

Instance Optimal Decoding and the Restricted Isometry Property

Nicolas Keriven, Rémi Gribonval

▶ **To cite this version:**

Nicolas Keriven, Rémi Gribonval. Instance Optimal Decoding and the Restricted Isometry Property. 2018. hal-01718411v1

HAL Id: hal-01718411

https://inria.hal.science/hal-01718411v1

Preprint submitted on 27 Feb 2018 (v1), last revised 1 Mar 2018 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Instance Optimal Decoding and the Restricted Isometry Property

Nicolas Keriven

École Normale Supérieure. 45 rue d'Ulm, 75005 Paris, France.

E-mail: nicolas.keriven@ens.fr

Rémi Gribonval

Université Rennes 1, Inria, CNRS, IRISA. F-35000 Rennes, France.

E-mail: remi.gribonval@inria.fr

Abstract. In this paper, we study the preservation of information in ill-posed non-linear inverse problems, where the measured data is assumed to live near a low-dimensional model set. We provide necessary and sufficient conditions for the existence of a so-called *instance optimal* decoder, i.e. that is robust to noise and modelling error. Inspired by existing results in Compressive Sensing, our analysis is based on a (Lower) *Restricted Isometry Property* (LRIP), formulated in a non-linear fashion. We also provide a characterization for non-uniform recovery when the encoding process is randomly drawn, with a new formulation of the LRIP. We finish by describing typical strategies to prove the LRIP in both linear and non-linear cases, and illustrate our results by studying the invertibility of a one-layer neural network with random weights.

1. Introduction

Inverse problems are ubiquitous in all areas of data science. While *linear* inverse problems have been arguably far more studied in the literature, some frameworks are intrinsically non-linear [14]. In this paper, we aim at giving a characterization of the *feasibility* of ill-posed inverse problems, regularized by the introduction of a "low-dimensional" *model set* close to which the data of interest is assumed to live. We consider a very general context, in particular by considering measurement operators that are possibly non-linear, and ambient spaces that can be any pseudometric set. Our main results show that the existence of a decoder that is robust to noise and modelling error is *equivalent* to a modified *Restricted Isometry Property* (RIP), which is a classical property in compressive sensing [9]. We thus outline the fundamental nature of the RIP in settings that are more general than previously studied.

The problem is formulated as follows. Let (E, d_E) be a set equipped with a pseudometric[1] d_E and $(F, \|\cdot\|_F)$ a seminormed[2] vector space, the space of *measurements*. Consider a (possibly non-linear) measurement map $\Psi : E \to F$. We have access to noisy measurements of the form:

$$\mathbf{y} = \Psi(\mathbf{x}^\star) + \mathbf{e} \tag{1}$$

[1] A pseudometric d satisfies all the requirements of a metric except $d(x, y) = 0 \Rightarrow x = y$.

[2] Similarly, a seminorm satisfy the requirements of a norm except that $\|x\| = 0$ does not imply $x = 0$.

where $\mathbf{e} \in F$ is measurement noise and $\mathbf{x}^\star$ is the true signal. Our goal is to characterize the possibility to approximately recover the signal $\mathbf{x}^\star$ from $\mathbf{y}$.

Regularization. In most interesting problems, the "dimension" of the measurement space F is far lower than that of the signal set E (in a loose sense: we recall that here none is required to be finite-dimensional, and E is not necessarily a vector space), which makes the problem *ill-posed*, meaning that there are information-theoretic limits that prevent us from recovering the underlying signal from the measurements, even in the noiseless case. A classical regularization technique is to introduce *prior knowledge* about the true signal $\mathbf{x}^\star$, here we consider a *model set* $\mathfrak{S} \subset E$ of "simple" signals, such that $\mathbf{x}^\star$ is likely to be close to $\mathfrak{S}$. For instance, sparsity, i.e. the assumption that the true signal is a linear combination of a few elements in a well-chosen dictionary, is a hugely-studied prior in modern signal processing, in particular in compressive sensing [15].

Instance Optimal Decoding. Ideally, a *decoder* $\Delta : F \to E$ must be able to exactly retrieve $\mathbf{x}^\star$ from $\mathbf{y}$ when the modelling is exact (i.e. $\mathbf{x}^\star \in \mathfrak{S}$) and the noise is zero ($\mathbf{e} = 0$). However, as these conditions are highly unrealistic in practice, it is desirable for this decoding process to be both robust to noise and stable to *modelling error*. In the literature, such a decoder is said to be *instance optimal* [12] (see Definition 1). In this paper, our goal is to characterize necessary and sufficient conditions for the existence of an instance optimal decoder for the problem (1). Note that we will not study the existence of efficient algorithms to solve (1), which is another significant achievement of compressive sensing [15], but only the *feasibility* of the problem.

