfies $c_{w_0}\leq c$.

A.3 Proof of Proposition 2

For any mixture model, kernel, and feature family, the set of normalized dipoles $\mathfrak{D}$ is included in the normalized secant set hence

$$\sup_{\nu\in\mathcal{S}_k}\|\nu\|_{\mathcal{F}} = \sup_{\omega\in\mathbb{R}^d}\sup_{\nu\in\mathcal{S}_k}|\langle\phi_\omega^w,\nu\rangle| \geq \sup_{\omega\in\mathbb{R}^d}\sup_{\nu\in\mathfrak{D}}|\langle\phi_\omega^w,\nu\rangle|.$$

For $x\in\Theta$ s.t. $0<\|x\|\leq 1$, define $\tilde{\iota}_x := \pi_x - \pi_0$. Since Θ contains a neighborhood of zero, $\pi_0/\|\pi_0\|_\kappa, \tilde{\iota}_x/\|\tilde{\iota}_x\|_\kappa\in\mathfrak{D}$ as soon as $\|x\|$ is small enough hence there is $0<\delta\leq 1$ such that

$$\sup_{\nu\in\mathcal{S}_k}\|\nu\|_{\mathcal{F}} \geq \sup_{\omega\in\mathbb{R}^d}\max\left(\frac{|\langle\phi_\omega^w,\pi_0\rangle|}{\|\pi_0\|_\kappa}, \sup_{0<\|x\|\leq\delta}\frac{|\langle\phi_\omega^w,\pi_x-\pi_0\rangle|}{\|\pi_x-\pi_0\|_\kappa}\right).$$

Since $\langle \phi_\omega^w, \pi_x \rangle = e^{2\pi\imath\langle\omega,x\rangle}\langle \phi_\omega^w, \pi_0\rangle$ and $\|\pi_x - \pi_0\|_\kappa^2 = 2\|\pi_0\|_\kappa^2(1-\bar\kappa(x))$ we have

$$\sup_{0<\|x\|\le\delta} \frac{|\langle \phi_\omega^w, \pi_x - \pi_0\rangle|}{\|\pi_x-\pi_0\|_\kappa} = |\langle \phi_\omega^w, \pi_0\rangle| \cdot \sup_{0<\|x\|\le\delta} \frac{|e^{2\pi\imath\langle\omega,x\rangle}-1|}{\|\pi_0\|_\kappa\sqrt{2(1-\bar\kappa(x))}}.$$

Now, since $\bar\kappa$ achieved its maximum at zero where it is C^2, we have for every $x \neq 0$

$$\lim_{t\to 0}\sqrt{\frac{1-\bar\kappa(tx)}{t^2\|x\|^2}} = \sqrt{\frac{-x^{\mathrm{T}}\nabla^2\bar\kappa(0)x}{2\|x\|^2}}.$$

By assumption, $\nabla^2\bar\kappa(0) \neq 0$ and we get

$$C := \sup_{0<\|x\|\le\delta}\lim_{t\to 0}\sqrt{\frac{1-\bar\kappa(tx)}{t^2\|x\|^2}} = \sqrt{\frac{\|\nabla^2\bar\kappa(0)\|_{\mathrm{op}}}{2}} > 0.$$

We obtain

$$\begin{aligned}\sup_{0<\|x\|\le\delta}\frac{|e^{2\pi\imath\langle\omega,x\rangle}-1|}{\sqrt{1-\bar\kappa(x)}} &\ge \sup_{0<\|x\|\le\delta}\lim_{t\to 0}\frac{|e^{2\pi\imath t\langle\omega,x\rangle}-1|}{\sqrt{1-\bar\kappa(tx)}} = \sup_{0<\|x\|\le\delta}\lim_{t\to 0}\frac{|2\pi\imath t\langle\omega,x\rangle|}{\sqrt{1-\bar\kappa(tx)}}\\ &= \sup_{0<\|x\|\le\delta}\frac{|\langle 2\pi\omega, x/\|x\|\rangle|}{\lim_{t\to 0}\sqrt{\frac{1-\bar\kappa(tx)}{t^2\|x\|^2}}} \ge \frac{1}{C}\sup_{0<\|x\|\le\delta}|\langle 2\pi\omega, x/\|x\|\rangle| = \frac{2\pi}{C}\|\omega\|_\star.\end{aligned}$$

Therefore

$$\sup_{\nu\in\mathcal{S}_k}\|\nu\|_{\mathcal{F}} \ge \sup_{\omega\in\mathbb{R}^d}\frac{|\langle\phi_\omega^w,\pi_0\rangle|}{\|\pi_0\|_\kappa}\max\left(1, \frac{\pi\sqrt{2}}{C}\|\omega\|_\star\right).$$

To conclude we use that $\pi\sqrt{2}/C = 2\pi/\sqrt{\|\nabla^2\bar\kappa(0)\|_{\mathrm{op}}}$.

A.4 Proof of Proposition 3

Let $\nu \in \mathcal{S}_k$. By Proposition 1, there exist $2k$ normalized dipoles $\iota_1, \dots, \iota_{2k} \in \mathfrak{D}$, with $(\iota_i, \iota_j) \in \mathfrak{D}^2_{\neq}$ when $i \neq j$, and coefficients $\alpha_1, \dots, \alpha_{2k} \ge 0$ such that $(1+c)^{-1} \le \sum_{i\in[2k]}\alpha_i^2 \le (1-c)^{-1}$ that satisfy $\nu = \sum_{i\in[2k]}\alpha_i\iota_i$. We notice that

$$\sum_{i\neq j\in[2k]}\alpha_i\alpha_j = (\sum_{i\in[2k]}\alpha_i)^2 - \sum_{i\in[2k]}\alpha_i^2 \le (\sqrt{2k}\|\alpha\|_2)^2 - \|\alpha\|_2^2 \le \frac{2k-1}{1-c}.$$

Since $\mathcal{A}\pi_\theta$ is well defined for probability distributions in the family $\mathcal{T}$, the action of $\mathcal{A}$ is well defined on k-mixtures, hence on elements of the normalized secant set. We have

$$\begin{aligned}\|\mathcal{A}\nu\|_2^2 - \sum_{i\in[2k]}\alpha_i^2 &= \sum_{i,j\in[2k]}\alpha_i\alpha_j\langle\mathcal{A}\iota_i, \mathcal{A}\iota_j\rangle - \sum_{i\in[2k]}\alpha_i^2\\ &= \sum_{i\in[2k]}\alpha_i^2(\|\mathcal{A}\iota_i\|_2^2 - 1) + \sum_{i\neq j\in[2k]}\alpha_i\alpha_j\langle\mathcal{A}\iota_i,\mathcal{A}\iota_j\rangle,\end{aligned}$$

we obtain

$$\|\mathcal{A}\nu\|_2^2 - 1 = \sum_{i\in[2k]}\alpha_i^2 - 1 + \sum_{i\in[2k]}\alpha_i^2(\|\mathcal{A}\iota_i\|_2^2 - 1) + \sum_{i\neq j\in[2k]}\alpha_i\alpha_j\langle\mathcal{A}\iota_i,\mathcal{A}\iota_j\rangle.$$

For $i \neq j$, since $\langle \mathcal{A}\iota_j, \mathcal{A}\iota_i \rangle$ is the complex conjugate of $\langle \mathcal{A}\iota_i, \mathcal{A}\iota_j \rangle$, we have $\langle \mathcal{A}\iota_i, \mathcal{A}\iota_j \rangle + \langle \mathcal{A}\iota_j, \mathcal{A}\iota_i \rangle = \mathfrak{Re}\langle \mathcal{A}\iota_i, \mathcal{A}\iota_j \rangle + \mathfrak{Re}\langle \mathcal{A}\iota_j, \mathcal{A}\iota_i \rangle$ hence $\sum_{i \neq j \in [2k]} \alpha_i \alpha_j \langle \mathcal{A}\iota_i, \mathcal{A}\iota_j \rangle = \sum_{i \neq j \in [2k]} \alpha_i \alpha_j \mathfrak{Re}\langle \mathcal{A}\iota_i, \mathcal{A}\iota_j \rangle$. As a result

$$\begin{aligned}\big|\|\mathcal{A}\nu\|_2^2 - 1\big| &\leq \Big|1 - \sum_{i \in [2k]} \alpha_i^2\Big| + \Big|\sum_{i \in [2k]} \alpha_i^2 (\|\mathcal{A}\iota_i\|_2^2 - 1)\Big| + \Big|\sum_{i \neq j \in [2k]} \alpha_i \alpha_j \mathfrak{Re}\langle \mathcal{A}\iota_i, \mathcal{A}\iota_j \rangle\Big| \\ &\leq \Big|1 - \sum_{i \in [2k]} \alpha_i^2\Big| + \sum_{i \in [2k]} \alpha_i^2 \sup_{\iota \in \mathfrak{D}} \big|\|\mathcal{A}\iota\|_2^2 - 1\big| + \sum_{i \neq j \in [2k]} \alpha_i \alpha_j \sup_{(\iota, \iota') \in \mathfrak{D}^2_{\neq}} \big|\mathfrak{Re}\langle \mathcal{A}\iota, \mathcal{A}\iota' \rangle\big|.\end{aligned}$$

Now, since $(1+c)^{-1} \leq \sum_{i \in [2k]} \alpha_i^2 \leq (1-c)^{-1}$ we have $\big|1 - \sum_{i \in [2k]} \alpha_i^2\big| \leq c/(1-c)$, hence

$$\big|\|\mathcal{A}\nu\|_2^2 - 1\big| \leq \frac{1}{1-c}\left(c + \sup_{\iota \in \mathfrak{D}} \big|\|\mathcal{A}\iota\|_2^2 - 1\big| + (2k-1) \sup_{(\iota, \iota') \in \mathfrak{D}^2_{\neq}} \big|\mathfrak{Re}\langle \mathcal{A}\iota, \mathcal{A}\iota' \rangle\big|\right).$$

Since this holds for every $\nu \in \mathcal{S}_k$, this establishes (44) using the definitions of $\delta(\cdot|\mathcal{A})$ (see (12)) and of $\mu(\mathfrak{D}^2_{\neq}|\mathcal{A})$ (see (18)). □

A.5 Proof of Proposition 4

To prove Proposition 4 we rely on a generic formula expressing $\|\mathcal{A}\iota\|_2^2$ for $\iota \in \mathfrak{D}$ that depends on a scalar $\alpha \in [0,1]$, which reflects how *balanced* the normalized dipole ι is, and on a vector x, which reflects the *relative position* between the supports of the two monopoles that form ι. We also exploit an expression of $\mathfrak{Re}\langle \mathcal{A}\iota, \mathcal{A}\iota' \rangle$ for $(\iota, \iota') \in \mathfrak{D}^2_{\neq}$ that depends on two scalars α, α' reflecting how balanced ι and ι' are respectively, and on two relative-position vectors x, x'. The proof of the following proposition is deferred to Appendix A.5.1.

Proposition 9. *Consider $\mathcal{T} = (\Theta, \rho, \mathcal{I})$ a location-based family with base distribution π_0 where $\rho(\cdot,\cdot) = \|\cdot - \cdot\|$ for some norm $\|\cdot\|$, and κ a normalized shift-invariant kernel that is locally characteristic with respect to $\mathcal{T}$. Consider $\mathcal{A}$ a WFF sketching operator (Definition 6) with frequencies $\omega_1, \dots, \omega_m$, Θ_{d} and $\psi(\omega)$ as defined in (59) and (52), and $\bar{\kappa}$ as defined in (15).*

- *For any normalized dipole $\iota \in \mathfrak{D}$, there exists $\alpha \in [0,1]$ and a vector $x \in \Theta_{\mathrm{d}}$ such that*

$$\|\mathcal{A}\iota\|_2^2 = \frac{1}{m} \sum_{j=1}^{m} \psi(\omega_j) \frac{(1-\alpha)^2 + 2\alpha(1 - \cos\langle \omega_j, x \rangle)}{(1-\alpha)^2 + 2\alpha(1 - \bar{\kappa}(x))}. \tag{121}$$

  The case of a normalized monopole $\iota \in \mathfrak{M}$ corresponds to $\alpha = 0$ (and arbitrary x), while the case of a balanced normalized dipole corresponds to $\alpha = 1$. Vice-versa, for any $\alpha \in [0,1]$ and $x \in \Theta_{\mathrm{d}}$ there is $\iota \in \mathfrak{D}$ such that this equality holds.

- *For $(\iota, \iota') \in \mathfrak{D}^2_{\neq}$, there exist $s, s' \in \{-1, 1\}$, $\alpha, \alpha' \in [0,1]$, $\theta_1, \theta_2, \theta_1', \theta_2' \in \Theta$ that satisfy*

$$\begin{aligned} \forall i, j \in \{1,2\}, \quad & \varrho(\theta_i, \theta_j') \geq 1, \\ 0 < \varrho(\theta_1, \theta_1') \leq 1, \quad & 0 < \varrho(\theta_2, \theta_2') \leq 1, \end{aligned} \tag{122}$$

such that

$$\mathfrak{Re}\langle \mathcal{A}\iota, \mathcal{A}\iota' \rangle = ss' \frac{a - \alpha' b - \alpha c + \alpha\alpha' d}{\sqrt{(1-\alpha)^2 + 2\alpha(1-\bar{\kappa}(x))}\sqrt{(1-\alpha')^2 + 2\alpha'(1-\bar{\kappa}(x'))}}, \tag{123}$$

where $x := \theta_1 - \theta_2 \in \Theta_{\mathrm{d}}$ *and* $x' := \theta_1' - \theta_2' \in \Theta_{\mathrm{d}}$*, and* a, b, c *and* d *are defined as follows*

$$\begin{aligned} a &= \frac{1}{\|\pi_0\|_\kappa^2}\mathfrak{Re}\langle \mathcal{A}\pi_{\theta_1}, \mathcal{A}\pi_{\theta_1'} \rangle, \quad b = \frac{1}{\|\pi_0\|_\kappa^2}\mathfrak{Re}\langle \mathcal{A}\pi_{\theta_1}, \mathcal{A}\pi_{\theta_2'} \rangle, \\ c &= \frac{1}{\|\pi_0\|_\kappa^2}\mathfrak{Re}\langle \mathcal{A}\pi_{\theta_2}, \mathcal{A}\pi_{\theta_1'} \rangle, \quad d = \frac{1}{\|\pi_0\|_\kappa^2}\mathfrak{Re}\langle \mathcal{A}\pi_{\theta_2}, \mathcal{A}\pi_{\theta_2'} \rangle. \end{aligned} \tag{124}$$

The case where ι *is a monopole (resp. a balanced dipole) corresponds to* $\alpha = 0$ *(resp.* $\alpha = 1$*) and similarly for* ι' *and* α'*.*

Proof of (49). Since normalized monopoles and balanced normalized dipoles are special cases of normalized dipoles, we have $\mathfrak{M} \subset \mathfrak{D}$ and $\hat{\mathfrak{D}} \subset \mathfrak{D}$ hence by the definition (12) of $\delta(\mathfrak{D}|\mathcal{A})$ as a supremum we trivially have

$$\delta(\mathfrak{D}|\mathcal{A}) \geq \max\big(\delta(\mathfrak{M}|\mathcal{A}), \delta(\hat{\mathfrak{D}}|\mathcal{A})\big).$$

To establish (49) we show the converse inequality. First observe that for any normalized monopole $\iota_{\mathfrak{M}} \in \mathfrak{M}$ we have $\|\mathcal{A}\iota_{\mathfrak{M}}\|_2^2 = \frac{1}{m}\sum_{j=1}^m \psi(\omega_j)$. Since ϱ is a norm, this is a direct consequence of Proposition 9 and shows that $\delta(\mathfrak{M}|\mathcal{A}) = |1 - \|\mathcal{A}\iota_{\mathfrak{M}}\|_2^2|$ independently of the choice of the monopole $\iota_{\mathfrak{M}}$.

Now, consider an arbitrary normalized dipole $\iota \in \mathfrak{D}$. If ι is either a normalized monopole or a balanced dipole, we trivially have $|1 - \|\mathcal{A}\iota\|_2^2| \leq \max(\delta(\mathfrak{M}|\mathcal{A}), \delta(\hat{\mathfrak{D}}|\mathcal{A}))$. Otherwise, by Proposition 9 again, there exists $\alpha \in (0,1)$ such that

$$\|\mathcal{A}\iota\|_2^2 = \frac{1}{m}\sum_{j=1}^m \psi(\omega_j)\frac{(1-\alpha)^2 + 2\alpha(1-\cos\langle\omega_j, x\rangle)}{(1-\alpha)^2 + 2\alpha(1-\bar{\kappa}(x))}.$$

Define a *balanced* normalized dipole as $\iota_{\hat{\mathfrak{D}}} := \frac{\pi_0 - \pi_x}{\|\pi_0 - \pi_x\|_\kappa} \in \hat{\mathfrak{D}}$. Proposition 9 again yields

$$\|\mathcal{A}\iota_{\hat{\mathfrak{D}}}\|_2^2 = \frac{1}{m}\sum_{j=1}^m \psi(\omega_j)\frac{1-\cos\langle\omega_j, x\rangle}{1-\bar{\kappa}(x)}$$

so that, with simple algebraic manipulations, we have

$$\begin{aligned} \|\mathcal{A}\iota\|_2^2 &= \frac{(1-\alpha)^2\frac{1}{m}\sum_{j=1}^m \psi(\omega_j) + 2\alpha\frac{1}{m}\sum_{j=1}^m \psi(\omega_j)(1-\cos\langle\omega_j, x\rangle)}{(1-\alpha)^2 + 2\alpha(1-\bar{\kappa}(x))} \\ &= \frac{(1-\alpha)^2\|\mathcal{A}\iota_{\mathfrak{M}}\|_2^2 + 2\alpha(1-\bar{\kappa}(x))\|\mathcal{A}\iota_{\hat{\mathfrak{D}}}\|_2^2}{(1-\alpha)^2 + 2\alpha(1-\bar{\kappa}(x))}. \end{aligned}$$

We deduce that

$$\min\big(\|\mathcal{A}\iota_{\mathfrak{M}}\|_2^2, \|\mathcal{A}\iota_{\hat{\mathfrak{D}}}\|_2^2\big) \leq \|\mathcal{A}\iota\|_2^2 \leq \max\big(\|\mathcal{A}\iota_{\mathfrak{M}}\|_2^2, \|\mathcal{A}\iota_{\hat{\mathfrak{D}}}\|_2^2\big).$$

Moreover, since $\iota_{\hat{\mathfrak{D}}} \in \hat{\mathfrak{D}}$, we have $\max(1-\|\mathcal{A}\iota_{\hat{\mathfrak{D}}}\|_2^2, \|\mathcal{A}\iota_{\hat{\mathfrak{D}}}\|_2^2 - 1) = |1-\|\mathcal{A}\iota_{\hat{\mathfrak{D}}}\|_2^2| \leq \delta(\hat{\mathfrak{D}}|\mathcal{A})$ hence

$$\begin{aligned}
1-\|\mathcal{A}\iota\|_2^2 &\leq 1-\min\left(\|\mathcal{A}\iota_{\mathfrak{M}}\|_2^2, \|\mathcal{A}\iota_{\hat{\mathfrak{D}}}\|_2^2\right) \\
&= \max\left(1-\|\mathcal{A}\iota_{\mathfrak{M}}\|_2^2, 1-\|\mathcal{A}\iota_{\hat{\mathfrak{D}}}\|_2^2\right) \leq \max\left(\delta(\mathfrak{M}|\mathcal{A}), \delta(\hat{\mathfrak{D}}|\mathcal{A})\right), \\
\|\mathcal{A}\iota\|_2^2 - 1 &\leq \max\left(\|\mathcal{A}\iota_{\mathfrak{M}}\|_2^2 - 1, \|\mathcal{A}\iota_{\hat{\mathfrak{D}}}\|_2^2 - 1\right) \leq \max\left(\delta(\mathfrak{M}|\mathcal{A}), \delta(\hat{\mathfrak{D}}|\mathcal{A})\right),
\end{aligned}$$

This shows that $|1-\|\mathcal{A}\iota\|_2^2| \leq \max\left(\delta(\mathfrak{M}|\mathcal{A}), \delta(\hat{\mathfrak{D}}|\mathcal{A})\right)$ and establishes (49) as claimed. □

Proof of (50). Recall that the families defined in (48) satisfy $\mathfrak{M}^2_{\neq}, \mathfrak{M}\times\hat{\mathfrak{D}}_{\neq}, \hat{\mathfrak{D}}^2_{\neq} \subset \mathfrak{D}^2_{\neq}$. Hence, by the definition (18) of $\mu(\cdot|\mathcal{A})$ as a supremum we trivially have

$$\mu(\mathfrak{D}^2_{\neq}|\mathcal{A}) \geq \max\left(\mu(\mathfrak{M}^2_{\neq}|\mathcal{A}), \mu(\mathfrak{M}\times\hat{\mathfrak{D}}_{\neq}|\mathcal{A}), \mu(\hat{\mathfrak{D}}^2_{\neq}|\mathcal{A})\right).$$

This yields the lower bound in (50). To establish the upper bound we will use a result which proof is postponed to Section A.5.2.