In [12, 8], the authors outlined the crucial role played by the *Restricted Isometry Property* (RIP), and more precisely by the *Lower*-RIP (LRIP), for the existence of instance optimal decoders in the linear case. In this paper, we extend these results to the non-linear case and to non-uniform probabilistic recovery.

Outline of the paper. The structure of the paper is as follows. In Section 2 we briefly outline some relevant references, keeping in mind that the field is tremendously large and we do not pretend to be exhaustive in this short paper. In Section 3 we state our main results relating instance optimal decoders (Def. 1) and the LRIP (Def. 2). In Section 4 we outline how one might typically prove the LRIP by extending a classical proof [2], and illustrate it on a simple example.

2. Related Work

Classical Compressive Sensing: the linear case. Instance Optimal decoding and the RIP are well-known notions in the field of compressive sensing [11, 10, 13]. We refer to the book of Foucart and Rauhut [15] for a review of the field, in particular to Chapters 6 and 11 for the topics of interest here. The interplay between the two notions was in particular studied in [12] in the finite-dimensional case. These results were later extended to more general models in [21], and to any linear measurement operators in [8], which is the main inspiration behind the present work.

Non-uniform decoding. In compressed sensing the measurements operator Ψ is often designed at random. Typical recovery results are therefore given *with high probability*. One can then distinguished between *uniform* guarantees, meaning that with high probability on the draw of Ψ all signals $\mathbf{x}^\star$ close to $\mathfrak{S}$ can be stably recovered, and non-uniform guarantees, *i.e.* for one *fixed* signal $\mathbf{x}^\star$ close to $\mathfrak{S}$, with high probability on Ψ the decoding is successful. In [12] the authors study non-uniform instance optimality, but only under the light of the classical uniform RIP.

In this paper we introduce a non-uniform version of the LRIP and prove that it is sufficient for non-uniform instance optimality.

Non-linear inverse problems. Non-linear inverse problems can be found in many areas of signal processing, see e.g. [14] for a review of some applications. They have also been considered by the compressive sensing community, often when *quantization* occurs [19, 7], in the so-called "1-bit" compressed sensing line of work [6]. Another focus is the development of efficient algorithms inspired by the linear case [3, 4]. In [4], the author assume that a locally linearized version of Ψ satisfy the classical RIP. In this paper we consider a different, "fully" non-linear RIP. We note that one notion does not imply the other.

3. Equivalence between IOP and LRIP

In this section we state our main results on instance optimal decoders and the LRIP. We distinguish the case where the operator Ψ is deterministic, or, equivalently, when it is random but one seeks so-called *uniform* recovery guarantees, and the case of non-uniform recovery.

3.1. Deterministic operator, uniform recovery

Recall that we consider a pseudometric set (E, d_E), a seminormed vector space $(F, \|\cdot\|_F)$, and measurements of the form $\mathbf{y} = \Psi(\mathbf{x}^\star) + \mathbf{e}$ where $\Psi : E \to F$. We consider a model set $\mathfrak{S} \subset E$, and are interested in characterizing the existence of a good *decoder* $\Delta(\Psi, \mathbf{y})$ that takes Ψ and $\mathbf{y}$ as inputs and return a signal $\widetilde{\mathbf{x}} \in E$ that is close to $\mathbf{x}^\star$. We want this decoder to be stable to modelling error and robust to noise, which is characterized by the notion of instance optimality.

Definition 1 (Instance Optimality Property (IOP))**.** *A decoder Δ satisfies the Instance Optimality Property for the operator Ψ and model $\mathfrak{S}$ with constants $A, B > 0$, pseudometrics d_E, d'_E on E and error $\lambda \geq 0$ if: for all signals $\mathbf{x}^\star \in E$ and noise $\mathbf{e} \in F$, denoting $\widetilde{\mathbf{x}} = \Delta(\Psi, \Psi(\mathbf{x}^\star) + \mathbf{e})$ the recovered signal, it holds that:*

$$d_E(\mathbf{x}^\star, \widetilde{\mathbf{x}}) \leq A d'_E(\mathbf{x}^\star, \mathfrak{S}) + B \|\mathbf{e}\|_F + \lambda \tag{2}$$

where $d'_E(\mathbf{x}, \mathfrak{S}) = \inf_{\mathbf{x}_\mathfrak{S} \in \mathfrak{S}} d'_E(\mathbf{x}, \mathbf{x}_\mathfrak{S})$.