Proposition 10. *Let $a,b,c,d \in \mathbb{R}$ and $e,f \in [0,1)$ and consider the function g defined on $[0,1]\times[0,1]$ by*

$$h(u,v) = \frac{a-bu-cv+duv}{\sqrt{1+u^2-2eu}\sqrt{1+v^2-2fv}}. \tag{125}$$

We have

$$\sup_{(u,v)\in[0,1]^2} \left|h(u,v)\right| \leq 3\max\left(|a|,|b|,|c|,|d|, \frac{|b-a|}{\sqrt{1-e}}, \frac{|d-c|}{\sqrt{1-e}}, \frac{|d-b|}{\sqrt{1-f}}, \frac{|c-a|}{\sqrt{1-f}}, \frac{|a-b-c+d|}{\sqrt{1-e}\sqrt{1-f}}\right).$$

Consider an arbitrary 1-separated pair of normalized dipoles, $(\iota,\iota') \in \mathfrak{D}^2_{\neq}$, and denote $s,s' \in \{-1,1\}$, $\alpha,\alpha' \in [0,1]$, $\theta_1,\theta_2,\theta_1',\theta_2' \in \Theta$, x, a,b,c,d the parameters satisfying (122)-(123)-(124) as given by Proposition 9. As $x,x' \in \Theta_{\mathrm{d}}$ we have $0 < \varrho(x,0) \leq 1$ and $0 < \varrho(x',0) \leq 1$, and since κ is *locally characteristic* [12] with respect to $\mathcal{T}$ and as $\kappa \geq 0$ implies $\bar{\kappa} \geq 0$ (cf (9) and (15)) , this yields $e := \bar{\kappa}(x'), f := \bar{\kappa}(x) \in [0,1)$ so that the expression (123) reads as $h(\alpha',\alpha)$ with h as in (125). Since $\alpha,\alpha' \in [0,1]$ and $e,f \in [0,1)$, by Proposition 10 applied to the absolute value of the expression (123) we get

$$\left|\mathfrak{Re}\langle\mathcal{A}\iota,\mathcal{A}\iota'\rangle\right| \leq 3\max\left(|a|,|b|,|c|,|d|, \frac{|b-a|}{\sqrt{1-e}}, \frac{|c-d|}{\sqrt{1-e}}, \frac{|b-d|}{\sqrt{1-f}}, \frac{|a-c|}{\sqrt{1-f}}, \frac{|b-a-c+d|}{\sqrt{1-e}\sqrt{1-f}}\right).$$

Now, observe that (124) in Proposition 9 implies that every $t \in \{a,b,c,d\}$ can be written as $t = \mathfrak{Re}\langle\mathcal{A}\nu,\mathcal{A}\nu'\rangle$ where $(\nu,\nu') \in \mathfrak{M}^2_{\neq}$, hence

$$\max(|a|,|b|,|c|,|d|) \leq \sup_{(\xi,\xi')\in\mathfrak{M}^2_{\neq}} \left|\mathfrak{Re}\langle\mathcal{A}\xi,\mathcal{A}\xi'\rangle\right| = \mu(\mathfrak{M}^2_{\neq}|\mathcal{A}).$$

Similarly, observe that (124) implies also that $|b-a|/\sqrt{1-e} = |\mathfrak{Re}\langle\mathcal{A}\nu,\mathcal{A}\nu'\rangle|$ where

$$\nu := \frac{\pi_{\theta_1}}{\|\pi_{\theta_1}\|_\kappa} \in \mathfrak{M}; \quad \nu' := \frac{\pi_{\theta_1'} - \pi_{\theta_2'}}{\|\pi_{\theta_1'} - \pi_{\theta_2'}\|_\kappa} \in \hat{\mathfrak{D}},$$

[12] See Definition 2.

and by (122) we have $(\nu, \nu') \in \mathfrak{M} \times \hat{\mathfrak{D}}_{\neq}$ hence

$$\frac{|b-a|}{\sqrt{1-e}} \leq \sup_{(xi,\xi')\in\mathfrak{M}\times\hat{\mathfrak{D}}_{\neq}} \left|\mathfrak{Re}\langle \mathcal{A}\xi, \mathcal{A}\xi'\rangle\right| =: \mu(\mathfrak{M}\times\hat{\mathfrak{D}}_{\neq}|\mathcal{A}).$$

By symmetry, the same argument is valid for $|c-d|/\sqrt{1-e}, |b-d|/\sqrt{1-f}$ and $|a-c|/\sqrt{1-f}$. Therefore

$$\max\Big(\frac{|b-a|}{\sqrt{1-e}}, \frac{|c-d|}{\sqrt{1-e}}, \frac{|b-d|}{\sqrt{1-f}}, \frac{|a-c|}{\sqrt{1-f}}\Big) \leq \mu(\mathfrak{M}\times\hat{\mathfrak{D}}_{\neq}|\mathcal{A}).$$

Finally, with $\nu := (\pi_{\theta_1} - \pi_{\theta_2})/\|\pi_{\theta_1} - \pi_{\theta_2}\|_\kappa$ and $\nu' := (\pi_{\theta_1'} - \pi_{\theta_2'})/\|\pi_{\theta_1'} - \pi_{\theta_2'}\|_\kappa$, we have $(\nu, \nu') \in \hat{\mathfrak{D}}^2_{\neq}$ (by (122)) and as a result

$$\frac{|b-a-c+d|}{\sqrt{1-e}\sqrt{1-f}} = \left|\mathfrak{Re}\langle \mathcal{A}\nu, \mathcal{A}\nu'\rangle\right| \leq \sup_{(\xi,\xi')\in\hat{\mathfrak{D}}^2_{\neq}} \left|\mathfrak{Re}\langle \mathcal{A}\xi, \mathcal{A}\xi'\rangle\right| =: \mu(\hat{\mathfrak{D}}^2_{\neq}|\mathcal{A}).$$

Combining all of the above yields

$$\left|\mathfrak{Re}\langle \mathcal{A}\iota, \mathcal{A}\iota'\rangle\right| \leq 3\max\big(\mu(\mathfrak{M}^2_{\neq}|\mathcal{A}), \mu(\mathfrak{M}\times\hat{\mathfrak{D}}_{\neq}|\mathcal{A}), \mu(\hat{\mathfrak{D}}^2_{\neq}|\mathcal{A})\big).$$

As this holds for every $(\iota, \iota') \in \mathfrak{D}^2_{\neq}$ this establishes (50). □

A.5.1 Proof of Proposition 9

Consider a normalized dipole $\iota \in \mathfrak{D}$. Since ρ is a norm we can apply [GBKT21b, Lemma C.1] hence there exists a dipole $\tilde{\iota}$ such that $\iota = \frac{\tilde{\iota}}{\|\tilde{\iota}\|_\kappa}$, with $\tilde{\iota} = \frac{s}{\|\pi_0\|_\kappa}(\pi_{\theta_1} - \alpha\pi_{\theta_2})$, where $s \in \{-1, 1\}$, $\alpha \in [0, 1]$ and $x := \theta_1 - \theta_2 \in \Theta - \Theta$ satisfies $0 < \|x\| \leq 1$. Since κ is locally characteristic we have $\|\tilde{\iota}\|_\kappa > 0$ hence the ratio $\tilde{\iota}/\|\tilde{\iota}\|_\kappa$ indeed makes sense. The case of a normalized monopole $\iota \in \mathfrak{M}$ corresponds to $\alpha = 0$ and an arbitrary x, while the case of a balanced dipole corresponds to $\alpha = 1$. Moreover, since κ is translation invariant and $\mathcal{T}$ is a location-based family by (21) we have $\|\pi_{\theta_1}\|_\kappa = \|\pi_{\theta_2}\|_\kappa = \|\pi_0\|_\kappa$, and $\langle \pi_{\theta_1}, \pi_{\theta_2}\rangle_\kappa = \bar{\kappa}(\theta_1, \theta_2) \cdot \|\pi_{\theta_1}\|_\kappa \|\pi_{\theta_2}\|_\kappa = \bar{\kappa}(x) \cdot \|\pi_0\|^2_\kappa$ where we recall that the $\mathcal{T}$-normalized kernel $\bar{\kappa}$ is defined in (15). Therefore

$$\|\tilde{\iota}\|^2_\kappa = (1-\alpha)^2 + 2\alpha(1 - \bar{\kappa}(x)),$$
$$\|\mathcal{A}\tilde{\iota}\|^2_2 = \frac{1}{m}\sum_{j=1}^m \psi(\omega_j)\Big((1-\alpha)^2 + 2\alpha(1 - \cos\langle \omega_j, x\rangle)\Big).$$

Since $\|\mathcal{A}\iota\|^2_2 = \|\mathcal{A}\tilde{\iota}\|^2_2/\|\tilde{\iota}\|^2_\kappa$, taking the quotient yields (121) as claimed. Vice-versa for $\alpha \in [0, 1]$ and $x \in \Theta_{\mathrm{d}}$, there are $\theta_1, \theta_2 \in \Theta^2$ such that $0 < \|\theta_1 - \theta_2\| \leq 1$ and setting $\iota = (\pi_{\theta_1} - \alpha\pi_{\theta_2})/\|\pi_{\theta_1} - \alpha\pi_{\theta_2}\|_\kappa$ yields a normalized dipole satisfying the desired expression.

Similarly, for any 1-separated pair of normalized dipoles $(\iota, \iota') \in \mathfrak{D}^2_{\neq}$ we write $\iota = \frac{\tilde{\iota}}{\|\tilde{\iota}\|_\kappa}$, $\iota' = \frac{\tilde{\iota}'}{\|\tilde{\iota}'\|_\kappa}$, with

$$\tilde{\iota} = \frac{s}{\|\pi_0\|_\kappa}(\pi_{\theta_1} - \alpha\pi_{\theta_2}), \quad \tilde{\iota}' = \frac{s'}{\|\pi_0\|_\kappa}(\pi_{\theta_1'} - \alpha'\pi_{\theta_2'}),$$

where $s, s' \in \{-1, 1\}$, $\alpha, \alpha' \in [0, 1]$ and $x := \theta_1 - \theta_2 \in \Theta - \Theta$ and $x' := \theta_1' - \theta_2' \in \Theta - \Theta$ satisfy $0 < \|x\| \leq 1$ and $0 < \|x'\| \leq 1$. The 1-separation assumption means that for

every $i,j \in \{1,2\}$ we have$\|\theta_i - \theta'_j\| \geq 1$. Since $\|\pi_{\theta_1}\|_\kappa = \|\pi_{\theta_2}\|_\kappa = \|\pi_{\theta'_1}\|_\kappa = \|\pi_{\theta'_2}\|_\kappa = \|\pi_0\|_\kappa$ and $\langle \pi_{\theta_1}, \pi_{\theta_2}\rangle_\kappa = \|\pi_0\|_\kappa^2 \bar\kappa(x)$, and $\langle \pi_{\theta'_1}, \pi_{\theta'_2}\rangle_\kappa = \|\pi_0\|_\kappa^2 \bar\kappa(x')$ we obtain

$$\begin{aligned}
\|\tilde\iota\|_\kappa^2 &= (1-\alpha)^2 + 2\alpha(1-\bar\kappa(x)),\\
\|\tilde\iota'\|_\kappa^2 &= (1-\alpha')^2 + 2\alpha'(1-\bar\kappa(x'))\\
\mathfrak{Re}\big(\langle \mathcal{A}\tilde\iota, \mathcal{A}\tilde\iota'\rangle\big) &= \frac{ss'}{\|\pi_0\|_\kappa^2}\mathfrak{Re}\Big(\big\langle \mathcal{A}(\pi_{\theta_1} - \alpha\pi_{\theta_2}), \mathcal{A}(\pi_{\theta'_1} - \alpha\pi_{\theta'_2})\big\rangle\Big) = ss'\left(a - \alpha' b - \alpha c + \alpha\alpha' d\right),
\end{aligned}$$

with a,b,c,d as in (124). Since $\mathfrak{Re}\big(\langle \mathcal{A}\iota, \mathcal{A}\iota'\rangle\big) = \mathfrak{Re}\big(\langle \mathcal{A}\tilde\iota, \mathcal{A}\tilde\iota'\rangle\big)/(\|\tilde\iota\|_\kappa\|\tilde\iota'\|_\kappa)$ we obtain (123). The special cases of monopoles and balanced dipoles, are proved similarly as above and left to the reader. □

A.5.2 Proof of Proposition 10

We will use a technical Lemma which proof is postponed to the end of the section.

Lemma 5. *Consider $\alpha,\beta \in \mathbb{R}$, $\gamma \in [0,1)$ and the function $\phi_{\alpha,\beta,\gamma}$ defined on $[0,1]$ by*

$$\phi_{\alpha,\beta,\gamma}(t) = \frac{\alpha - \beta t}{\sqrt{1 + t^2 - 2\gamma t}}. \tag{126}$$

We have

$$\forall t \in [0,1],\ \ |\phi_{\alpha,\beta,\gamma}(t)| \leq \sqrt{3}\max\Big(|\alpha|, |\beta|, \frac{|\beta-\alpha|}{\sqrt{2(1-\gamma)}}\Big). \tag{127}$$

Consider $(u,v) \in [0,1]^2$, $f \in [0,1)$, and define $g := \frac{1}{\sqrt{1+v^2-2fv}}$. We have

$$h(u,v) = \frac{ga - gbu - gcv + gduv}{\sqrt{1+u^2-2eu}} = \frac{(ga-gcv) - (gb-gdv)u}{\sqrt{1+u^2-2eu}} = \phi_{ga-gcv,gb-gdv,e}(u) \tag{128}$$

hence, by Lemma 5

$$|h(u,v)| = |\phi_{ga-gcv,gb-gdv,e}(u)| \leq \sqrt{3}\max\Big(|ga-gcv|, |gb-gdv|, \frac{|gb-gdv-ga+gcv|}{\sqrt{1-e}}\Big).$$

Now, we can use Lemma 5 again to get

$$\begin{aligned}
|ga-gcv| &= \frac{|a-cv|}{\sqrt{1+v^2-2fv}} = |\phi_{a,c,f}(v)| \leq \sqrt{3}\max\Big(|a|, |c|, \frac{|c-a|}{\sqrt{1-f}}\Big),\\
|gb-gdv| &= \frac{|b-dv|}{\sqrt{1+v^2-2fv}} = |\phi_{b,d,f}(v)| \leq \sqrt{3}\max\Big(|b|, |d|, \frac{|d-b|}{\sqrt{1-f}}\Big),\\
|gb-gdv-ga+gcv| &= |gb-ga+(gc-gd)v| = \frac{|b-a-(d-c)v|}{\sqrt{1+v^2-2fv}} = |\phi_{b-a,d-c,f}(v)|\\
&\leq \sqrt{3}\max\Big(|b-a|, |d-c|, \frac{|(d-c)-(b-a)|}{\sqrt{1-f}}\Big).
\end{aligned}$$

Combining the above inequalities we obtain

$$|h(u,v)| \leq 3\max\Big(|a|,|b|,|c|,|d|, \frac{|b-a|}{\sqrt{1-e}}, \frac{|d-c|}{\sqrt{1-e}}, \frac{|d-b|}{\sqrt{1-f}}, \frac{|c-a|}{\sqrt{1-f}}, \frac{|a-b-c+d|}{\sqrt{1-e}\sqrt{1-f}}\Big).$$ □

Proof of Lemma 5. Equivalently, we bound $c := \sup_{t\in[0,1]} g(t)$ where $g(t) := |\phi_{\alpha,\beta,\gamma}(t)|^2 = P(t)/Q(t)$, $P(t) := (\beta t-\alpha)^2$, and $Q(t) := 1+t^2-2\gamma t$. The bound $c \le 3\max(\alpha^2, \beta^2, (\beta-\alpha)^2/(1-\gamma))$ is trivial if $\alpha = \beta = 0$, so we now assume $(\alpha,\beta) \neq 0$. We have $g(0) = \alpha^2$ and $g(1) = (\beta-\alpha)^2/(2(1-\gamma))$ so the bound is also trivial if the maximum is achieved at a boundary point, so to conclude we now assume that c is achieved at an interior point $t \in (0,1)$, which must satisfy $g'(t) = 0$. Since $g' = (P'Q - PQ')/Q^2$ the fact that $g'(t) = 0$ reads as

$$\begin{aligned}
0 &= P'(t)Q(t) - P(t)Q'(t) \\
&= 2\beta(\beta t-\alpha)(1+t^2-2\gamma t) - (\beta t-\alpha)^2 2(t-\gamma) = 2(\beta t-\alpha)\big(\beta(1+t^2-2\gamma t) - (\beta t-\alpha)(t-\gamma)\big) \\
&= 2(\beta t-\alpha)\big(\beta + \cancel{\beta t^2} - 2\beta\gamma t - \cancel{\beta t^2} + \alpha t + \beta\gamma t - \alpha\gamma\big) = 2(\beta t-\alpha)\big((\alpha-\beta\gamma)t - (\alpha\gamma-\beta)\big).
\end{aligned}$$

Since we assume $(\alpha,\beta) \neq (0,0)$, we have $P(t)/Q(t) = g(t) \ge \max(g(0), g(1)) = \max(\alpha^2, (\beta-\alpha)^2/(2(1-\gamma))) > 0$, hence $P(t) \neq 0$, i.e. $\beta t - \alpha \neq 0$, thus the location of the maximum satisfies

$$(\alpha-\beta\gamma)t = \alpha\gamma - \beta.$$

This implies that $\alpha \neq \beta\gamma$ (otherwise we would have both $\alpha = \beta\gamma$ and $\alpha\gamma = \beta$, hence $\beta = \alpha\gamma = (\beta\gamma)\gamma = \beta\gamma^2$, and similarly $\alpha = \alpha\gamma^2$; since $0 \le \gamma < 1$ this would contradict the fact that $(\alpha,\beta) \neq (0,0)$). Moreover, since $P'(t)Q(t) - P(t)Q'(t) = 0$ we have $g(t) = P(t)/Q(t) = P'(t)/Q'(t) = 2\beta(\beta t-\alpha)/(2(t-\gamma))$. Since $g(t) > 0$ this shows that $\beta \neq 0$, and we conclude that

$$\begin{aligned}
g(t) &= \frac{\beta(\beta t-\alpha)}{t-\gamma} = \beta\frac{(\alpha-\beta\gamma)(\beta t-\alpha)}{(\alpha-\beta\gamma)(t-\gamma)} = \beta\frac{(\alpha\gamma-\beta)\beta - (\alpha-\beta\gamma)\alpha}{(\alpha\gamma-\beta) - (\alpha-\beta\gamma)\gamma} \\
&= \beta\frac{2\alpha\beta\gamma - \alpha^2 - \beta^2}{\beta(\gamma^2-1)} = \frac{\alpha^2+\beta^2-2\alpha\beta\gamma}{1-\gamma^2} = \frac{(\alpha-\beta)^2 + 2\alpha\beta(1-\gamma)}{(1+\gamma)(1-\gamma)} \\
&\le \frac{(\alpha-\beta)^2}{1-\gamma} + 2|\alpha\beta| \le \frac{(\alpha-\beta)^2}{1-\gamma} + 2\max\left(\alpha^2,\beta^2\right) \le 3\max\left(\alpha^2, \beta^2, \frac{(\alpha-\beta)^2}{1-\gamma}\right). \square
\end{aligned}$$

A.6 Proof of Proposition 5

Equalities (58), (59), (60), (61) and (62) are straightforward applications of the following lemma.