As indicated by the r.h.s. of (2), the decoding error between the true signal and the recovered one is bounded by the amplitude of the noise and the distance from $\mathbf{x}^\star$ to the model set, which indicates how well $\mathbf{x}^\star$ is modelled by $\mathfrak{S}$. We also consider the possibility that the decoding error includes a fixed additional error $\lambda \geq 0$, which may be unavoidable in some cases (due to algorithmic precision for instance). Ideally, one has $\lambda = 0$.

Let us now turn to the proposed non-linear version of the LRIP. As described in [8], the LRIP is just one side of the classical RIP, which states that the measurement operator Ψ approximately preserves distances between elements of the model $\mathfrak{S}$.

Definition 2 (Lower Restricted Isometry Property (LRIP))**.** *The operator Ψ satisfies the Lower Restricted Isometry Property for the model $\mathfrak{S}$ with constant $\alpha > 0$, pseudometric d_E and error $\eta \geq 0$ if: for all $\mathbf{x}, \mathbf{x}' \in \mathfrak{S}$ it holds that*

$$d_E(\mathbf{x}, \mathbf{x}') \leq \alpha \left\| \Psi(\mathbf{x}) - \Psi\left(\mathbf{x}'\right) \right\|_F + \eta. \tag{3}$$

The LRIP expresses the fact that Ψ must not collapse two elements of the model together. Like the IOP, we allow for a possible additional fixed error $\eta \geq 0$ in the LRIP. Note that this type of error is often considered when introducing quantization [7, 19]. Ideally, one has $\eta = 0$, however in some cases it can be considerably simpler to prove that the LRIP holds with a non-zero η [20]. The reader would have noted that the classical RIP is often expressed with a constant $\alpha = (1 - t)^{-1}$ where $t < 1$ is a small as possible.

We now state our main result. The proof, rather direct, can be found in Appendix A.

Theorem 1 (Equivalence between IOP and LRIP.). *Consider an operator Ψ and a model $\mathfrak{S}$.*

(i) If there exists a decoder Δ which satisfies the Instance Optimality Property for Ψ and $\mathfrak{S}$ with constants $A, B > 0$, pseudometrics d_E, d'_E and error $\lambda \geq 0$, then the operator Ψ satisfies the LRIP for $\mathfrak{S}$ with constant $\alpha = B$, pseudometric d_E and error $\eta = 2\lambda$.

(ii) If the operator Ψ satisfies the LRIP for the model $\mathfrak{S}$ with constant α, pseudometric d_E and error $\eta \geq 0$, then the decoder Δ defined as[3]

$$\Delta(\Psi, \mathbf{y}) = \underset{\mathbf{x} \in \mathfrak{S}}{\arg\min} \left\| \Psi(\mathbf{x}) - \mathbf{y} \right\|_F \tag{4}$$

satisfies the Instance Optimality Property for the operator Ψ and model $\mathfrak{S}$ with constants $A = 1$ and $B = 2\alpha$, pseudometrics d_E and d'_E where d'_E is defined by $d'_E(\mathbf{x}, \mathbf{x}') := d_E(\mathbf{x}, \mathbf{x}') + 2\alpha \left\| \Psi(\mathbf{x}) - \Psi(\mathbf{x}') \right\|_F$, and error $\lambda = \eta$.

Theorem 1 states that if the LRIP is satisfied, then the decoder that returns the element in the model that best matches the measurement is instance optimal, with a special metric d'_E. On the other hand, if *some* instance optimal decoder exists, then the LRIP must be satisfied. In other words, when the LRIP is satisfied, then we know that a negligible amount of information is lost when encoding a signal well-modeled by $\mathfrak{S}$. Conversely, if the LRIP is not satisfied, one has no hope of deriving an instance optimal decoder.

3.2. Random operator, from uniform recovery to non-uniform recovery

In the vast majority of the compressive sensing literature, the measurement process is drawn at random: for instance, in the finite dimensional case, it is an open problem to find deterministic matrices that satisfies the RIP with an optimal number of measurements ([15], pp. 27), while on the contrary many classes of random matrices satisfy the RIP with high probability [2].

A well-studied concept is that of uniform recovery guarantees, where one shows that, with high probability on the draw of Ψ, the LRIP holds. It follows by Theorem 1 that there is a decoder such that, with high probability on the draw of Ψ, all signals from $\mathfrak{S}$ can be stably recovered. There is also a notion of *non-uniform* recovery, where one considers a decoder Δ and wonders if, given an arbitrary signal close to $\mathfrak{S}$, this signal is stably recovered (with high probability on the draw of Ψ) from $\Psi(\mathbf{x}^\star) + \mathbf{e}$. In this section we introduce a non-uniform version of the LRIP, and show that it is a sufficient condition for the existence of a non-uniform instance optimal decoder. We start by discussing a notion of projection on the model.

Remark 1 (Projection.). *As we will see, in non-uniform recovery the distance from $\mathbf{x}^\star$ to $\mathfrak{S}$ is replaced by the distance from $\mathbf{x}^\star$ to a particular element $\mathbf{x}_{\mathfrak{S}} = \mathrm{P}_{\mathfrak{S}}(\mathbf{x}^\star)$, where $\mathrm{P}_{\mathfrak{S}} : E \to \mathfrak{S}$ is some "approximate" projection function with respect to some metric d'_E. In full generality, it is not guaranteed that there exists $\mathrm{P}_{\mathfrak{S}}$ such that $d'_E(\mathbf{x}^\star, \mathrm{P}_{\mathfrak{S}}(\mathbf{x}^\star)) = d'_E(\mathbf{x}^\star, \mathfrak{S})$, but one can always define it such that for all $\mathbf{x}^\star$, $d'_E(\mathbf{x}^\star, \mathrm{P}_{\mathfrak{S}}(\mathbf{x}^\star)) \leq d'_E(\mathbf{x}^\star, \mathfrak{S}) + \varepsilon$ for an arbitrary small ε.*

Let us now introduce the proposed non-uniform IOP and LRIP.

Definition 3 (Non-uniform IOP). *A decoder Δ satisfies the non-uniform Instance Optimality Property for the (random) mapping Ψ, model $\mathfrak{S}$ and projection function $\mathrm{P}_{\mathfrak{S}}$, with constants $A, B > 0$, pseudometrics d_E, d'_E, probability $1 - \rho$ and error $\lambda \geq 0$ if:*

$$\forall \mathbf{x}^\star \in E, \quad \mathbb{P}_\Psi \left[\forall \mathbf{e} \in F,\ d_E(\mathbf{x}^\star, \widetilde{\mathbf{x}}) \leq A d'_E(\mathbf{x}^\star, \mathrm{P}_{\mathfrak{S}}(\mathbf{x}^\star)) + B \left\| \mathbf{e} \right\|_F + \lambda \right] \geq 1 - \rho \tag{5}$$

where $\Delta(\Psi, \Psi(\mathbf{x}^\star) + \mathbf{e})$ is denoted by $\widetilde{\mathbf{x}}$.

[3] In this paper we assume that the minimization problem $\arg\min_{\mathbf{x} \in \mathfrak{S}} \left\| \Psi(\mathbf{x}) - \mathbf{y} \right\|_F$ has at least one solution, for simplicity (ties can be broken arbitrarily). When this is not the case, it is possible to consider a decoder that returns any element that approaches the infimum with a fixed precision, at the expense of having this precision in the decoding error λ, as in [8].

Note that in this definition the IOP is non-uniform with respect to the data $\mathbf{x}^\star$ but uniform with respect to the noise $\mathbf{e}$, meaning that with high probability on the draw of Ψ the (fixed) data can be stably recovered from any noisy measurement vector.

Definition 4 (Non-uniform LRIP)**.** *The operator Ψ satisfies the non-uniform LRIP for the model $\mathfrak{S}$ with constant $\alpha > 0$, pseudometric d_E, probability $1-\rho$ and error $\eta \geq 0$ if:*

$$\forall \mathbf{x} \in \mathfrak{S}, \quad \mathbb{P}_\Psi\Big[\forall \mathbf{x}' \in \mathfrak{S},\ d_E(\mathbf{x},\mathbf{x}') \leq \alpha \left\|\Psi(\mathbf{x}) - \Psi(\mathbf{x}')\right\|_F + \eta\Big] \geq 1-\rho. \tag{6}$$

This LRIP is in fact "semi"-uniform: it is non-uniform with respect to one element $\mathbf{x}$ but uniform with respect to $\mathbf{x}'$. A "fully" non-uniform LRIP would, in fact, be almost always valid (see Section 4), and thus probably too weak to yield recovery guarantees.