Lemma 6. *Under the assumptions and notations of Proposition 5, there exist sets* $\Theta_{\mathrm{md}} \subset \Theta_{\mathrm{d}} \times \Theta_{\mathrm{mm}}$ *and* $\Theta_{\mathrm{dd}} \subset \Theta_{\mathrm{d}} \times \Theta_{\mathrm{d}} \times \Theta_{\mathrm{mm}}$ *such that,*

1. *For every* $\iota \in \mathfrak{M}$*, we have* $1 - \|\mathcal{A}\iota\|_2^2 = 1 - \Psi_{\mathrm{m}}(\boldsymbol{\Omega})$*, for every* $\boldsymbol{\Omega}$ *and* $\mathcal{A} = \mathcal{A}_{\boldsymbol{\Omega}}$*.*
2. *For every* $\iota \in \hat{\mathfrak{D}}$*, there is* $x \in \Theta_{\mathrm{d}}$ *such that* $1 - \|\mathcal{A}\iota\|_2^2 = 1 - \Psi_{\mathrm{d}}(x|\boldsymbol{\Omega})$ *for every* $\boldsymbol{\Omega}$ *and* $\mathcal{A} = \mathcal{A}_{\boldsymbol{\Omega}}$*; and vice-versa.*
3. *For* $(\iota,\iota') \in \mathfrak{M}^2_{\neq}$*, there are* $s \in \{-1,1\}$*,* $y \in \Theta_{\mathrm{mm}}$ *s.t.* $\mathfrak{Re}\langle \mathcal{A}\iota, \mathcal{A}\iota'\rangle = s\Psi_{\mathrm{mm}}(y|\boldsymbol{\Omega})$ *for every* $\boldsymbol{\Omega}$ *and* $\mathcal{A} = \mathcal{A}_{\boldsymbol{\Omega}}$*, and vice-versa.*
4. *For* $(\iota,\iota') \in \mathfrak{M} \times \hat{\mathfrak{D}}_{\neq}$ *there are* $s \in \{-1,1\}$*,* $(x,y) \in \Theta_{\mathrm{md}}$ *s.t.* $\mathfrak{Re}\langle \mathcal{A}\iota, \mathcal{A}\iota'\rangle = s\Psi_{\mathrm{md}}(x,y|\boldsymbol{\Omega})$ *for every* $\boldsymbol{\Omega}$ *and* $\mathcal{A} = \mathcal{A}_{\boldsymbol{\Omega}}$*, and vice-versa.*
5. *For* $(\iota,\iota') \in \hat{\mathfrak{D}}^2_{\neq}$ *there are* $s \in \{-1,1\}$*,* $(x,x',y) \in \Theta_{\mathrm{dd}}$ *s.t.* $\mathfrak{Re}\langle \mathcal{A}\iota, \mathcal{A}\iota'\rangle = s\Psi_{\mathrm{dd}}(x,x',y|\boldsymbol{\Omega})$*, for every* $\boldsymbol{\Omega}$ *and* $\mathcal{A} = \mathcal{A}_{\boldsymbol{\Omega}}$*, and vice-versa.*

Proof of item 1: Proposition 9 yields $\|\mathcal{A}\iota\|_2^2 = \frac{1}{m}\sum_{j=1}^m \psi(\omega_j)$ for every $\iota \in \mathfrak{M}$. $\square$

Proof of item 2: Proposition 9 yields that for every $\iota \in \hat{\mathfrak{D}}$ there is $x \in \Theta_{\rm d}$ such that $1 - \|\mathcal{A}\iota\|_2^2 = 1 - \frac{1}{m}\sum_{j=1}^m \psi(\omega_j)\frac{1-\cos\langle\omega_j,x\rangle}{1-\bar\kappa(x)} = 1 - \frac{1}{m}\sum_{j=1}^m \psi(\omega_j) f_d(x|\omega_j)$, and vice-versa, where we used that for $t \in \mathbb{R}$ we have $1 - \cos t = 2\sin^2(t/2)$. □

The remaining items use the second part of Proposition 9 which gives a generic formula for $\mathfrak{Re}\langle\mathcal{A}\iota, \mathcal{A}\iota'\rangle$: when $(\iota,\iota') \in \mathfrak{D}^2_{\neq}$ there are $s, s' \in \{-1, 1\}$, $\alpha, \alpha' \in [0, 1]$, $\theta_1, \theta_2, \theta'_1, \theta'_2 \in \Theta$ satisfying (122)-(123) with $x := \theta_1 - \theta_2 \in \Theta_{\rm d}$, $x' := \theta'_1 - \theta'_2 \in \Theta_{\rm d}$, and a, b, c and d are defined by (124). We will also use that given that $\mathcal{A}$ is a (κ, w)-FF sketching operator (cf Definition 6) and π_θ is obtained by translation of π_0 we have $\mathfrak{Re}\langle\mathcal{A}\pi_{\theta_i}, \mathcal{A}\pi_{\theta'_j}\rangle = \frac{1}{m}\sum_{j=1}^m |\langle\pi_0, \phi_{\omega_j}\rangle|^2 \cos\langle\omega_j, \theta_i - \theta'_j\rangle$, and that $y := \theta_1 - \theta'_1$ satisfies $y \in \Theta - \Theta$ and $\|y\| = \|\theta_1 - \theta'_1\| \geq 1$ by (122), thus $y \in \Theta_{\rm mm}$.

Proof of item 3 When $(\iota,\iota') \in \mathfrak{M}^2_{\neq}$, we have $\alpha = \alpha' = 0$ hence (123)-(124) yield

$$\mathfrak{Re}\langle\mathcal{A}\iota, \mathcal{A}\iota'\rangle = ss'a = \frac{ss'}{\|\pi_0\|_\kappa^2}\frac{1}{m}\sum_{j=1}^m |\langle\pi_0, \phi_{\omega_j}\rangle|^2 \cos\big(\langle\omega_j, \theta_1 - \theta'_1\rangle\big) = \frac{ss'}{m}\sum_{j=1}^m \psi(\omega_j)\cos\big(\langle\omega_j, y\rangle\big).$$

Vice-versa for such s, s', y it is easy to exhibit $(\iota,\iota') \in \mathfrak{M}^2_{\neq}$ satisfying the same expression. □

Proof of item 4 When $(\iota,\iota') \in \mathfrak{M} \times \hat{\mathfrak{D}}_{\neq}$, $\alpha = 0$ and $\alpha' = 1$, which similarly yields

$$\mathfrak{Re}\langle\mathcal{A}\iota, \mathcal{A}\iota'\rangle = ss'\frac{a-b}{\sqrt{2(1-\bar\kappa(x'))}} = \frac{ss'}{m}\sum_{j=1}^m \psi(\omega_j)\frac{\cos\big(\langle\omega_j, \theta_1 - \theta'_1\rangle\big) - \cos\big(\langle\omega_j, \theta_1 - \theta'_2\rangle\big)}{\sqrt{2(1-\bar\kappa(x'))}}.$$

Now, using the identity $\cos(u) - \cos(v) = -2\sin((u-v)/2)\sin((u+v)/2)$, we get

$$\begin{aligned}\cos\big(\langle\omega_j, \theta_1 - \theta'_1\rangle\big) - \cos\big(\langle\omega_j, \theta_1 - \theta'_2\rangle\big) &= \cos\big(\langle\omega_j, y\rangle\big) - \cos\big(\langle\omega_j, y + x'\rangle\big)\\ &= 2\sin\big(\langle\omega_j, x'\rangle/2\big)\sin\big(\langle\omega_j, y + x'/2\rangle\big).\end{aligned}$$

Thus

$$\mathfrak{Re}\langle\mathcal{A}\iota, \mathcal{A}\iota'\rangle = ss'\Psi_{\rm md}((x', y)|\mathbf{\Omega}), \tag{129}$$

where $x' = \theta'_1 - \theta'_2$ and $y = \theta_1 - \theta'_1$ satisfy $x' \in \Theta_{\rm d}$ and $y \in \Theta_{\rm mm}$. We define $\Theta_{\rm md}$ as the set of all couples $(x', y) \in \Theta_{\rm d} \times \Theta_{\rm mm}$ that satisfy (129) for some $(\iota,\iota') \in (\mathfrak{M} \times \hat{\mathfrak{D}})_{\neq}$ and $s, s' \in \{-1, 1\}$.

Proof of item 5 When $(\iota,\iota') \in \hat{\mathfrak{D}}^2_{\neq}$, $\alpha = 1$ and $\alpha' = 1$, hence (123)-(124) similarly yield

$$\begin{aligned}\mathfrak{Re}\langle\mathcal{A}\iota, \mathcal{A}\iota'\rangle &= ss'\frac{a-b-c+d}{\sqrt{2(1-\bar\kappa(x))}\sqrt{2(1-\bar\kappa(x'))}}\\ &= \frac{ss'}{m}\sum_{j=1}^m \psi(\omega_j)\frac{\cos\big(\langle\omega_j, \theta_1 - \theta'_1\rangle\big) - \cos\big(\langle\omega_j, \theta_1 - \theta'_2\rangle\big) - \cos\big(\langle\omega_j, \theta_2 - \theta'_1\rangle\big) + \cos\big(\langle\omega_j, \theta_2 - \theta'_2\rangle\big)}{\sqrt{2(1-\bar\kappa(x))}\sqrt{2(1-\bar\kappa(x'))}}.\end{aligned}$$

Since $\cos(u) - \cos(v) = -2\sin(\frac{u-v}{2})\sin(\frac{u+v}{2})$ and $\sin(u) - \sin(v) = 2\sin(\frac{u-v}{2})\cos(\frac{u+v}{2})$, and denoting $y := \theta_1 - \theta'_1$, $x := \theta_1 - \theta_2$, and $x' := \theta'_1 - \theta'_2$, we get $x, x' \in \Theta_{\rm d}$, $y \in \Theta_{\rm mm}$

and

$$\begin{aligned}
&\cos\big(\langle\omega_j,\theta_1-\theta_1'\rangle\big)-\cos\big(\langle\omega_j,\theta_1-\theta_2'\rangle\big)-\cos\big(\langle\omega_j,\theta_2-\theta_1'\rangle\big)+\cos\big(\langle\omega_j,\theta_2-\theta_2'\rangle\big)\\
&\qquad=\cos\big(\langle\omega_j,y\rangle\big)-\cos\big(\langle\omega_j,y+x'\rangle\big)-\cos\big(\langle\omega_j,y-x\rangle\big)+\cos\big(\langle\omega_j,y-x+x'\rangle\big)\\
&\qquad=2\sin\big(\langle\omega_j,x'/2\rangle\big)\sin\big(\langle\omega_j,y+x'/2\rangle\big)-2\sin\big(\langle\omega_j,x'/2\rangle\big)\sin\big(\langle\omega_j,y-x+x'/2\rangle\big)\\
&\qquad=2\sin\big(\langle\omega_j,x'/2\rangle\big)\Big(\sin\big(\langle\omega_j,y+x'/2\rangle\big)-\sin\big(\langle\omega_j,y-x+x'/2\rangle\big)\Big)\\
&\qquad=4\sin\big(\langle\omega_j,x/2\big)\sin\big(\langle\omega_j,x'/2\rangle\big)\cos\big(\langle\omega_j,y+x'/2-x/2\rangle\big).
\end{aligned}$$

Thus

$$\mathfrak{Re}\langle\mathcal{A}\iota,\mathcal{A}\iota'\rangle=ss'\Psi_{\rm dd}((x,x',y)|\mathbf{\Omega}). \tag{130}$$

We define $\Theta_{\rm dd}$ as the set of all triplets $(x,x',y)\in\Theta_{\rm d}\times\Theta_{\rm d}\times\Theta_{\rm mm}$ that satisfy (130) for some $(\iota,\iota')\in\hat{\mathfrak{D}}^2_{\neq}$ and $s,s'\in\{-1,1\}$.

A.7 Proof of Theorem 3

Consider $\ell\in\{\rm d,md,mm,dd\}$ and $z,z'\in\Theta_\ell$, and denote

$$\Delta_{\ell,\mathbf{\Omega}}(z,z'):=\left|\Psi_\ell(z|\mathbf{\Omega})-\Psi_\ell(z'|\mathbf{\Omega})\right|.$$

The result will follow if we exhibit $z_1,\dots,z_T$, $2\le T\le 6$ such that $z_1=z$, $z_T=z'$ such that

$$\max_{1\le t\le T-1}\Delta_\ell(z_t,z_{t+1})\le\Delta_\ell(z,z') \tag{131}$$

and if we can find smooth functions $u\in[0,1]\mapsto f_{\ell,t}(u|\omega)$ such that for every $\omega\in\mathbb{R}^d$ and every t

$$f_{\ell,t}(0|\omega)=f_\ell(z_t|\omega),\quad f_{\ell,t}(1|\omega)=f_\ell(z_{t+1}|\omega)\,, \tag{132}$$

$$\sup_{0<u<1}|f'_{\ell,t}(u|\omega)|\le G(\omega)\Delta_\ell(z_t,z_{t+1}) \tag{133}$$

$$\text{where}\quad G(\omega):=C_\kappa\sum_{i=1}^3\|\omega\|^i_{a,\star}. \tag{134}$$

Indeed, this will imply by the triangle inequality, the mean value theorem, and (67) that

$$\begin{aligned}
\Delta_{\ell,\mathbf{\Omega}}(z,z')=\Delta_{\ell,\mathbf{\Omega}}(z_1,z_T)&\le\sum_{t=1}^{T-1}\Delta_{\ell,\mathbf{\Omega}}(z_t,z_{t+1})=\sum_{t=1}^{T-1}\left|\frac1m\sum_{j=1}^m\psi_m(\omega_j)[f_{\ell,t}(1|\omega_j)-f_{\ell,t}(0|\omega_j)]\right|\\
&\le\sum_{t=1}^{T-1}\frac1m\sum_{j=1}^m\psi_m(\omega_j)\left|[f_{\ell,t}(1|\omega_j)-f_{\ell,t}(0|\omega_j)]\right|\le\frac1m\sum_{t=1}^{T-1}\sum_{j=1}^m\psi_m(\omega_j)\sup_{0<u<1}|f'_{\ell,t}(u|\omega_j)|\\
&\le\frac1m\sum_{t=1}^{T-1}\sum_{j=1}^m\psi_m(\omega_j)G(\omega_j)\Delta_\ell(z_t,z_{t+1})=\left(\frac1m\sum_{j=1}^m\psi_m(\omega_j)G(\omega_j)\right)\cdot\sum_{t=1}^{T-1}\Delta_\ell(z_t,z_{t+1})\\
&\le\left(\frac1m\sum_{j=1}^m\psi_m(\omega_j)G(\omega_j)\right)\cdot 6\Delta_\ell(z,z')=6\Psi_0(\mathbf{\Omega})\cdot C_\kappa\cdot\ \Delta_\ell(z,z').
\end{aligned}$$

A.7.1 Construction of z_t, $1 \le t \le T$

First we focus on the construction of $z_1, \dots, z_T$ satifying the inequality (131):

- When $\ell = \mathrm{d}$, we have $z = x$ and $z' = x'$ where $x, x' \in \Theta_{\mathrm{d}}$ (cf (59)). Observe that $r := \|x\|_a, r' := \|x'\|_a$ satisfy $r, r' > 0$, and that $n := x/r$, $n' := x'/r'$ satisfy $\|n\|_a = \|n'\|_a = 1$. We set $z_1 = x$, $z_2 = \tilde{x}$, $z_3 = x'$ where $\tilde{x} := \|x'\|_a \cdot (x/\|x\|_a)$ satisfies $\|\tilde{x}\|_a = \|x'\|_a$. Given the definition (68) we have $\Delta_\ell(z_1, z_2) = |r - r'| = |\|x\|_a - \|x'\|_a| \le \Delta_\ell(x, x') = \Delta_\ell(z, z')$ and similarly $\Delta_\ell(z_2, z_3) = \|n - n'\|_a \le \Delta_\ell(z, z')$ as claimed.
- When $\ell = \mathrm{mm}$, $z = y$, $z' = y'$, where $y, y' \in \Theta_{\mathrm{mm}}$ and we simply set $z_1 = z$, $z_2 = z'$.
- When $\ell = \mathrm{md}$, $z = (x, y)$ and $z' = (x', y')$ where $x, x' \in \Theta_{\mathrm{d}}$, $y, y' \in \Theta_{\mathrm{mm}}$. Setting $\tilde{x}$ as above we define $z_1 = z = (x, y)$, $z_2 = (x, y')$, $z_3 = (\tilde{x}, y')$ and $z_4 = (x', y')$. It is not difficult to check that the inequality (131) holds given the definition (70) of Δ_{md}.
- Finally, when $\ell = \mathrm{dd}$, $z = (x_1, x_2, y)$ and $z' = (x'_1, x'_2, y')$ where $x_i \in \Theta_d$ and $y, y' \in \Theta_{\mathrm{mm}}$. It is easy to check (131) with $z_1 = (x_1, x_2, y)$, $z_2 = (x_1, x_2, y')$, $z_3 = (\tilde{x}_1, x_2, y')$, $z_4 = (x'_1, x_2, y')$, $z_5 = (x'_1, \tilde{x}_2, y')$, $z_6 = (x'_1, x'_2, y')$ given the definition (71) of Δ_{dd}.

To complete the proof of Theorem 3 we now exhibit $f_{\ell,t}$ satisfying (132)-(133) by treating each case $\ell = \mathrm{d}, \ell = \mathrm{mm}, \ell = \mathrm{md}, \ell = \mathrm{dd}$ and pair (z_t, z_{t+1}). First we build the functions and show that they satisfy (132). Then observing their common structure we establish the bound (133).