Before stating our result, let us remark that the definition of the metric d'_E in Theorem 1 (*ii*) involves the operator Ψ, which is potentially problematic when it is random. To solve this, [12] introduces a so-called *Boundedness Property* (BP) in the classical sparse setting in finite dimension. We extend this notion in the considered context here.

Definition 5 (Boundedness property (BP))**.** *The operator Ψ satisfies the Boundedness Property with constant β, pseudometric d_G and probability $1-\rho$ if:*

$$\forall \mathbf{x} \in E,\ \forall \mathbf{x}_{\mathfrak{S}} \in \mathfrak{S}, \quad \mathbb{P}_\Psi\Big[\ \|\Psi(\mathbf{x}) - \Psi(\mathbf{x}_{\mathfrak{S}})\|_F \leq \beta d_G(\mathbf{x},\mathbf{x}_{\mathfrak{S}})\ \Big] \geq 1-\rho. \tag{7}$$

We then have the following result, proved in Appendix B.

Theorem 2 (The non-uniform LRIP and BP implies the non-uniform IOP)**.** *Consider a random operator Ψ. Assume that:*

(i) the operator Ψ satisfies the non-uniform LRIP for the model $\mathfrak{S}$ with constant $\alpha > 0$, pseudometric d_E, probability $1-\rho_1$ and error $\eta \geq 0$;

(ii) the operator Ψ satisfies the non-uniform Boundedness Property with constant β, pseudometric d_G and probability $1-\rho_2$;

Then, the decoder Δ defined by (4) *satisfies the non-uniform Instance Optimality Property for the operator Ψ, model $\mathfrak{S}$ and any projection function $\mathrm{P}_{\mathfrak{S}}$ with constants $A = 1$, $B = 2\alpha$, pseudometrics d_E and $d'_E := d_E + 2\alpha\beta d_G$, probability $1-\rho_1-\rho_2$ and error $\lambda = \eta$.*

Compared with the result in [12], which proves non-uniform recovery under a *uniform* LRIP and non-uniform BP in the finite-dimensional case, our result holds under weaker hypotheses. For the converse implication, unlike Theorem 1, the non-uniform IOP does not seem to directly imply the non-uniform LRIP.

4. A typical proof of the LRIP

In this section, we outline a possible strategy to prove the LRIP, inspired by the proof for random matrices in [2]. This relatively simple proof has two steps: first, a pointwise concentration result, and second, an extension by covering arguments. For a set S and a metric d, we denote by $\mathcal{N}(S,d,\delta) \in \mathbb{N} \cup \{+\infty\}$ the minimum number of balls of radius δ, with centers that belong to S, required to cover S.

4.1. Linear case

We start with the linear case, which follows closely the proof in [2]. We treat the uniform case (Def. 2), with no error ($\eta = 0$). Assume (E, d_E) and $(F, \|\cdot\|_F)$ are both vector spaces, and that we have a random *linear* operator $\Psi : E \to F$ such that the following *concentration* result holds:

$$\forall \mathbf{x},\mathbf{x}' \in \mathfrak{S} \text{ s.t. } d_E(\mathbf{x},\mathbf{x}') > 0, \quad \mathbb{P}_\Psi\left(\frac{\|\Psi(\mathbf{x}-\mathbf{x}')\|_F}{d_E(\mathbf{x},\mathbf{x}')} - 1 \leq -t\right) \leq e^{-c(t)} \tag{8}$$

for an increasing *concentration function* $c(t) > 0$. Typically, the "bigger" the space F is (*i.e.* the more measurements we collect), the higher the concentration function is: often, for m measurements ($F = \mathbb{R}^m$ or $\mathbb{C}^m$), classical concentration inequalities yield $c(t) \propto mt^2$.