A.7.2 Construction of $f_{\ell,t}$ satisfying (132)

The case $\ell = \mathrm{d}$. For $u \in (0,1)$, we define $\bar{r}(u) := r + u(r' - r)$ and $\bar{n}(u) := n + u(n' - n)$, which satisfies $\|\bar{n}(u)\|_a = \|(1-u)n + un'\|_a \le (1-u)\|n\|_a + u\|n'\|_a = 1$, and we set

$$f_{\mathrm{d},1}(u|\omega) := 2\frac{\sin^2(\langle \omega, \bar{r}(u)n\rangle/2)}{1 - \tilde{\kappa}(\bar{r}(u))} \overset{(64)}{=} 2\alpha^2(\bar{r}(u))\frac{\sin^2(\bar{r}(u)\langle \omega, n\rangle/2)}{(\bar{r}(u))^2}, \tag{135}$$

$$f_{\mathrm{d},2}(u|\omega) := \frac{2\sin^2(r'\langle \omega, \bar{n}(u)\rangle/2)}{1 - \tilde{\kappa}(r')} \overset{(64)}{=} 2\alpha^2(r')\frac{\sin^2(r'\langle \omega, \bar{n}(u)\rangle/2)}{(r')^2} \tag{136}$$

Property (132) follows from (53) and (63) since $z_1 = x = r \cdot n = \bar{r}(0) \cdot n$, $z_2 = \tilde{x} = r' \cdot n = \bar{r}(1) \cdot n = r'\bar{n}(0)$, $z_3 = x' = r' \cdot n' = r' \cdot \bar{n}(1)$ and $\|n\|_a = \|n'\|_a = 1$ so that $\bar{\kappa}(z_2) = \bar{\kappa}(z_3) = \tilde{\kappa}(r')$.

The case $\ell = \mathrm{mm}$. For $u \in (0,1)$, we define $\bar{y}(u) := y + u(y' - y)$ and

$$f_{\mathrm{mm},1}(u|\omega) := \cos(\langle \omega, \bar{y}(u)\rangle). \tag{137}$$

Property (132) follows by (54) since $z_1 = y = \bar{y}(0)$ and $z_2 = y' = \bar{y}(1)$.

The case $\ell = md$. For any x, y, ω, by (55), (63), (64), since $2\sin v \sin w = \cos(v - w) - \cos(v + w)$

$$f_{\mathrm{md}}(x, y|\omega) = \sqrt{2}\frac{\sin\left(\langle\omega, x\rangle/2\right)\sin\left(\langle\omega, y + x/2\rangle\right)}{\sqrt{1 - \tilde{\kappa}(\|x\|_a)}} = \sqrt{2}\alpha(\|x\|_a)\frac{\sin\left(\langle\omega, x\rangle/2\right)\sin\left(\langle\omega, y + x/2\rangle\right)}{\|x\|_a}$$
$$= \frac{1}{\sqrt{2}}\alpha(\|x\|_a)\frac{\cos(\langle\omega, y\rangle) - \cos(\langle\omega, y\rangle + \langle\omega, x\rangle)}{\|x\|_a}.$$

Since $z_1 = (x, y) = (x, \bar{y}(0))$, $z_2 = (x, y') = (x, \bar{y}(1)) = (\bar{r}(0) \cdot n, y')$, $z_3 = (\tilde{x}, y') = (\bar{r}(1) \cdot n, y') = (r' \cdot \bar{n}(0), y')$, $z_4 = (x', y') = (r' \cdot \bar{n}(1), y')$, Property (132) holds with

$$f_{\mathrm{md},1}(u|\omega) := \sqrt{2}\alpha(r)\frac{\sin\left(r\langle\omega, n\rangle/2\right)\sin\left(\langle\omega, \bar{y}(u) + x/2\rangle\right)}{r}, \tag{138}$$

$$f_{\mathrm{md},2}(u|\omega) := \sqrt{2}\alpha(\bar{r}(u))\frac{\sin\left(\bar{r}(u)\langle\omega, n\rangle/2\right)\sin\left(\langle\omega, y' + \bar{r}(u) \cdot n/2\rangle\right)}{\bar{r}(u)}, \tag{139}$$

$$f_{\mathrm{md},3}(u|\omega) := \frac{1}{\sqrt{2}}\alpha(r')\frac{\cos(\langle\omega, y'\rangle) - \cos(\langle\omega, y'\rangle + r'\langle\omega, \bar{n}(u)\rangle)}{r'}. \tag{140}$$

The case $\ell = \mathrm{dd}$. For any x_1, x_2, y, ω, (56) yields using (63)-(64)

$$f_{\mathrm{dd}}((x_1, x_2, y)|\omega) = 2\frac{\sin\left(\langle\omega, x_1\rangle/2\right)\sin\left(\langle\omega, x_2\rangle/2\right)\cos\left(\langle\omega, y + x_2/2 - x_1/2\rangle\right)}{\sqrt{1 - \tilde{\kappa}(\|x_1\|_a)}\sqrt{1 - \tilde{\kappa}(\|x_2\|_a)}}$$
$$= 2\alpha(\|x_1\|_a)\alpha(\|x_2\|_a)\frac{\sin\left(\langle\omega, x_1\rangle/2\right)\sin\left(\langle\omega, x_2\rangle/2\right)\cos\left(\langle\omega, y + x_2/2 - x_1/2\rangle\right)}{\|x_1\|_a\|x_2\|_a}$$

Reasoning as above establishes Property (132) holds with $z_1 = (x_1, x_2, y)$, $z_2 = (x_1, x_2, y')$, $z_3 = (\tilde{x}_1, x_2, y')$, $z_4 = (x_1', x_2, y')$, $z_5 = (x_1', \tilde{x}_2, y')$, $z_6 = (x_1', x_2', y')$, $\bar{y}(u) := y + u(y' - y)$, where $\tilde{x}_i, r_i, r_i', \bar{r}_i(u)$, $n_i, n_i', \bar{n}_i(u)$, $i \in \{1, 2\}$ were defined in the same way as in the case $\ell = \mathrm{d}$, and

$$f_{\mathrm{dd},1}(u|\omega) := 2\alpha(r_1)\alpha(r_2)\frac{\sin\left(r_1\langle\omega, n_1\rangle/2\right)\sin\left(r_2\langle\omega, n_2\rangle/2\right)\cos\left(\langle\omega, \bar{y}(u) + r_2n_2/2 - r_1n_1/2\rangle\right)}{r_1r_2} \tag{141}$$

$$f_{\mathrm{dd},2}(u|\omega) := 2\alpha(\bar{r}_1(u))\alpha(r_2)\frac{\sin\left(\bar{r}_1(u)\langle\omega, n_1\rangle/2\right)\sin\left(r_2\langle\omega, n_2\rangle/2\right)\cos\left(\langle\omega, y' + r_2n_2/2 - \bar{r}_1(u) \cdot n_1/2\rangle\right)}{\bar{r}_1(u)r_2} \tag{142}$$

$$f_{\mathrm{dd},3}(u|\omega) := 2\alpha(r_1')\alpha(r_2)\frac{\sin\left(r_1'\langle\omega, \bar{n}_1(u)\rangle/2\right)\sin\left(r_2\langle\omega, n_2\rangle/2\right)\cos\left(\langle\omega, y' + r_2n_2/2 - r_1' \cdot \bar{n}_1(u)/2\rangle\right)}{r_1'r_2} \tag{143}$$

$$f_{\mathrm{dd},4}(u|\omega) := 2\alpha(r_1')\alpha(\bar{r}_2(u))\frac{\sin\left(r_1'\langle\omega, n_1'\rangle/2\right)\sin\left(\bar{r}_2(u)\langle\omega, n_2\rangle/2\right)\cos\left(\langle\omega, y' + \bar{r}_2(u)n_2/2 - r_1'n_1'/2\rangle\right)}{r_1'\bar{r}_2(u)} \tag{144}$$

$$f_{\mathrm{dd},5}(u|\omega) := 2\alpha(r_1')\alpha(r_2')\frac{\sin\left(r_1'\langle\omega, n_1'\rangle/2\right)\sin\left(r_2'\langle\omega, \bar{n}_2(u)\rangle/2\right)\cos\left(\langle\omega, y' + r_2'\bar{n}_2(u)/2 - r_1'n_1'/2\rangle\right)}{r_1'r_2'} \tag{145}$$

A.7.3 Proof of the bound (133)

To continue we gather a few observations. First, since $\operatorname{sinc}(t) := \sin(t)/t = \int_0^1 \cos(xt)dx$ for every $t \neq 0$ (and $\operatorname{sinc}(0) = 1$) we have $\operatorname{sinc}'(t) = \int_0^1 -x\sin(xt)dx$ hence $\max(|\operatorname{sinc}(t)|, |\operatorname{sinc}'(t)|) \leq 1$. Now, by definition of the dual norm $\|\cdot\|_{a,\star}$, for every $\omega, v \in \mathbb{R}^d$ we have $|\langle \omega, v\rangle| \leq \|\omega\|_{a,\star}\|v\|_a$. Thus, by definition (65) of C_κ for every $0 < t \leq R$ we have

$$\forall v \in \mathbb{R}^d, \quad |\alpha(t)\sin(t\langle\omega, v\rangle/2)/t| = |\alpha(t)\tfrac{\langle\omega,v\rangle}{2}\operatorname{sinc}(t\langle\omega, v\rangle/2)| \leq \frac{\sqrt{C_\kappa}}{2}\|\omega\|_{a,\star}\|v\|_a. \tag{146}$$

Recalling that $\bar{y}(u) := y + u(y' - y)$, $\bar{n}(u) := n + u(n' - n)$, $\bar{r}(u) := r + u(r' - r)$, and $\omega \in \mathbb{R}^d$ we now bound the following auxiliary functions and their derivatives, with arbitrary $\phi \in \mathbb{R}$

$$g_{0,\phi}(u) := \cos(\langle\omega, \bar{y}(u)\rangle + \phi)\,,\ g_1(u) := \alpha(\bar{r}(u))\frac{\sin(\bar{r}(u)\langle\omega, n\rangle/2)}{\bar{r}(u)}\,,\ g_2(u) := \alpha(r')\frac{\sin(r'\langle\omega, \bar{n}(u)\rangle/2)}{r'}\,.$$

Since $0 < r' \leq R$ we have

$$|g'_{0,\phi}(u)| = |\sin(\langle\omega, \bar{y}(u)\rangle + \phi)\cdot\langle\omega, y' - y\rangle| \leq |\langle\omega, y' - y\rangle| \leq \|\omega\|_{a,\star}\cdot\|y' - y\|_a\,,$$
$$|g'_2(u)| = |\alpha(r')\cos(r'\langle\omega, \bar{n}(u)\rangle/2)\cdot\tfrac{\langle\omega,n'-n\rangle}{2}| \leq \frac{\sqrt{C_\kappa}\|\omega\|_{a,\star}}{2}\cdot\|n' - n\|_a\,.$$

As $g_1(u) = \alpha(\bar{r}(u))\frac{\langle\omega,n\rangle}{2}\operatorname{sinc}(\bar{r}(u)\langle\omega, n\rangle/2)$, $\|n\|_a = 1$, and $\bar{r}(u) \leq \max(r, r') \leq R$ we get

$$|g'_1(u)| = \left|\alpha'(\bar{r}(u))\operatorname{sinc}(\bar{r}(u)\langle\omega, n\rangle/2) + \alpha(\bar{r}(u))\frac{\langle\omega, \cdot n\rangle}{2}\operatorname{sinc}'(\bar{r}(u)\langle\omega, n\rangle/2)\right|\cdot\left|\frac{\langle\omega, n\rangle}{2}\cdot(r' - r)\right|$$
$$\leq \frac{\sqrt{C_\kappa}\|\omega\|_{a,\star}}{2}\cdot(1 + \|\omega\|_{a,\star}/2)\cdot|r' - r|$$

Since $\|\bar{n}(u)\|_a \leq \max(\|n\|_a, \|n'\|_a) = 1$ we also get using (146)

$$\max(|g_1(u)|, |g_2(u)|) \leq \frac{\sqrt{C_\kappa}}{2}\|\omega\|_{a,\star}\max(\|n\|_a, \|\bar{n}(u)\|_a) = \frac{\sqrt{C_\kappa}}{2}\|\omega\|_{a,\star}\,.$$

We are now equipped to proceed.

The case $\ell = \mathrm{d}$. Expressions (135)-(136) yield $f_{\mathrm{d},1}(u|\omega) = 2g_1^2(u)$, $f_{\mathrm{d},2}(u|\omega) = 2g_2^2(u)$. By (68) and the choice of $z_1 = x, z_2 = \tilde{x}, z_3 = x'$ we have $\Delta_d(z_1, z_2) = |r' - r|$, $\Delta_d(z_2, z_3) = \|n' - n\|_a$ and we obtain the bound (133) since

$$|f'_{\mathrm{d},1}(u)| = |4g_1(u)g'_1(u)| \leq C_\kappa(\|\omega\|_{a,\star}^2 + \|\omega\|_{a,\star}^3/2)\cdot|r' - r| \leq G(\omega)\cdot\Delta_d(z_1, z_2)\,,$$
$$|f'_{\mathrm{d},2}(u)| = |4g_2(u)g'_2(u)| \leq C_\kappa\|\omega\|_{a,\star}^2\cdot\|n' - n\|_a \leq G(\omega)\cdot\Delta_d(z_2, z_3)\,.$$

The case $\ell = \mathrm{mm}$. By the expression (137) we have $f_{\mathrm{mm},1} = g_0$. The bound (133) follows from

$$|f'_{\mathrm{mm},1}(u)| = |g'_{0,0}(u)| \leq \|\omega_j\|_{a,\star}\|y' - y\|_a \overset{(65)\&(69)}{\leq} G(\omega)\Delta_{\mathrm{mm}}(z_1, z_2).$$

The case $\ell = \mathrm{md}$. By the identity $\cos(\theta - \pi/2) = \sin(\theta)$ we have $f_{\mathrm{md},1}(u) \overset{(138)}{=} \sqrt{2} g_1(0) g_{0,\phi}(u)$ with $\phi := \langle \omega, x\rangle/2 - \pi/2$, and by (139) $f_{\mathrm{md},2}(u) = \sqrt{2} g_1(u)\sin(\psi(u))$ with $\psi(u) := \langle \omega, y' + \bar{r}(u) n/2\rangle$, and $\psi'(u) = \langle \omega, n\rangle (r' - r)/2$. Combining with (140) we obtain the bound (133) since

$$
\begin{aligned}
|f'_{\mathrm{md},1}(u)| &= |\sqrt{2} g_1(0) g'_{0,\phi}(u)| \le \frac{\sqrt{C_\kappa}}{\sqrt{2}} \|\omega_j\|_{a,\star}^2 \|y' - y\|_a \overset{C_\kappa \ge 1 \& (70)}{\le} G(\omega) \Delta_{\mathrm{md}}(z_1, z_2)\,, \\
|f'_{\mathrm{md},2}(u)| &= \sqrt{2} |g'_1(u)\sin(\psi(u)) + g_1(u)\cos(\psi(u))\langle \omega, n\rangle (r' - r)/2| \\
&\le \sqrt{2}\left(\frac{\sqrt{C_\kappa}\|\omega\|_{a,\star}}{2} (1 + \|\omega\|_{a,\star}/2)|r' - r| + \frac{\sqrt{C_\kappa}\|\omega\|_{a,\star}}{2} \|\omega\|_{a,\star} |r' - r|/2 \right) \\
&\le \frac{C_\kappa \|\omega\|_{a,\star}}{\sqrt{2}} (1 + \|\omega\|_{a,\star}) \cdot |r' - r| \le G(\omega) \Delta_{\mathrm{md}}(z_2, z_3)\,, \\
|f'_{\mathrm{md},3}(u)| &= \frac{1}{\sqrt{2}} |\alpha(r') \sin(\langle \omega, y'\rangle + r' \langle \omega, \bar{n}(u)\rangle) \cdot \langle \omega, n' - n\rangle| \\
&\le \frac{\sqrt{C_\kappa}\|\omega\|_{a,\star}}{\sqrt{2}} \cdot \|n' - n\|_a \le G(\omega) \Delta_{\mathrm{md}}(z_2, z_3)\,.
\end{aligned}
$$

The case $\ell = \mathrm{dd}$ Denote $g_{i,j}$, $i, j = 1, 2$ the functions defined as g_i with r_j, r'_j, etc. instead of r, r' etc. By (141) we have $f_{\mathrm{dd},1}(u) = 2 g_{1,1}(0) g_{1,2}(0) g_{0,\phi}(u)$ with $\phi := r_2 n_2/2 - r_1 n_1/2$, and by (142), $f_{\mathrm{dd},2}(u) = 2 g_{1,1}(u) g_{1,2}(0)\cos(\psi(u))$ with $\psi_2(u) := \langle \omega, y' + r_2 n_2/2\rangle - \bar{r}_1(u)\langle \omega, n_1\rangle/2$, $\psi'_2(u) = -\langle \omega, n_1\rangle (r'_1 - r_1)/2$ hence

$$
\begin{aligned}
|f'_{\mathrm{dd},1}(u)| &= |2 g_{1,1}(0) g_{1,2}(0) g'_{0,\phi}(u)| \le \frac{C_\kappa}{2} \|\omega\|_{a,\star}^3 \cdot \|y' - y\|_a \le G(\omega) \Delta_{\mathrm{dd}}(z_1, z_2)\,, \\
|f'_{\mathrm{dd},2}(u)| &= |2 g_{1,2}(0)| \cdot |g'_{1,1}(u)\cos(\psi(u)) + g_{1,1}(u)\sin(\psi(u))\langle \omega, n_1\rangle (r'_1 - r_1)/2| \\
&\le \sqrt{C_\kappa}\|\omega\|_{a,\star} \left(\frac{\sqrt{C_\kappa}\|\omega\|_{a,\star}}{2} (1 + \|\omega\|_{a,\star}/2)|r'_1 - r_1| + \frac{\sqrt{C_\kappa}\|\omega\|_{a,\star}}{2} \|\omega\|_{a,\star} |r'_1 - r_1|/2 \right) \\
&\le \frac{C_\kappa}{2} \|\omega\|_{a,\star}^2 (1 + 3\|\omega\|_{a,\star}/4)\, |r'_1 - r_1| \le G(\omega) \Delta_{\mathrm{dd}}(z_2, z_3)\,.
\end{aligned}
$$

Similarly by (144) $f_{\mathrm{dd},4}(u) = 2 g_{1,1}(1) g_{1,2}(u)\cos(\psi_4(u))$ with $\psi_4(u) := \langle \omega, y' - r'_1 n'_1/2\rangle + \bar{r}_2(u)\langle \omega, n_2\rangle/2$ and the same reasoning yields the same bound. This establishes the bound (133) for $t = 1, 2, 4$.

By (143) and the identity $2\sin v \cos w = \sin(v + w) + \sin(v - w)$ we write

$$
\begin{aligned}
f_{\mathrm{dd},3}(u) &= g_{1,2}(0) \frac{\alpha(r'_1)}{r'_1} 2\sin\left(r'_1 \langle \omega, \bar{n}_1(u)\rangle/2\right) \cos\left(\langle \omega, y' + r_2 n_2/2 - r'_1 \cdot \bar{n}_1(u)/2\rangle\right) \\
&= g_{1,2}(0) \frac{\alpha(r'_1)}{r'_1} \Big(\sin(\langle \omega, y' + r_2 n_2/2\rangle) + \sin(\underbrace{r'_1 \langle \omega, \bar{n}_1(u)\rangle - \langle \omega, y' + r_2 n_2/2\rangle}_{\psi_3(u)}) \Big) \\
|f'_{\mathrm{dd},3}(u)| &= |g_{1,2}(0) \frac{\alpha(r'_1)}{r'_1} \cos(\psi_3(u)) \psi'_3(u)| \le \frac{\sqrt{C_\kappa}\|\omega\|_{a,\star}}{2} \frac{\sqrt{C_\kappa}}{r'_1} |\psi'_3(u)| = \frac{C_\kappa \|\omega\|_{a,\star}}{2} |\langle \omega, n'_1 - n_1\rangle| \\
&\le \frac{C_\kappa \|\omega\|_{a,\star}^2}{2} \cdot \|n'_1 - n_1\|_a \le G(\omega) \Delta_{\mathrm{dd}}(z_3, z_4).
\end{aligned}
$$

The same reasoning works from (145) for $t = 5$. This establishes the bound (133) for $t = 3, 5$.

A.7.4 Proof of Proposition 6

We denote by $\mathbb{B}_a$ (resp. $\mathbb{S}_a$) the unit ball (resp. unit sphere) with respect to $\|.\|_a$, $R_d := \sup_{x\in\Theta_d} \|x\|_a$, and $R_{\rm mm} := \sup_{x\in\Theta_{\rm mm}} \|x\|_a$. Observe that

$$\max(R_d, R_{\rm mm}) = \sup_{x\in\Theta_d\cup\Theta_{\rm mm}} \|x\|_a \overset{(59)-(60)}{=} \sup_{x\in\Theta-\Theta} \|x\|_a = \mathrm{diam}_a(\Theta) =: D.$$

An upper bound of $\mathcal{N}_d(\tau)$. Denote $I := (0, R_d] \subseteq \mathbb{R}$. We will soon show that

$$\mathcal{N}_d(\tau) := \mathcal{N}(\Theta_d, \Delta_d, \tau) \le \mathcal{N}(I, |\cdot|, \tau/2) \times \mathcal{N}\big(\mathbb{S}_a, \|\cdot\|_a, \tfrac{\tau}{2(R_d+1)}\big). \tag{147}$$

By [GBKT21b, Lemma A.1], since $\mathbb{S}_a \subset \mathbb{B}_a$, we have $\mathcal{N}\big(\mathbb{S}_a, \|\cdot\|_a, \frac{\tau}{2(R_d+1)}\big) \le \mathcal{N}\big(\mathbb{B}_a, \|.\|_a, \frac{\tau}{4(R_d+1)}\big)$. Moreover, by [Wai19, Lemma 5.7] the inequality $\mathcal{N}\big(\mathbb{B}_a, \|.\|_a, \frac{\tau}{(4(R_d+1)}\big) \le \big(1+8(R_d+1)/\tau\big)^d$ is valid for every $\tau > 0$. Finally, since $\mathcal{N}(I, |\cdot|, \tau/2) \le 1+2R_d/\tau \le 1 + 8(R_d+1)/\tau$ we obtain

$$\mathcal{N}_d(\tau) \le \big(1 + 8(R_d+1)/\tau\big)^{d+1} \le \big(1 + 8(D+1)/\tau\big)^{d+1}.$$

We now establish (147). Denote $N_1 := \mathcal{N}(I, |\cdot|, \tau/2)$, $N_2 := \mathcal{N}(\mathbb{S}_a, \|\cdot\|_a, \tau/(2(R_d+1)))$, and consider $(r_i)_{i\in[N_1]}$ a covering of I with respect to $|\cdot|$ at scale $\tau/2$ and $(s_i)_{i\in[N_2]}$ a covering of $\mathbb{S}_a$ with respect to $\|\cdot\|_a$ at scale $\tau/(2(R_d+1))$. We show that the family $(r_i \cdot s_j)_{(i,j)\in[N_1]\times[N_2]}$ is a covering of Θ_d with respect to the metric Δ_d at scale τ. For this, consider an arbitrary $x \in \Theta_d$ (recall the definition (59)) and define $r := \|x\|_a$ and $n := x/r$. By definition of R_d and I we have $r \in I$, and $\|n\|_a \in \mathbb{S}_a$, hence there are $i \in [N_1]$, $j \in [N_2]$ such that $|r - r_i| \le \tau/2$ and $\|n - s_j\|_a \le \tau/(2(R_d+1))$. To reach the conclusion we show that $\Delta_d(x, r_i s_j) \le \tau$. Indeed

$$\begin{aligned}\Delta_d(x, r_i s_j) = \Delta_d(rn, r_i s_j) \overset{(68)}{=} \|rn - r_i s_j\|_a + \|n - s_j\|_a &\le \|rn - r_i n\|_a + \|r_i n - r_i s_j\|_a + \|n - s_j\|_a \\ &\le |r - r_i|\|n\|_a + (r_i+1)\|n - s_j\|_a \le \tau/2 + (R_d+1)\tau/(2(R_d+1)) \le \tau.\end{aligned}$$

An upper bound of $\mathcal{N}_{\rm mm}(\tau)$. By definition of $R_{\rm mm}$ we have $\Theta_{\rm mm} \subset R_{\rm mm} \cdot \mathbb{B}_a$. By [GBKT21b, Lemma A.1] [Wai19, Lemma 5.7] and (69) we have

$$\begin{aligned}\mathcal{N}_{\rm mm}(\tau) := \mathcal{N}(\Theta_{\rm mm}, \|\cdot\|_a, \tau) &\le \mathcal{N}\big(R_{\rm mm} \cdot \mathbb{B}_a, \|.\|_a, \tau/2\big) = \mathcal{N}\big(\mathbb{B}_a, \|.\|_a, \tau/(2R_{\rm mm})\big) \\ &\le \big(1 + 4R_{\rm mm}/\tau\big)^d \le \big(1 + 4D/\tau\big)^d.\end{aligned}$$

An upper bound of $\mathcal{N}_{\rm md}(\tau)$. By (61), we have $\Theta_{\rm md} \subset \Theta_d \times \Theta_{\rm mm}$. Thus by [GBKT21b, Lemma A.1], we have

$$\mathcal{N}_{\rm md}(\tau) = \mathcal{N}(\Theta_{\rm md}, \Delta_{\rm md}, \tau) \le \mathcal{N}(\Theta_d \times \Theta_{\rm mm}, \Delta_{\rm md}, \tau/2).$$

Now, given the definition (70) of $\Delta_{\rm md}$ we have

$$\mathcal{N}_{\rm md}(\tau) \le \mathcal{N}(\Theta_d \times \Theta_{\rm mm}, \Delta_{\rm md}, \tau/2) \le \mathcal{N}(\Theta_d, \Delta_d, \tau/4)\mathcal{N}(\Theta_{\rm mm}, \Delta_{\rm mm}, \tau/4). \tag{148}$$

Indeed, with $(x_i)_{i\in[\mathcal{N}_d(\tau/4)]}$ a covering of Θ_d with respect to Δ_d at scale $\tau/4$ and $(y_j)_{j\in[\mathcal{N}_{\rm mm}(\tau/4)]}$ a covering of $\Theta_{\rm mm}$ with respect to $\Delta_{\rm mm}$ at scale $\tau/4$ it is straightforward to show using (70) that $(x_i, y_j)_{(i,j)\in[\mathcal{N}_d(\tau/4)]\times[\mathcal{N}_{\rm mm}(\tau/4)]}$ covers $\Theta_d \times \Theta_{\rm mm}$ with respect to $\Delta_{\rm md}$ at scale $\tau/2$. Combined with the above estimates we get

$$\mathcal{N}_{\rm md}(\tau) \le \mathcal{N}_d(\tau/4)\mathcal{N}_{\rm mm}(\tau/4) \le \big(1 + 32(D+1)/\tau\big)^{d+1}\big(1 + 16D/\tau\big)^d.$$

An upper bound of $\mathcal{N}_{\mathrm{dd}}(\tau)$ By (62), we have $\Theta_{\mathrm{dd}} \subset \Theta_d \times \Theta_d \times \Theta_{\mathrm{mm}}$, thus a similar argument yields

$$\begin{aligned}\mathcal{N}_{\mathrm{dd}}(\tau) \le \mathcal{N}(\Theta_d \times \Theta_d \times \Theta_{\mathrm{mm}}, \Delta_{\mathrm{dd}}, \tau/2) &\overset{(71)}{\le} \mathcal{N}(\Theta_d, \Delta_d, \tau/8)^2 \mathcal{N}(\Theta_{\mathrm{mm}}, \Delta_{\mathrm{mm}}, \tau/4) \\ &= \mathcal{N}_d(\tau/8)\mathcal{N}_d(\tau/8)\mathcal{N}_{\mathrm{mm}}(\tau/4) \\ &\le \big(1 + 64(D+1)/\tau\big)^{2(d+1)} \big(1 + 16D/\tau\big)^d \\ &\le \big(1 + 64(D+1)/\tau\big)^{3d+2}.\end{aligned}$$

To conclude, observe that the last bound dominates all the previous ones.

A.8 Proof of Theorem 4

Since the average marginal density of the ω_j's satisfies (33), we have

$$\mathbb{E}\Psi_{\mathrm{m}}(\boldsymbol{\Omega}) = 1 \tag{149}$$

$$\mathbb{E}\Psi_{\mathrm{d}}(z|\boldsymbol{\Omega}) = 1, \quad \forall z \in \Theta_{\mathrm{d}} \tag{150}$$

$$|\mathbb{E}\Psi_\ell(z|\boldsymbol{\Omega})| \le \mu \quad \forall \ell \in \{\mathrm{mm}, \mathrm{md}, \mathrm{dd}\},\ \forall z \in \Theta_\ell. \tag{151}$$

The proof of (151) ((149) and (150) are obtained similarly) is postponed to the end of this section. By (151) we have

$$|\Psi_\ell(z|\boldsymbol{\Omega})| \le |\Psi_\ell(z|\boldsymbol{\Omega}) - \mathbb{E}\Psi_\ell(z|\boldsymbol{\Omega})| + \mu, \quad \forall \ell \in \{\mathrm{mm}, \mathrm{md}, \mathrm{dd}\},\ \forall z \in \Theta_\ell$$

hence (74),(75),(76) yield

$$\mathbb{P}\Big(|\Psi_{\mathrm{m}}(\boldsymbol{\Omega}) - 1| > \frac{\tau}{4}\Big) \le 2\exp\Big(-\frac{m}{v}\Big), \tag{152}$$

$$\forall z \in \Theta_{\mathrm{d}},\ \mathbb{P}\Big(|\Psi_{\mathrm{d}}(z|\boldsymbol{\Omega}) - 1| > \frac{\tau}{8}\Big) \le 2\exp\Big(-\frac{m}{v}\Big), \tag{153}$$

$$\forall \ell \in \{\mathrm{mm}, \mathrm{md}, \mathrm{dd}\},\ \forall z \in \Theta_\ell,\ \mathbb{P}\Big(|\Psi_\ell(z|\boldsymbol{\Omega})| > \mu + \frac{\tau}{16k}\Big) \le 2\exp\Big(-\frac{m}{v}\Big). \tag{154}$$

Now, observe that the mutual incoherence assumption implies that κ is locally characteristic with respect to $\mathcal{T}$ and that its $2k$-coherence is bounded by $c = (2k-1)\mu$. Since $\kappa \ge 0$, all assumptions of Propositions 3 to 5 and Theorem 3 thus hold. For each $\ell \in \{\mathrm{d}, \mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$ consider a covering $(z_i^\ell)_{i \in [\mathcal{N}_\ell(\tau')]}$ of Θ_ℓ with respect to Δ_ℓ at scale

$$\tau' := \frac{\tau}{96Mk \cdot C_\kappa}. \tag{155}$$

By Theorem 3, if

$$\Psi_0(\boldsymbol{\Omega}) \le M\,, \tag{156}$$

$$\max_{i \in [\mathcal{N}_{\mathrm{d}}(\tau')]} |\Psi_{\mathrm{d}}(z_i^{\mathrm{d}}|\boldsymbol{\Omega}) - 1| \le \tau/8 \tag{157}$$

$$\max_{\ell \in \{\mathrm{mm},\mathrm{md},\mathrm{dd}\}} \max_{i \in [\mathcal{N}_\ell(\tau')]} |\Psi_\ell(z_i^\ell|\boldsymbol{\Omega})| \le \mu + \frac{\tau}{16k} \tag{158}$$

then by (66), for each $\ell \in \{\mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$ and $z \in \Theta_\ell$ there is $i \in [\mathcal{N}_\ell(\tau')]$ such that

$$\begin{aligned}|\Psi_\ell(z|\boldsymbol{\Omega})| \le |\Psi_\ell(z_i^\ell|\boldsymbol{\Omega})| + |\Psi_\ell(z|\boldsymbol{\Omega}) - \Psi_\ell(z_i^\ell|\boldsymbol{\Omega})| &\le \mu + \frac{\tau}{16k} + 6M \cdot C_\kappa \cdot \underbrace{\Delta_\ell(z, z_i^\ell)}_{\le \tau'} \\ &\le \mu + \frac{\tau}{16k} + \frac{\tau}{16k} \le \mu + \frac{\tau}{8k}\,,\end{aligned}$$

and similarly for each $z \in \Theta_{\mathrm{d}}$ there is $i \in [\mathcal{N}_{\mathrm{d}}(\tau')]$ such that

$$|\Psi_{\mathrm{d}}(z|\mathbf{\Omega}) - 1| \leq |\Psi_\ell(z_i^{\mathrm{d}}|\mathbf{\Omega}) - 1| + |\Psi_{\mathrm{d}}(z|\mathbf{\Omega}) - \Psi_\ell(z_i^{\mathrm{d}}|\mathbf{\Omega})| \leq \frac{\tau}{8} + \frac{\tau}{8} \leq \frac{\tau}{4}.$$

Thus, when (156)-(157)-(158) hold we have

$$\sup_{z\in\Theta_{\mathrm{d}}} |\Psi_{\mathrm{d}}(z|\mathbf{\Omega}) - 1| \leq \frac{\tau}{4} \quad \text{and} \quad \max_{\ell\in\{\mathrm{mm},\mathrm{md},\mathrm{dd}\}} \sup_{z\in\Theta_\ell} |\Psi_\ell(z|\mathbf{\Omega})| \leq \mu + \frac{\tau}{8k}.$$

If in addition we have

$$|\Psi_{\mathrm{m}}(\mathbf{\Omega}) - 1| \leq \frac{\tau}{4} \tag{159}$$

then by Proposition 5 and the bound (51), which follows from Propositions 3 and 4, we obtain

$$\begin{aligned}
\max\big(\delta(\mathfrak{M}|\mathcal{A}), \delta(\hat{\mathfrak{D}}|\mathcal{A})\big) &\leq \tau/4, \\
(2k-1)\max\big(\mu(\mathfrak{M}^2_{\neq}|\mathcal{A}), \mu(\hat{\mathfrak{D}}^2_{\neq}|\mathcal{A}), \mu(\mathfrak{M}\times\hat{\mathfrak{D}}_{\neq}|\mathcal{A})\big)\big) &\leq (2k-1)\mu + \tau/4 = c + \tau/4, \\
\delta(\mathcal{S}_k|\mathcal{A}) &\leq \frac{1}{1-c}\big(c + \frac{\tau}{4} + 3(c+\frac{\tau}{4})\big) = \frac{4c+\tau}{1-c}.
\end{aligned}$$

Observe that, since $0 < \tau < 1 - 5c$, we have $4c + \tau < 1 - c$ hence $(4c+\tau)/(1-c) < 1$.

To conclude, we bound the probability p that one of the inequalities (156)-(157)-(158)-(159) fails to hold. Using (155), denoting $D := \mathrm{diam}_a(\Theta)$, we have $1 + 64(D+1)/\tau' = 1 + 6144Mk(D+1)C_\kappa/\tau = 1 + C/\tau$. By a union bound combining (73)-(152)-(153)-(154) and by Proposition 6 we obtain

$$\begin{aligned}
p &\leq 2\exp\Big(-\frac{m}{v}\Big)\cdot\Big(2 + \sum_{\ell\in\{\mathrm{d},\mathrm{mm},\mathrm{md},\mathrm{dd}\}} \mathcal{N}_\ell(\tau')\Big) \\
&\leq 2\exp\Big(-\frac{m}{v}\Big)\cdot\Big(2 + 4\big(1 + 64(D+1)/\tau'\big)^{3d+2}\Big) \leq 12\exp\Big(-\frac{m}{v}\Big)\cdot\big(1 + C/\tau\big)^{3d+2}.
\end{aligned}$$

Proof of (151). Let $\ell \in \{\mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$ and let $z \in \Theta_\ell$. First, by Lemma 6, there exists $(\iota, \iota') \in \mathfrak{D}^2_{\neq}$ and $\xi \in \{-1, 1\}$ such that $\Psi_\ell(z|\mathbf{\Omega}) = \xi\langle\mathcal{A}\iota, \mathcal{A}\iota'\rangle$ for any choice of $\mathbf{\Omega}$ (and of the corresponding sketching operator $\mathcal{A} = \mathcal{A}_{\mathbf{\Omega}}$). In the following, we show that $\mathbb{E}_{\mathbf{\Omega}}\langle\mathcal{A}\iota, \mathcal{A}\iota'\rangle = \langle\iota, \iota'\rangle_\kappa$, so that, by the definition (17) of the mutual coherence with respect to $\mathcal{T}$, we get $|\mathbb{E}_{\mathbf{\Omega}}\Psi_\ell(z|\mathbf{\Omega})| = |\langle\iota, \iota'\rangle_\kappa| \leq \mu$. Remember that $\langle\mathcal{A}\iota, \mathcal{A}\iota'\rangle = \sum_{j=1}^m \langle\phi_{\omega_j}, \iota\rangle\overline{\langle\phi_{\omega_j}, \iota'\rangle}/m$, so that

$$\begin{aligned}
\mathbb{E}_{\mathbf{\Omega}}\langle\mathcal{A}\iota, \mathcal{A}\iota'\rangle = \frac{1}{m}\sum_{j=1}^m \mathbb{E}_{\mathbf{\Omega}}\langle\phi_{\omega_j}, \iota\rangle\overline{\langle\phi_{\omega_j}, \iota'\rangle} &= \frac{1}{m}\sum_{j=1}^m \int_{\mathbb{R}^d} \Lambda_j(\omega)\langle\phi_\omega, \iota\rangle\overline{\langle\phi_\omega, \iota'\rangle}\mathrm{d}\omega \\
&= \int_{\mathbb{R}^d} \frac{1}{m}\sum_{j=1}^m \Lambda_j(\omega)\langle\phi_\omega, \iota\rangle\overline{\langle\phi_\omega, \iota'\rangle}\mathrm{d}\omega \\
&\overset{(33)}{=} \int_{\mathbb{R}^d} w(\omega)^2\hat{\kappa}(\omega)\langle\phi_\omega, \iota\rangle\overline{\langle\phi_\omega, \iota'\rangle}\mathrm{d}\omega = \langle\iota, \iota'\rangle_\kappa.
\end{aligned}$$

A.9 Proof of Theorem 5 and its corollaries

In this section, we prove Theorem 5, Corollary 1 and Corollary 2. We start by establishing Lemma 2, and a few lemmas to deal with sub-exponential random variables.

A.9.1 Proof of Lemma 2

The case $\ell = \mathrm{d}$. With $x = z \in \Theta_d$, $x' := x/\|x\|_a$ we have $\|x'\|_a = 1$. As $|\sin t| \le |t|$ for all t

$$|f_{\mathrm{d}}(z|\omega)| \overset{(53)}{=} 2\frac{\sin^2(\langle\omega, x\rangle)/2}{1-\tilde{\kappa}(\|x\|_a)} \overset{(64)}{=} 2\alpha^2(\|x\|_a)\frac{\sin^2(\langle\omega, x\rangle/2)}{\|x\|_a^2} \overset{(65)}{\le} \frac{1}{2}C_\kappa\langle\omega, x'\rangle^2 \le C_\kappa|\langle\omega, x'\rangle|^2.$$

This establishes (82) with $p = p_{\mathrm{d}} = 2$.

The case $\ell = \mathrm{mm}$. Denote $y = z \in \Theta_{\mathrm{mm}}$. We have $|f_{\mathrm{mm}}(z|\omega)| \overset{(54)}{=} |\cos(\langle\omega, y\rangle)| \le 1$. This establishes (82) with $p = p_{\mathrm{mm}} = 0$.

The case $\ell = \mathrm{md}$. With $(x, y) = z \in \Theta_{\mathrm{md}}$ and $x' := x/\|x\|_a$ we have $\|x'\|_a = 1$ and

$$\begin{aligned}|f_{\mathrm{md}}(z|\omega)| &\overset{(55)}{=} 2\left|\frac{\sin\big(\langle\omega, x/2\rangle\big)\sin\big(\langle\omega, y + x/2\rangle\big)}{\sqrt{2(1-\bar{\kappa}(x))}}\right| \overset{(64)}{\le} \sqrt{2}\alpha(\|x\|_a)\left|\frac{\sin\big(\langle\omega, x/2\rangle\big)}{\|x\|_a}\right| \\ &\overset{(65)}{\le} \sqrt{2}\sqrt{C_\kappa}\left|\Big\langle\omega, \frac{x}{2\|x\|_a}\Big\rangle\right| \le \sqrt{C_\kappa}|\langle\omega, x'\rangle|. \end{aligned} \tag{160}$$

This establishes (82) with $p = p_{\mathrm{md}} = 1$.

The case $\ell = \mathrm{dd}$. With $(x_1, x_2, y) = z \in \Theta_{\mathrm{dd}}$ and $x_i' = x_i/\|x_i\|_a$ we have $\|x_i'\|_a = 1$ and

$$\begin{aligned}|f_{\mathrm{dd}}(z|\omega)| &\overset{(56)}{=} 4\left|\frac{\sin\big(\langle\omega, x_1/2\rangle\big)\sin\big(\langle\omega, x_2/2\rangle\big)\cos\big(\langle\omega, y + x_2/2 - x_1/2\rangle\big)}{\sqrt{2(1-\bar{\kappa}(x_1))}\sqrt{2(1-\bar{\kappa}(x_2))}}\right| \\ &\overset{(64)}{\le} 2\alpha(\|x_1\|_a)\alpha(\|x_2\|_a)\left|\frac{\sin\big(\langle\omega, x_1/2\rangle\big)\sin\big(\langle\omega, x_2/2\rangle\big)}{\|x_1\|_a\|x_2\|_a}\right| \\ &\overset{(65)}{\le} \frac{C_\kappa}{2}\big|\langle\omega, x_1/\|x_1\|_a\rangle\langle\omega, x_2/\|x_2\|_a\rangle\big| = \frac{C_\kappa}{2}\big|\langle\omega, x_1'\rangle\langle\omega, x_2'\rangle\big| \le \frac{C_\kappa}{4}\Big(\langle\omega, x_1'\rangle^2 + \langle\omega, x_2'\rangle^2\Big).\end{aligned}$$

A.9.2 Some properties of sub-exponential random variables

Proof of Lemma 3. Denote $E := 2\mathbb{E}Y$. If $E = 0$ then $Y = X = 0$ almost surely, hence X (and Y) are both sub-exp(ν', β') for any choice of $\nu', \beta' \ge 0$ so the result is trivial. Assume now that $E > 0$. Since $|X - \mathbb{E}X| \le |X| + |\mathbb{E}X| \le Y + \mathbb{E}Y$ almost surely, we get

$$\mathbb{E}e^{\lambda(X-\mathbb{E}X)} = \sum_{q=0}^{+\infty}\frac{\lambda^q}{q!}\mathbb{E}(X-\mathbb{E}X)^q \le 1+\sum_{q=2}^{+\infty}\frac{|\lambda|^q}{q!}\mathbb{E}|X-\mathbb{E}X|^q \le 1+\sum_{q=2}^{+\infty}\frac{|\lambda|^q}{q!}\mathbb{E}(Y+\mathbb{E}Y)^q.$$

Using the binomial formula this yields

$$\begin{aligned}\mathbb{E}e^{\lambda(X-\mathbb{E}X)} &\le 1 + \sum_{q=2}^{+\infty}\frac{|\lambda|^q}{q!}\sum_{j=0}^{q}\binom{q}{j}\mathbb{E}(Y-\mathbb{E}Y)^j E^{q-j} \\ &= 1 + \sum_{j=0}^{\infty}\frac{|\lambda|^j}{j!}\mathbb{E}(Y-\mathbb{E}Y)^j \cdot \sum_{q\ge\max(j,2)}\frac{|\lambda|^{q-j}j!}{q!}\binom{q}{j}E^{q-j}.\end{aligned}$$

Observe that $\mathbb{E}(Y-\mathbb{E}Y)^0 = 1$, $\mathbb{E}(Y-\mathbb{E}Y)^1 = 0$, and

$$\sum_{q\geq \max(j,2)} \frac{|\lambda|^{q-j} j!}{q!}\binom{q}{j} E^{q-j} = \sum_{q\geq \max(j,2)} \frac{|\lambda|^{q-j}}{(q-j)!} E^{q-j} = \begin{cases} \sum_{k\geq 0} \frac{(|\lambda|E)^k}{k!} = e^{|\lambda|E}, & j \geq 2 \\ \sum_{k\geq 1} \frac{(|\lambda|E)^k}{k!} = e^{|\lambda|E} - 1 - |\lambda|E, & j = 0. \end{cases}$$

Since Y is sub-exp(ν,β), when $|\lambda| \leq 1/\beta$ we obtain (using again that $\mathbb{E}(Y-\mathbb{E}Y)^j = 0$ for $j = 1$)

$$\begin{aligned} \mathbb{E}e^{\lambda(X-\mathbb{E}X)} &\leq 1 + (e^{|\lambda|E} - 1 - |\lambda|E) + \sum_{j=2}^{\infty} \frac{|\lambda|^j}{j!}\mathbb{E}(Y-\mathbb{E}Y)^j e^{|\lambda|E} \\ &= \Big(1 + \sum_{j=2}^{\infty} \frac{|\lambda|^j}{j!}\mathbb{E}(Y-\mathbb{E}Y)^j\Big) e^{|\lambda|E} - |\lambda|E = \mathbb{E}e^{|\lambda|(Y-\mathbb{E}Y)} e^{|\lambda|E} - |\lambda|E \\ &\leq e^{\nu^2\lambda^2/2 + |\lambda|E} - |\lambda|E. \end{aligned}$$

To conclude we use a technical lemma which proof is postponed to Appendix A.9.6.

Lemma 7. *For $\alpha > 0$, we have*

$$\forall t \geq 0,\ \ e^{\alpha t^2/2 + t} - t \leq e^{(\alpha+2)t^2}. \tag{161}$$

Applying the lemma with $\alpha = (\nu/E)^2$, $t = |\lambda|E$, we obtain

$$|\lambda| \leq 1/\beta \implies \mathbb{E}e^{\lambda(X-\mathbb{E}X)} \leq e^{\frac{\nu^2\lambda^2}{2} + |\lambda|E} - |\lambda|E = e^{\frac{\alpha t^2}{2} + t} - t \leq e^{(\alpha+2)t^2} = e^{\frac{2(\nu^2+2E^2)\lambda^2}{2}}. \tag{162}$$

This shows that X is sub-exp(ν',β) with $(\nu')^2 = 2(\nu^2 + 2E^2) = 2\nu^2 + 16(\mathbb{E}Y)^2$. □

Lemma 8. *If $X \sim \mathcal{N}(0,1)$ then X^2 is sub-exp$(2,4)$ and $|X|$ is sub-exp$(4,4)$.*

Proof. Since $Y := X^2$ follows the chi-squared distribution of one degree of freedom, by [FEHP11] we have

$$\mathbb{E}e^{\lambda Y} = \frac{1}{\sqrt{1-2\lambda}},\ \ \forall|\lambda| < \frac{1}{2}. \tag{163}$$

In particular, since $\mathbb{E}Y = 1$, then

$$\mathbb{E}e^{\lambda(Y-\mathbb{E}Y)} = \mathbb{E}e^{\lambda(Y-1)} = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{\frac{2^2\lambda^2}{2}},\ \ \forall|\lambda| \leq \frac{1}{4}.$$

This is the definition of a sub-exp$(2,4)$ random variable.

Now, considering $Z := |X|$, since $|t| \leq t^2 + 1/4$ for each $t \in \mathbb{R}$, we have

$$Z = |X| \leq \frac{1}{4} + X^2.\ \text{a.s.}$$

Since X^2 is sub-exp$(2,4)$, $X^2+1/4$ is also sub-exp$(2,4)$. As $\mathbb{E}(X^2+1/4) = 5/4 \leq \sqrt{2}$, by Lemma 3, X is sub-exp$(\nu,4)$, with $\nu^2 := 2\times 2^2 + 4(5/4)^2 \leq 16$, hence Z is sub-exp$(4,4)$.

□

Lemma 9. *Consider X_i, $i = 1, 2$ two real-valued random variables (possibly non independent), assumed to be respectively sub-exp(ν_i, β_i). Then X_1+Y_2 is sub-exp(ν, β) where*

$$\nu := \nu_1 + \nu_2, \text{ and } \beta := \max\Big(\frac{\beta_1(\nu_1+\nu_2)}{\nu_1}, \frac{\beta_2(\nu_1+\nu_2)}{\nu_2}\Big). \tag{164}$$

Proof. Let $p = (\nu_1+\nu_2)/\nu_1$ and $q = (\nu_1+\nu_2)/\nu_2$, so that $1/p+1/q = 1$. By Hölder's inequality and Definition 9, if $|\lambda| \leq \min(\frac{1}{pb_1}, \frac{1}{qb_2}) = \min\Big(\frac{\nu_1}{b_1(\nu_1+\nu_2)}, \frac{\nu_2}{b_2(\nu_1+\nu_2)}\Big)$ then

$$\mathbb{E}e^{\lambda(X_1+X_2)} \leq (\mathbb{E}e^{\lambda p X_1})^{1/p}(\mathbb{E}e^{\lambda q X_2})^{1/q} \leq e^{\frac{\lambda^2 p^2 \nu_1^2}{2p}} e^{\frac{\lambda^2 q^2 \nu_2^2}{2q}} \leq e^{\frac{\lambda^2(p\nu_1^2+q\nu_2^2)}{2}} = e^{\frac{\lambda^2(\nu_1+\nu_2)^2}{2}}. \square$$

A.9.3 Proof of Theorem 5

The proof of Theorem 5 leverages Theorem 4. Before exploiting this theorem we check that the basic required assumptions are met: i) $\mathcal{T} = (\Theta, \rho, \mathcal{I})$, $\kappa \geq 0$ and $\|\cdot\|_a$ are satisfying the assumptions required in Theorem 3, ii) κ is assumed to have its mutual coherence with respect to $\mathcal{T}$ bounded by $0 < \mu < 1/10$, and k satisfies $1 \leq k < \frac{1}{10\mu}$, iii) w is a κ-compatible weight function and the average marginal density of the ω_j's satisfies (33), iv) the assumption (73) holds.

Now, we move to check the more technical assumptions: (74),(75),(76) with v defined in (86). For this purpose, we prove that $\Psi_m(\mathbf{\Omega})$ is sub-exp$(\nu'/\sqrt{m/b}, \beta'/(m/b))$ and for $\ell \in \{\mathrm{d}, \mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$, and for $z \in \Theta_\ell$, $\Psi_\ell(z|\mathbf{\Omega})$ is sub-exp$(\nu'\sqrt{m/b}, \beta'/(m/b))$, where we recall that b is the block size, and ν', β' will be specified in due time to derive (74),(75),(76) using (81).

First, consider $\ell \in \{\mathrm{d}, \mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$ and $z \in \Theta_\ell$, and observe that

$$\Psi_\ell(z|\mathbf{\Omega}) = \frac{1}{m/b}\sum_{j=1}^{m/b} \underbrace{\frac{1}{b}\sum_{i=1}^{b} \psi(\omega_{(j-1)b+i}) f_\ell(z|\omega_{(j-1)b+i})}_{=:X_j},$$

where by assumption the random variables X_j are independent and identically distributed. A well-known property of sub-exp random variables is that if $X_1, \ldots, X_n$ are independent sub-exp(ν, β) then $\frac{1}{n}\sum_{j=1}^n X_j$ is sub-exp$(\nu/\sqrt{n}, \beta/n)$. Thus, in order to prove that $\Psi_\ell(z|\mathbf{\Omega})$ is sub-exp$(\nu'/\sqrt{m/b}, \beta'/(m/b))$, it is enough to prove that the random variable

$$X := \frac{1}{b}\sum_{i=1}^{b} \psi(\omega_i) f_\ell(z|\omega_i)$$

is sub-exp(ν', β'). For this purpose, we make use of Lemmas 2 and 3.

Consider arbitrary $x_t' \in \mathbb{R}^d$, $t \in \{0, 1, 2\}$, s.t. $\|x_t'\|_a = 1$, and denote $Z_{t,p} := \frac{1}{b}\sum_{i=1}^b \psi(\omega_i)(\sqrt{C_\kappa}|\langle \omega_i, x_t'\rangle|)^p$, $t, p \in \{0, 1, 2\}$. By assumption, each $Z_{t,p}$ is sub-exp(ν, β) with $|\mathbb{E}(Z_{t,p})| \leq B$. We distinguish two cases

- if $\ell \in \{\mathrm{d}, \mathrm{mm}, \mathrm{md}\}$, since $\psi(.) \geq 0$, the first claim of Lemma 2 implies that $|X| \leq Z_{0,p}$ for some choice of x_0', hence by Lemma 3 X is sub-exp(ν', β), where

$$\nu' := \sqrt{2\nu^2 + 16B^2}. \tag{165}$$

- if $\ell = \mathrm{dd}$ by the second claim of Lemma 2 we similarly get the existence of $x'_t \in \mathbb{R}^d$ satisfying $\|x'_t\|_a = 1$, $t = 1, 2$, such that $|X| \leq Z'$ with $Z' := (Z_{1,2} + Z_{2,2})/4$. By Lemma 9, $Z_{1,2} + Z_{2,2}$ is sub-exp$(2\nu, 2\beta)$, and $|\mathbb{E}(Z_{1,2} + Z_{2,2})| \leq 2B$, hence $Z' := (Z_{1,2} + Z_{2,2})/4$ is also sub-exp(ν, β) with $|\mathbb{E}(Z')| \leq B$. By Lemma 3 we also get that X is sub-exp(ν', β) with ν' as in (165).

Now, consider $\ell = \mathrm{m}$. Similarly, to prove that $\Psi_m(\mathbf{\Omega})$ is sub-exp$(\nu'/\sqrt{m/b}, \beta'/(m/b))$, it is enough to prove that $X := \frac{1}{b}\sum_{i=1}^b \psi(\omega_i)$ is sub-exp(ν', β'). Since $\psi(\omega) = Z_{0,0}$ for any choice of x'_0, this is indeed true with ν' as in (165) and $\beta' = \beta$.

Now, we use (81) applied to $\Psi_{\mathrm{m}}(\mathbf{\Omega})$ with $t = \frac{\tau}{4}$, and to $\Psi_{\mathrm{d}}(z|\mathbf{\Omega})$ with $t = \frac{\tau}{8}$, and to $\Psi_\ell(z|\mathbf{\Omega})$ with $t = \frac{\tau}{16k}$ for $\ell \in \{\mathrm{d}, \mathrm{mm}, \mathrm{md}, \mathrm{dd}\}$ and $z \in \Theta_\ell$, to get that the left hand side of (74),(75),(76) is bounded from above by

$$\max_{t \in \{\frac{\tau}{4}, \frac{\tau}{8}, \frac{\tau}{16k}\}} 2\exp\left(-\frac{(m/b)t^2}{2\nu'^2 + \beta t}\right) = 2\exp\left(-\frac{(m/b)\left(\frac{\tau}{16k}\right)^2}{2\nu'^2 + \beta\left(\frac{\tau}{16k}\right)}\right) = 2\exp\left(-\frac{m\tau^2}{256bk^2(2\nu'^2 + \beta\left(\frac{\tau}{16k}\right)}\right)$$

where we used that $t \mapsto t^2/(2(\nu')^2 + \beta t)$ is an increasing function. We conclude by observing that, with v as defined in (86), we have

$$256bk^2(2\nu'^2 + \beta\frac{\tau}{16k}) \leq 256bk^2(2\nu'^2 + \beta\tau) = \tau^2 v,$$

To prove the variant of the theorem, we first reason as above to show that the modified assumptions imply that for arbitrary $x'_t \in \mathbb{R}^d$, $t \in \{0, 1, 2\}$, s.t. $\|x'_t\|_a = 1$ the random variable $Y_p := \frac{1}{b}\sum_{i=1}^b B_\psi(\sqrt{C_\kappa}|\langle\omega, x'_0\rangle|)^p$ is sub-exp$(B_\psi\nu, B_\psi\beta)$ with $|\mathbb{E}Y_p| \leq B_\psi\beta$, and similarly for $Y' := \frac{1}{b}\sum_{i=1}^b B_\psi C_\kappa(\langle\omega, x'_1\rangle^2 + \langle\omega, x'_2\rangle^2)/4$. Then, using the definition of B_ψ we obtain that: for $\ell = \mathrm{m}$, $X := \frac{1}{b}\sum_{i=1}^b \psi(\omega_i)$ satisfies $|X| \leq Y_0$; for $\ell \in \{\mathrm{d}, \mathrm{mm}, \mathrm{md}\}$ and $z \in \Theta_\ell$, $X := \frac{1}{b}\sum_{i=1}^b \psi(\omega_i) f_\ell(z|\omega_i)$ satisfies $|X| \leq Y_p$ for an appropriate choice of x'_0, $p \in \{0, 1, 2\}$; for $\ell = \mathrm{dd}$ and $z \in \Theta_\ell$, $X := \frac{1}{b}\sum_{i=1}^b \psi(\omega_i) f_\ell(z|\omega_i)$ satisfies $|X| \leq Y'$ for an appropriate choice of x'_1, x'_2. The same reasoning as above yields that X is sub-exp$(B_\psi\nu', B_\psi\beta)$ with ν' as in (165). We conclude similarly once we observe that $256bk^2(2B_\psi^2\nu'^2 + B_\psi\beta\tau) = \tau^2 v'$.

A.9.4 Proofs of Corollary 1 and Corollary 2

Corollary 1 and Corollary 2 are direct consequences of Theorem 5. To see why, we check that the assumptions of these theorems are met in these settings.

Checking that the assumptions on $\mathcal{T}$ and κ hold, and controlling C_κ, μ. First, observe that, in the setting of Gaussian (resp. Dirac) mixtures of Example 2 (resp. of Example 1), the kernel satisfies $\kappa \geq 0$, and (63) holds with $\|.\|_a = \|.\|_{\mathbf{\Sigma}}$ (resp. with $\|\cdot\|_a = \|\cdot\|_2$) and $\tilde{\kappa}(r) = e^{-r^2/\sigma^2}$ with $\sigma := \sqrt{2(2 + s^2)}$ (resp. with $\sigma := \sqrt{2}s$). In both settings we have $\varrho = \|\cdot\|_a/\epsilon$, and by the definition of Θ_{d} (see (59), with $\|\cdot\| := \|\cdot\|_a/\epsilon$), we have $R := \sup_{x \in \Theta_{\mathrm{d}}} \|x\|_a \leq \epsilon$. Thus, by Lemma 10, the constant C_κ from (65) satisfies $C_\kappa \leq \max\left(1, \sqrt[4]{3}R, \sqrt[4]{3}\sigma\right)^2 \leq \max\left(1, \sqrt{3}\epsilon^2, \sqrt{3}\sigma^2\right)$. Now we proceed separately for the two settings.

- For Gaussian mixtures, as $\sqrt{2 + s^2} \leq \epsilon(4\sqrt{\log(ek)})^{-1}$, we get $\sigma = \sqrt{2(2 + s^2)} \leq \epsilon$, and by [GBKT21b, Theorem 5.16, Lemma 6.10] κ is locally characteristic with mutual coherence with respect to $\mathcal{T}$ bounded by $\mu \leq 12/(16(10k - 1))$.

- For mixtures of Diracs, since $s \leq \epsilon(4\sqrt{\log(5ek)})^{-1}$, we get $\sigma := \sqrt{2}s \leq \epsilon$ and the bound on the coherence holds by [GBKT21b, Theorem 5.16, Lemma 6.10].

In both cases, we get $\mu \leq 12/(16(10k-1))m \leq 1/(10k)$ and $\sigma \leq \epsilon$. The latter implies

$$C_\kappa \leq \max\left(1, \sqrt{3}\epsilon^2, \sqrt{3}\sigma^2\right) = \max(1, \sqrt{3}\epsilon^2). \tag{166}$$

Since $B_\psi := \sup_{\omega\in\mathbb{R}^d} \psi(\omega)$ is finite in both settings, we may use the variant of Theorem 5. Indeed, by (91) (resp. by (96)) $B_\psi = (1+2s^{-2})^{d/2}$ for Gaussian mixtures (resp. $B_\psi = 1$ for mixtures of Diracs). Now, we check that the assumptions expressed in Item 1 (in its variant involving Z'_p) and Item 2 of Theorem 5 hold in both settings.

Checking the variant of Item 1 in Theorem 5 We show the existence of $\nu, \beta, B > 0$ such that the random variables Z'_p defined in (90) are sub-exp(ν, β) with $|\mathbb{E}(Z_p)| \leq B$. Since the frequencies $\omega_1, \ldots, \omega_m$ are i.i.d., we consider a block size $b = 1$ and $Z'_p = (\sqrt{C_\kappa}|\langle\omega, x\rangle|)^p$ with $\omega \sim \Lambda$. We will indeed prove that for each $x \in \mathbb{R}^d$ s.t. $\|x\|_a = 1$, the random variables $|\sqrt{C_\kappa}\langle\omega, x\rangle|^p$, $p \in \{0, 1, 2\}$ are sub-exp(ν, β) with $|\mathbb{E}Z_p| \leq B$, where

$$B := \max(1, C_\kappa s^{-2}), \quad \textit{and} \quad \nu = \beta = 4B. \tag{167}$$

This is done in the following. To handle both settings in a common framework, define $\mathbf{\Sigma} = \mathbb{I}_d$ for the setting of a mixture of Diracs, so that in both cases we have $\|.\|_a = \|.\|_{\mathbf{\Sigma}}$. Thus, a vector $x \in \mathbb{R}^d$ satisfies $\|x\|_a = \|x\|_{\mathbf{\Sigma}} = 1$ if, and only if, $\|\mathbf{\Sigma}^{-1/2}x\|_2 = 1$. Since $\omega \sim \mathcal{N}(0, s^{-2}\mathbf{\Sigma}^{-1})$ we have $s\mathbf{\Sigma}^{1/2}\omega \sim \mathcal{N}(0, \mathbb{I}_d)$ hence $s\langle\omega, x\rangle = s\langle\mathbf{\Sigma}^{1/2}\omega, \mathbf{\Sigma}^{-1/2}x\rangle \sim \mathcal{N}(0, 1)$, so that $(\mathbb{E}|\langle\omega, x\rangle|)^2 \leq \mathbb{E}|\langle\omega, x\rangle|^2 = 1/s^2$, hence using also that $|\langle\omega, x\rangle|^0 \equiv 1$ we obtain

$$\max_{p\in\{0,1,2\}} \mathbb{E}|\sqrt{C_\kappa}\langle\omega, x\rangle|^p \leq \max(1, \sqrt{C_\kappa}s^{-1}, C_\kappa s^{-2}) = B.$$

By Lemma 8, $s|\langle\omega, x\rangle|$ is sub-exp$(4, 4)$ and $s^2\langle\omega, x\rangle^2$ is sub-exp$(2, 4)$, hence $|\sqrt{C_\kappa}\langle\omega, x\rangle|$ is sub-exp$(4\sqrt{C_\kappa}/s, 4\sqrt{C_\kappa}/s)$, and $C_\kappa\langle\omega, x\rangle^2$ is sub-exp$(2C_\kappa/s^2, 4C_\kappa/s^2)$. Observe that $\max(4\sqrt{C_\kappa}/s, 2C_\kappa/s^2) \leq 4B$ and $\max(4\sqrt{C_\kappa}/s, 4C_\kappa/s^2) \leq 4B$ to conclude that $|\sqrt{C_\kappa}\langle\omega, x\rangle|$ and $C_\kappa\langle\omega, x\rangle^2$ are indeed both sub-exp(ν, b) with $\nu = b = 4B$. The same holds for $p = 0$ since $|\sqrt{C_\kappa}\langle\omega, x\rangle|^0 \equiv 1$.

NB: in the setting of Gaussian mixtures, for $x \in \mathbb{R}^d$ such that $\|x\|_{\mathbf{\Sigma}} = 1$, $\psi(\omega)$ and $\psi(\omega)|\langle\omega, x\rangle|$ and $\psi(\omega)\langle\omega, x\rangle^2$ are bounded and we may alternatively have used Hoeffding's inequality [Hoe94] instead of Theorem 5. We chose to use the latter as it allows to encompass both settings under the same reasoning.

We move now to check that Item 2 holds in both settings.

Identifying M such that Item 2 in Theorem 5 holds for Gaussian mixtures with any $v > 0$. We show that

$$|\psi(\omega)f_0(\omega)| \leq M := 4B_\psi, \ \forall\omega, \tag{168}$$

hence (85) holds for any $v > 0$. In particular, (85) holds for $v = v_k(\tau)$, with $v_k(\tau)$ defined in (94). Indeed, for every $t \geq 0$ we have $t^2 \leq (t + t^3)/2$ (since $t(t-1)^2 \geq 0$),

hence given the definition (67) of f_0 and since $\|\cdot\|_{a,\star} = \|\cdot\|_{\mathbf{\Sigma}^{-1}}$ we have $f_0(\omega) \leq \frac{3}{2}(\|\omega\|_{\mathbf{\Sigma}^{-1}} + \|\omega\|^3_{\mathbf{\Sigma}^{-1}})$ for every ω, so that

$$|\psi(\omega) f_0(\omega)| \overset{(91)}{=} B_\psi e^{-\omega^{\mathrm{T}}\mathbf{\Sigma}\omega/2} f_0(\omega) \leq \frac{3}{2} B_\psi e^{-\|\omega\|^2_{\mathbf{\Sigma}^{-1}}/2}(\|\omega\|_{\mathbf{\Sigma}^{-1}} + \|\omega\|^3_{\mathbf{\Sigma}^{-1}}) \leq \frac{3}{2} B_\psi \varphi(\|\omega\|_{\mathbf{\Sigma}^{-1}}) \leq 4B_\psi$$

since $\varphi(t) := (t + t^3)e^{-t^2/2}$ satisfies $\varphi(t) \leq 8/3$ for $t \in \mathbb{R}$. Indeed, we have

$$\sup_{t\in\mathbb{R}} \varphi(t) \leq \sup_{t\in\mathbb{R}} \varphi_1(t) + \sup_{t\in\mathbb{R}} \varphi_2(t),$$

where $\varphi_1(t) := te^{-t^2/2}$, and $\varphi_2(t) := t^3 e^{-t^2/2}$, and it is easy to prove that $\sup_{t\in\mathbb{R}} \varphi_1(t) = \varphi_1(1) = e^{-1/2} \leq 1$, and $\sup_{t\in\mathbb{R}} \varphi_2(t) = \varphi_2(\sqrt{3}) = 3\sqrt{3}e^{-3/2} \approx 1.16 \leq 5/3 \approx 1.66$.

Identifying M such that Item 2 holds for mixtures of Diracs with $v = v_k(\tau)$. Since $\psi(\omega) = 1$ (cf (96)) we have $\Psi_0(\mathbf{\Omega}) = \frac{1}{m}\sum_{j=1}^m f_0(\omega_j)$, with $f_0(\omega)$ defined in (67). Since $\|\cdot\|_a = \|\cdot\|_2$ we have $\|\cdot\|_{a,\star} = \|\cdot\|_2$ hence $f_0(\omega) = \sum_{t=1}^3 \|\omega\|_2^t$. As $\omega \sim \mathcal{N}(0, s^{-2}\mathbb{I}_d)$, Proposition 11 yields

$$\forall t \geq 0, \quad \mathbb{P}\Big(\Psi_0(\mathbf{\Omega}) > \tilde{M}(t)\Big) \leq \exp\left(-\frac{(mt)^{2/3}}{2}\right),$$

with $\tilde{M}(t) := 1 + \frac{8}{s^3}\big(\sqrt{\frac{8}{\pi}}d^{3/2} + t\big)$. We define

$$C_0 := 7B_\psi B \tag{169}$$

and recall that $C_0 \geq 7$ since $B_\psi \geq 1$ and $B \geq 1$. Therefore, for $0 < \tau \leq 1$, and $v_k(\tau)$ as defined in (99), we have

$$\mathbb{P}\Big(\Psi_0(\mathbf{\Omega}) > \tilde{M}(\tau^{3/2}m^{1/2})\Big) \leq \exp\left(-\frac{m\tau}{2}\right) \leq 2\exp(-m/v_k(\tau)),$$

since $2/\tau \leq 2/\tau^2 \leq 2(C_0/\tau)^2 \leq v_k(\tau)$.

In other words, (85) holds with $M = \tilde{M}(\tau^{3/2}m^{1/2})$ and $v = v_k(\tau)$.

Wrapping up the proof.

To complete the proof of Corollary 1 and Corollary 2 we use Theorem 5 and the last step is to give explicit upper bounds of the constants C and v' respectively defined in (88) and (89).

We start with v', and we show that it is upper bounded by $v_k(\tau)$ defined in (94) (resp. in (99)). By (167), $\nu = \beta = 4B$ hence, by (86) $\nu' = \sqrt{2}\sqrt{\nu^2 + 8B^2} = \sqrt{48B^2} \leq \sqrt{49B^2} = 7B$. Hence, $B_\psi^2\nu'^2 \leq (7B_\psi B)^2 = C_0^2$ and $B_\psi\beta = 4B_\psi B = \frac{4}{7}C_0$ where C_0 is defined in (169). Since the block size is $b = 1$, by (89) we obtain

$$v' = \frac{256k^2 b(2B_\psi^2\nu'^2 + B_\psi\beta\tau)}{\tau^2} \leq 512k^2\left((C_0/\tau)^2 + \frac{2}{7}(C_0/\tau)\right) \leq 512k^2\left((C_0/\tau)^2 + \frac{1}{3}(C_0/\tau)\right).$$

This matches the expressions of $v_k(\tau)$ used in (94) and (99). Soon we will also prove that C_0 satisfies the bounds expressed in (94) and (99).

As for C, observe that by definition (88), $C = 6144MC_\kappa \cdot k(1 + \mathrm{diam}_a(\Theta))$, so that we only need to give an upper bound of $6144MC_\kappa$ to get the expressions that appear in (95) and (100).

To conclude we bound C_0 and $6144MC_\kappa$. First, by (166), we have $C_\kappa \leq \max(1, \sqrt{3}\epsilon^2)$, and by (167) $B = \max(1, C_\kappa s^{-2})$. We study separately the two settings:

- For mixtures of Gaussians, by (92) we have $\epsilon \geq \max(s, 1)$ hence $C_\kappa \leq \sqrt{3}\epsilon^2$, $B \leq \sqrt{3}(\epsilon/s)^2$. Since $B_\psi = (1 + 2s^{-2})^{d/2}$, we get by (169) $C_0 = 7BB_\psi \leq 7\sqrt{3}\epsilon^2 s^{-2}(1 + 2s^{-2})^{d/2}$ as claimed in (94). Since $M = 4B_\psi$ by (168), we get $6144MC_\kappa \leq (4 \times 6144\sqrt{3})\epsilon^2 B_\psi \leq 43000\epsilon^2 B_\psi$ as claimed in (95).
- For mixtures of Diracs, by (97) we have $\epsilon \geq s$ hence $\sqrt{3}\epsilon^2 s^{-2} \geq 1$. As $C_\kappa s^{-2} \leq \max(s^{-2}, \sqrt{3}\epsilon^2 s^{-2})$, it follows that $B = \max(1, C_\kappa s^{-2}) \leq \max(1, s^{-2}, \sqrt{3}\epsilon^2 s^{-2}) = \max(s^{-2}, \sqrt{3}\epsilon^2 s^{-2}) = s^{-2}\max(1, \sqrt{3}\epsilon^2)$. Since $B_\psi = 1$ it follows by (169) that $C_0 = 7BB_\psi = 7B \leq 7s^{-2}\max(1, \sqrt{3}\epsilon^2)$ as claimed in (99). Finally observe that $c := \sqrt{8/\pi} \approx 1.6$ so that $cd^{3/2} + t \geq 1$ for every $t > 0$ hence $\tilde{M}(t) = 1 + \frac{8}{s^3}(cd^{3/2} + t) \leq (1 + 8/s^3)(cd^{3/2} + t) \leq (1 + 2/s)^3(cd^{3/2} + t) \leq (1 + 2/s)^3(2d^{3/2} + t)$. Since $M = \tilde{M}(m^{1/2}\tau^{3/2})$ it follows that $6144MC_\kappa \leq 6144(1 + 2/s)^3 \max(1, \sqrt{3}\epsilon^2)(2d^{3/2} + \sqrt{m}\tau^{3/2})$ as claimed in (100).

A.9.5 Some helpful results

Lemma 10. *Consider $\sigma > 0$ and for $r \geq 0$ define $\alpha_\sigma(r) := \frac{r}{\sqrt{1 - e^{-r^2/\sigma^2}}}$. For each $R > 0$ we have*

$$\sup_{r \in [0,R]} |\alpha_\sigma(r)| \leq \sqrt[4]{3}\max(\sigma, R), \quad \text{and} \quad \sup_{r \in (0,R]} |\alpha'_\sigma(r)| \leq 1. \tag{170}$$

Proof. First we show that $\alpha_\sigma(r) = \sigma\alpha_1(r/\sigma) \leq (1 - e^{-1})^{-1/2}\max(\sigma, r)$ when $r \geq 0$. This implies the first bound as $(1 - e^{-1})^{-1/2} \approx 1.26 \leq 1.316 \approx \sqrt[4]{3}$. When $\sigma \leq r$ we have

$$|\alpha_\sigma(r)| = \frac{r}{\sqrt{1 - e^{-r^2/\sigma^2}}} \leq \frac{r}{\sqrt{1 - e^{-1}}} \leq (1 - e^{-1})^{-1/2} r = (1 - e^{-1})^{-1/2}\max(\sigma, r).$$

When $0 \leq r \leq \sigma$, we prove below that $|\alpha_1(t)| \leq (1 - e^{-1})^{-1/2}$ for every $t \in [0, 1]$, so that

$$|\alpha_\sigma(r)| = \sigma|\alpha_1(r/\sigma)| \leq (1 - e^{-1})^{-1/2}\sigma = (1 - e^{-1})^{-1/2}\max(\sigma, r).$$

To show that $|\alpha_1(t)| \leq (1 - e^{-1})^{-1/2}$ on $[0, 1]$ observe that since $u \mapsto e^{-u}$ is convex, it has non-decreasing slopes so that $(e^{-u} - 1)/u \leq (e^{-1} - 1)/1$, $\forall u \in [0, 1]$. This reads $u/(1 - e^{-u}) \leq 1/(1 - e^{-1})$ and implies $|\alpha_1(t)|^2 = t^2/(1 - e^{-t^2}) \leq 1/(1 - e^{-1})$ for $t \in [0, 1]$, thus $|\alpha_1(t)| \leq 1/\sqrt{1 - e^{-1}}$.

We now prove that $|\alpha'_1(t)| \leq 1$, $\forall t > 0$. This implies the second bound since $\alpha'_\sigma(r) = \alpha'_1(r/\sigma)$. Writing $\alpha_1(t) = t[v(t)]^{-1/2}$ with $v(t) := 1 - e^{-t^2}$ we get

$$\alpha'_1(t) = [v(t)]^{-1/2} - \frac{t}{2}v'(t)[v(t)]^{-3/2} = [v(t)]^{-3/2}\left(v(t) - tv'(t)/2\right) = [v(t)]^{-3/2}\left(1 - e^{-t^2}(1 + t^2)\right)$$

For each $t \geq 0$, since $e^{t^2} \geq 1 + t^2$, it follows that $|\alpha'_1(t)| = \alpha'_1(t)$. When $t \geq 1$, since $0 < x := e^{-t^2} \leq 1/e \approx 0.368 < 1/2 < 0.618 \approx (\sqrt{5} - 1)/2$ we have

$(x-1)^3+(1-2x)^2 = x^3+x^2-x = x(x^2+x-1) = x[x+(1+\sqrt{5})/2][x+(1-\sqrt{5})/2] \leq 0$ hence $(1-2x)^2 \leq (1-x)^3$, and since $1-2x>0$ it follows that $1-2x \leq (1-x)^{3/2}$. Thus, we obtain

$$|\alpha_1'(t)| = \frac{1-e^{-t^2}(1+t^2)}{\sqrt{1-e^{-t^2}}^3} \leq \frac{1-2e^{-t^2}}{\sqrt{1-e^{-t^2}}^3} = \frac{1-2x}{(1-x)^{3/2}} \leq 1,$$

Now, when $0 < t \leq 1$, since $[v(t)]^{-3/2} = (\alpha_1(t)/t)^3$, using that $|\alpha_1(t)| \leq (1-e^{-1})^{-1/2}\max(1,t) = (1-e^{-1})^{-1/2}$ and $(1-e^{-1})^{-3/2} \approx 1.99 \leq 2$ we get

$$|\alpha_1'(t)| \leq (1-e^{-1})^{-3/2}\frac{1-e^{-t^2}(1+t^2)}{t^3} \leq 2\frac{1-e^{-t^2}(1+t^2)}{t^3}.$$

It is enough to show that $g(t) := (1-e^{-t^2}(1+t^2))/t^3 \leq 1/2$, $\forall t \in (0,1]$. We have $g'(t) = t^{-4}(-3+e^{-t^2}(3+3t^2+2t^4))$ hence $\texttt{sign}(g'(t)) = -\texttt{sign}(e^{t^2}-(1+t^2+\frac{2}{3}t^4))$. Thus there is a neighborhood of zero in which g is increasing, since $\texttt{sign}(g'(t)) = -\texttt{sign}(1+t^2+\frac{1}{2}t^4+O(t^6)-(1+t^2+\frac{2}{3}t^4)) = -\texttt{sign}(-\frac{1}{6}t^4+O(t^6)) = +1$ for t small enough. Since g is continuously differentiable, its supremum on $(0,1]$ is thus either equal to $g(1) = 1-2/e \approx 0.264 < 1/2$ or to $g(t_*)$, for some local maximum $0 < t_* \leq 1$ which must satisfy $g'(t_*) = 0$. To conclude without further characterizing the existence or value of such a root, we establish that any such root must satisfy $g(t_*) \leq 1/2$. Indeed using that $g'(t_*) = 0$ if, and only if $e^{t_*^2} = 1+t_*^2+\frac{2}{3}t_*^4$, we obtain

$$g(t_*) = \frac{1-e^{-t_*^2}(1+t_*^2)}{t_*^3} = \frac{e^{t_*^2}-(1+t_*^2)}{e^{t_*^2}t_*^3} = \frac{1+t_*^2+\frac{2}{3}t_*^4-(1+t_*^2)}{(1+t_*^2+\frac{2}{3}t_*^4)t_*^3} = \frac{\frac{2}{3}t_*}{1+t_*^2+\frac{2}{3}t_*^4}.$$

We finally distinguish two cases: i) if $t_* < 1/\sqrt{2}$ then $g(t_*) \leq \frac{2}{3}t_* \leq \frac{\sqrt{2}}{3} \leq 1/2$; ii) if $1/\sqrt{2} \leq t_* \leq 1$ then $g(t_*) \leq \frac{\frac{2}{3}}{1+\frac{1}{2}+\frac{2}{3}(1/2)^2} = 0.4 < 1/2$. □

Proposition 11. *Consider $s>0$, and let $\omega_1,\dots,\omega_m$ be i.i.d. samples from $\mathcal{N}(0,\frac{1}{s^2}\mathbb{I}_d)$. Then*

$$\forall \tau \geq 0,\ \mathbb{P}\Big(\frac{1}{m}\sum_{j=1}^m \|\omega_j\|^3 > \frac{4}{s^3}\big(d^{3/2}\sqrt{8/\pi}+\tau\big)\Big) \leq \exp\Big(-\frac{(m\tau)^{2/3}}{2}\Big). \tag{171}$$

Moreover, as a consequence we have for $\tau > 0$

$$\mathbb{P}\Big(\frac{1}{m}\sum_{j=1}^m\sum_{t=1}^3 \|\omega_j\|_2^t > 1+\frac{8}{s^3}\big(d^{3/2}\sqrt{8/\pi}+\tau\big)\Big) \leq \exp\Big(-\frac{(m\tau)^{2/3}}{2}\Big). \tag{172}$$

Proof. First, to prove (171) when $\omega_1,\dots,\omega_m$ are i.i.d. samples from $\mathcal{N}(0,\frac{1}{s^2}\mathbb{I}_d)$, it is enough to deal with the case $s=1$. Next, for $s=1$, denoting $\Omega \in \mathbb{R}^{dm}$ the concatenation of $\omega_1,\dots,\omega_m \in \mathbb{R}^d$, the vector Ω has independent standard normal random entries when $\omega_1,\dots,\omega_m$ are i.i.d. samples from $\mathcal{N}(0,\mathbb{I}_d)$, and for any $2 \leq p < \infty$ the function $f_p : \mathbb{R}^{dm} \to \mathbb{R}$ defined by $f_p(\Omega) := (\sum_{j=1}^m \|\omega_j\|_2^p)^{1/p}$ is (as we will soon show) 1-Lipschitz with respect to the Euclidean norm in $\mathbb{R}^{dm}$. Therefore, we may use the Tsirelson-Ibragimov-Sudakov inequality [TIS76], a.k.a. concentration of a random variable that writes as a Lipschitz function of a Gaussian vector [BLM13a,

Theorem 5.6], to obtain $\mathbb{P}(f_p(\Omega) - \mathbb{E}f_p(\Omega) \geq t) \leq \exp(-t^2/2)$, for each $t \geq 0$, or equivalently

$$\forall t \geq 0,\ \mathbb{P}\Big(\big[f_p(\Omega)\big]^p \geq \big[\mathbb{E}f_p(\Omega) + \tau\big]^p\Big) \leq \exp\big(-t^2/2\big).$$

By convexity we have $(a+b)^p = 2^p(a/2+b/2)^p \leq 2^{p-1}(a^p+b^p)$ for every $a, b \in \mathbb{R}_+$, hence

$$\forall t \geq 0,\ \mathbb{P}\Big(\big[f_p(\Omega)\big]^p \geq 2^{p-1}\big(\big[\mathbb{E}f_p(\Omega)\big]^p + t^p\big)\Big) \leq \exp\big(-t^2/2\big).$$

We now show that $\big[\mathbb{E}f_p(\Omega)\big]^p \leq md^{p/2}\mathbb{E}|g|^p$. The convexity of $t \mapsto t^{p/2}$ ($p \geq 2$) on $\mathbb{R}_+$ yields

$$\mathbb{E}\|\omega\|_2^p = d^{p/2}\mathbb{E}\Big(\frac{1}{d}\sum_{i=1}^d \omega_i^2\Big)^{p/2} \leq d^{p/2}\mathbb{E}\left[\frac{1}{d}\sum_{i=1}^d |\omega_i|^p\right] = d^{p/2}\mathbb{E}|g|^p$$

where $\omega \sim \mathcal{N}(0, \mathbb{I}_d)$, $g \sim \mathcal{N}(0,1)$. B convexity of $t \mapsto t^p$ and Jensen's inequality, it follows that

$$\big[\mathbb{E}f_p(\Omega)\big]^p \leq \mathbb{E}\big[f_p(\Omega)\big]^p = \sum_{j=1}^m \mathbb{E}\|\omega_j\|_2^p = m\mathbb{E}\|\omega\|_2^p \leq md^{p/2}\mathbb{E}|g|^p.$$

As a result

$$\forall t \geq 0,\ \mathbb{P}(\big[f_p(\Omega)\big]^p \geq 2^{p-1}(md^{p/2}\mathbb{E}|g|^p + t^p)) \leq \exp(-t^2/2),$$

or equivalently $\mathbb{P}\Big(\frac{1}{m}\sum_{j=1}^m \|\omega_j\|^p \geq 2^{p-1}\big(d^{p/2}\mathbb{E}|g|^p + \tau\big)\Big) \leq \exp(-(m\tau)^{2/p}/2)$ for each $\tau \geq 0$. Since $\mathbb{E}|g|^3 = \sqrt{8/\pi}$ [FEHP11, Chapter 11], considering $p=3$ yields (171) (for $s=1$) as claimed. Now, observe that for $t \geq 0$ we have $t^3 - t - (t^2-1) = (t^2-1)(t-1) = (t-1)^2(t+1) \geq 0$ hence $t^3+1 \geq t^2+t$ and $t + t^2 + t^3 \leq 1 + 2t^3$. Thus $\sum_{i=1}^3 \|\omega_j\|_2^i \leq 1 + 2\|\omega_j\|_2^3$, and we deduce (172) from (171).

To complete the proof, we show that f_p is 1-Lipschitz with respect to $\|.\|_2$. Denoting $v_\Omega := (\|\omega_j\|_2)_{j\in[m]} \in \mathbb{R}^m$, observe that $f_p(\Omega) = \|v_\Omega\|_p$. Thus, for $\Omega, \Omega' \in \mathbb{R}^{dm}$, since $p \geq 2$, we have

$$\big|f_p(\Omega) - f_p(\Omega')\big| = \big|\|v_\Omega\|_p - \|v_{\Omega'}\|_p\big| \leq \big\|v_\Omega - v_{\Omega'}\big\|_p \leq \big\|v_\Omega - v_{\Omega'}\big\|_2.$$

Finally,

$$\big\|v_\Omega - v_{\Omega'}\big\|_2^2 = \sum_{j=1}^m (\|\omega_j\|_2 - \|\omega_j'\|_2)^2 \leq \sum_{j=1}^m \|\omega_j - \omega_j'\|_2^2 = \|\Omega - \Omega'\|_2^2. \qquad \square$$

A.9.6 Proof of Lemma 7

Denote $c := 2(\alpha+2)$ and $\varphi(t) := e^{-(c-\alpha)t^2/2+t} - te^{-ct^2/2}$. Since $\varphi(0) = 1$, it is enough to prove that φ is non-increasing on $\mathbb{R}_+$. Since φ is $\mathcal{C}^1$, we study the sign of $\varphi'(t) = \Big(-(c-\alpha)t + 1 + (ct^2-1)e^{-\alpha t^2/2 - t}\Big)e^{-(c-\alpha)t^2/2+t}$ which is the sign of $\psi(t) := 1 - (c-\alpha)t + (ct^2-1)e^{-\alpha t^2/2-t}$. To show that $\psi(t) \leq 0$ for each $t \in \mathbb{R}_+$ we study its sign on the intervals $(0, \frac{1}{c-\alpha})$ and $(\frac{1}{c-\alpha}, +\infty)$. As a preliminary we record that since $\alpha > 0$, we have

$$\sqrt{c}/(c-\alpha) = \sqrt{2(2+\alpha)}/(\alpha+4) \leq 1/2. \tag{173}$$

Case of $t \in (0, \frac{1}{c-\alpha})$. Since $1-(c-\alpha)t > 0$, we will get that $\psi(t) \leq 0$ if we show that

$$\frac{1-ct^2}{1-(c-\alpha)t} \geq e^{\alpha t^2/2+t}. \tag{174}$$

Since $t \in (0, 1/(c-\alpha))$, using (173) we have $\sqrt{c}t < (c-\alpha)t < 1$ hence

$$\frac{1-ct^2}{1-(c-\alpha)t} = \frac{(1-\sqrt{c}t)(1+\sqrt{c}t)}{1-(c-\alpha)t} \geq 1+\sqrt{c}t. \tag{175}$$

Denoting $h(u) := (1+\sqrt{c}u)e^{-\alpha u^2/2-u}$, it is enough to prove that $h(t) \geq 1 = h(1)$, which will follow if we establish that $h'(u) = \Big(\sqrt{c}(1-u-\alpha u^2)-(\alpha u+1)\Big)e^{-\alpha u^2/2-u} \geq 0$ on $(0, 1/(c-\alpha))$, or equivalently that the quadratic function $\sqrt{c}(1-u-\alpha u^2)-(\alpha u+1)$ takes non-negative values at $u=0$ and at $u = 1/(c-\alpha)$. Indeed, its evaluation at $u=0$ yields $\sqrt{c}-1 = \sqrt{2(2+\alpha)}-1 > 0$, while its evaluation on $1/(c-\alpha) = 1/(\alpha+4)$ is lower bounded by $8/(\alpha+4)^2 > 0$.

Case of $t \in (\frac{1}{c-\alpha}, +\infty)$. Since $1-(c-\alpha)t < 0$, we get that $\psi(t) \leq 0$ as soon as

$$\frac{ct^2-1}{(c-\alpha)t-1} \leq e^{\alpha t^2/2+t}. \tag{176}$$

Since $t > 1/(c-\alpha)$, we have $(c-\alpha)t > 1$, and using (173) we get

$$\frac{\sqrt{c}t-1}{(c-\alpha)t-1} \leq \frac{\sqrt{c}t}{(c-\alpha)t} \leq \frac{1}{2}.$$

Therefore

$$\frac{ct^2-1}{(c-\alpha)t-1} = \frac{(\sqrt{c}t-1)(\sqrt{c}t+1)}{(c-\alpha)t-1} \leq \frac{1}{2}+\frac{1}{2}\sqrt{c}t \leq \frac{e^{\alpha t^2/2+t}}{2}+\frac{\sqrt{c}t}{2}. \tag{177}$$

Denoting $g(u) := ue^{-\alpha u^2/2-u}$, it is thus enough to show that $\sqrt{c}g(t) \leq 1$ to conclude. Since $g'(u) = -(-1+u+\alpha u^2)e^{-\alpha u^2/2-u}$, the unique $u \geq 0$ such that $g'(u) = 0$ is $u_\alpha := 2/(\sqrt{4\alpha+1}+1)$, which satisfies $\alpha u_\alpha^2 + u_\alpha - 1 = 0$, and the maximum of $g(u)$ on $\mathbb{R}_+$ is at $u = u_\alpha$. As a result

$$\sqrt{c}g(t) \leq \sqrt{c}g(u_\alpha) = 2\frac{\sqrt{2(2+\alpha)}}{\sqrt{4\alpha+1}+1}e^{-1/2-u_\alpha/2}. \tag{178}$$

To conclude, we show that the r.h.s. is bounded by one, by distinguishing two cases. On the one hand, if $\alpha \geq 2$, we have $2\sqrt{2(2+\alpha)}/(\sqrt{4\alpha+1}+1) \leq \sqrt{2}$, and since $u_\alpha \geq 0$ the r.h.s. of (178) is upper bounded by $e^{-1/2}\sqrt{2} \leq 1$. On the other hand, if $\alpha \leq 2$, we have $u_\alpha/2 \geq 1/4$ and $2\sqrt{2(2+\alpha)}/(\sqrt{4\alpha+1}+1) \leq 2$ (the latter inequality holds for any $\alpha \geq 0$), so that the r.h.s. of (178) is upper bounded by $2e^{-3/4} \leq 1$. In both cases, we get as claimed that $\sqrt{c}g_\alpha(t) \leq 1$.

References

[AC06] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563, 2006.

[Ach01] D. Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281, 2001.

[AKM+17] H. Avron, M. Kapralov, C. Musco, Ch. Musco, A. Velingker, and A. Zandieh. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International conference on machine learning*, pages 253–262. PMLR, 2017.

[Bac17] F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

[BKZ20] O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626. PMLR, 2020.

[BLM13a] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[BLM13b] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[BM01] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250, 2001.

[BTA11] A. Berlinet and Ch. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[Can08] E. J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

[CGK18] A. Chatalic, R. Gribonval, and N. Keriven. Large-scale high-dimensional clustering with fast sketching. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4714–4718. IEEE, 2018.

[Cha20] A. Chatalic. *Efficient and privacy-preserving compressive learning*. PhD thesis, Université Rennes 1, 2020.

[CRS+18] K. Choromanski, M. Rowland, T. Sarlós, V. Sindhwani, R. Turner, and A. Weller. The geometry of random features. In *International Conference on Artificial Intelligence and Statistics*, pages 1–9. PMLR, 2018.

[CRW17] K. M. Choromanski, M. Rowland, and A. Weller. The unreasonable effectiveness of structured random orthogonal embeddings. *Advances in neural information processing systems*, 30, 2017.

[CS01] F. Cucker and S. Smale. On the mathematical foundations of learning. *BULLETIN*, 39, 11 2001.

[CS16] K. Choromanski and V. Sindhwani. Recycling randomness with structure for sublinear time kernel expansions. In *International Conference on Machine Learning*, pages 2502–2510. PMLR, 2016.

[DDSR17] T. Dao, Ch. De Sa, and Ch. Ré. Gaussian quadrature for kernel features. *Advances in neural information processing systems*, 30, 2017.

[FEHP11] C. Forbes, M. Evans, N. Hastings, and B. Peacock. *Statistical distributions*. John Wiley & Sons, 2011.

[FR13] S. Foucart and H. Rauhut. An invitation to compressive sensing. In *A mathematical introduction to compressive sensing*, pages 1–39. Springer, 2013.

[GBKT21a] R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin. Compressive statistical learning with random feature moments. *Mathematical Statistics and Learning*, 3(2):113–164, 2021.

[GBKT21b] Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, and Yann Traonmilin. Statistical learning guarantees for compressive clustering and compressive mixture modeling. *Mathematical Statistics and Learning*, 3(2):165–257, 2021.

[GBR+12] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[GCK+20] R. Gribonval, A. Chatalic, N. Keriven, V. Schellekens, L. Jacques, and P. Schniter. Sketching datasets for large-scale learning (long version). *arXiv preprint arXiv:2008.01839*, 2020.

[Hal50] P. R. Halmos. Measure theory. (University Series in Higher Mathematics) New York: D. Van Nostrand Co., Inc.; London: Macmillan & Co., Ltd., XII, 304 p. (1950)., 1950.

[Hoe94] W. Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[KBGP18] N. Keriven, A. Bourrier, R. Gribonval, and P. Pérez. Sketching for large-scale learning of mixture models. *Information and Inference: A Journal of the IMA*, 7(3):447–508, 2018.

[KTTG17] N. Keriven, N. Tremblay, Y. Traonmilin, and R. Gribonval. Compressive k-means. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373. IEEE, 2017.

[LHCS21] F. Liu, X. Huang, Y. Chen, and J. A.K. Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148, 2021.

[LSS13] Q. Le, T. Sarlós, and A. Smola. Fastfood-approximating kernel expansions in loglinear time. In *Proceedings of the international conference on machine learning*, volume 85, page 8, 2013.

[LTOS19] Z. Li, JF. Ton, D. Oglic, and D. Sejdinovic. Towards a unified analysis of random fourier features. In *International conference on machine learning*, pages 3905–3914. PMLR, 2019.

[MFS+17] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[MKBO18] M. Munkhoeva, Y. Kapushev, E. Burnaev, and I. Oseledets. Quadrature-based features for kernel approximation. *Advances in neural information processing systems*, 31, 2018.

[MM09] O. Maillard and R. Munos. Compressed least-squares regression. *Advances in neural information processing systems*, 22, 2009.

[MM12] O. Maillard and R. Munos. Linear regression with random projections. *Journal of Machine Learning Research*, 13(1):2735–2772, 2012.

[RR07] A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

[Rud17] W. Rudin. *Fourier analysis on groups*. Courier Dover Publications, 2017.

[Sar06] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th annual IEEE symposium on foundations of computer science (FOCS'06)*, pages 143–152. IEEE, 2006.

[SGF+10] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

[SS15a] B. Sriperumbudur and Z. Szabó. Optimal rates for random fourier features. *Advances in neural information processing systems*, 28, 2015.

[SS15b] D. J. Sutherland and J. Schneider. On the error of random fourier features. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, page 862–871, Arlington, Virginia, USA, 2015. AUAI Press.

[SZ21] I. Steinwart and J. F. Ziegel. Strictly proper kernel scores and characteristic kernels on compact spaces. *Applied and Computational Harmonic Analysis*, 51:510–542, 2021.

[TIS76] B. S. Tsirelson, I. A. Ibragimov, and V. N. Sudakov. Norms of gaussian sample functions. In *Proceedings of the Third Japan—USSR Symposium on Probability Theory*, pages 20–41. Springer, 1976.

[Ver18] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[Wai19] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[Wen04] H. Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[YSAM14] J. Yang, V. Sindhwani, H. Avron, and M. Mahoney. Quasi-monte carlo feature maps for shift-invariant kernels. In *International Conference on Machine Learning*, pages 485–493. PMLR, 2014.

[ZM15] J. Zhao and D. Meng. Fastmmd: Ensemble of circular discrepancy for efficient two-sample test. *Neural computation*, 27(6):1345–1372, 2015.