This property proves a "pointwise" (or "fully" non-uniform) LRIP: for two given $\mathbf{x}, \mathbf{x}'$, the quantity $\|\Psi(\mathbf{x} - \mathbf{x}')\|_F$ is a good approximation of $d_E(\mathbf{x}, \mathbf{x})$ with high probability. We now invert the quantifiers by covering arguments. From the formulation of the concentration (8) we see that a particular set of interest is the so-called *normalized secant set* [22]:

$$\mathcal{S} = \left\{ \tfrac{\mathbf{x}-\mathbf{x}'}{d_E(\mathbf{x},\mathbf{x}')} \mid \mathbf{x}, \mathbf{x}' \in \mathfrak{S},\ d_E(\mathbf{x}, \mathbf{x}') > 0 \right\} \subset E \tag{9}$$

The proof of the following result is in Appendix C.

Proposition 1. *Consider $0 < t < 1$. Assume that the concentration property (8) holds, that $\mathcal{S}$ has finite covering numbers, and that for any draw of Ψ and any $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$ we have $\|\Psi(\mathbf{s} - \mathbf{s}')\|_F \leq C d_E(\mathbf{s}, \mathbf{s}')$. Set $\delta = \frac{t}{2C}$. Define the probability of failure*

$$\rho = \mathcal{N}(\mathcal{S}, d_E, \delta) \cdot e^{-c(t/2)}$$

Then the operator Ψ satisfies the uniform LRIP for the model $\mathfrak{S}$ with constant $\alpha = (1-t)^{-1}$, metric d_E, probability $1 - \rho$ and error $\eta = 0$.

This proof of the RIP has been used for instance in classical compressive sensing [2] or for random linear embeddings of Radon measures [18]. It is also used in a constructive manner to build appropriate operators Ψ in [22].

4.2. Non-linear case

It is possible to adapt the previous proof to non-linear operators, by distinguishing the case where $\mathbf{x}$ and $\mathbf{x}'$ are "close", for which we resort to a linearization of Ψ and properties of the normalized secant set, and the case where $\mathbf{x}$ and $\mathbf{x}'$ are distant from each other, for which we use directly the covering numbers of the model. We treat here the non-uniform case (Def. 4).

Assume again that (E, d_E) and $(F, \|\cdot\|_F)$ are vector spaces. Assume that we have a random map $\Psi : E \to F$ such that the concentration property (8) holds. Next, suppose that there exists $\varepsilon_0 > 0$ such that for any fixed $\mathbf{x} \in \mathfrak{S}$:

(i) for all $\mathbf{x}' \in \mathfrak{S}$ and any draw of Ψ, $\|\Psi(\mathbf{x}) - \Psi(\mathbf{x}')\|_F \leq C_1 d_E(\mathbf{x}, \mathbf{x}')$

(ii) the model $\mathfrak{S}$ has finite covering numbers with respect to d_E, and in particular it also has finite diameter $M_{\mathfrak{S}} = \sup_{\mathbf{x},\mathbf{x}' \in \mathfrak{S}} d_E(\mathbf{x}, \mathbf{x}')$.

(iii) for all $0 < \varepsilon \leq \varepsilon_0$, the following version of the normalized secant set $\mathcal{S}_\varepsilon := \left\{ \tfrac{\mathbf{x}-\mathbf{x}'}{d_E(\mathbf{x},\mathbf{x}')} \mid \mathbf{x}' \in \mathfrak{S},\ 0 < d_E(\mathbf{x}, \mathbf{x}') \leq \varepsilon \right\}$ has finite covering numbers.

(iv) for all $\mathbf{x}' \in \mathfrak{S}$ such that $d_E(\mathbf{x}, \mathbf{x}') \leq \varepsilon_0$, and any draw of Ψ, we have $|\Psi(\mathbf{x}) - \Psi(\mathbf{x}') - D_\Psi(\mathbf{x} - \mathbf{x}')| \leq C_2 d_E(\mathbf{x}, \mathbf{x}')^2$, where $D_\Psi : E \to F$ is a linear map such that for all $\mathbf{s}, \mathbf{s}' \in \mathcal{S}_{\varepsilon_0}$, $\|D_\Psi(\mathbf{s} - \mathbf{s}')\|_F \leq C_3 d_E(\mathbf{s}, \mathbf{s}')$.

The following result is proved in Appendix D.

Proposition 2. *Assume that the properties $(i) - (iv)$ above are satisfied. Set $\varepsilon \leq \min\left(\varepsilon_0, \frac{t}{8C_2}\right)$, $\delta' \leq \frac{t}{4C_3}$ and $\delta \leq \frac{t\varepsilon^2/(4C_1)}{\varepsilon + M_{\mathfrak{S}}}$. Define the probability of failure*

$$\rho := \left[\mathcal{N}(\mathfrak{S}, d_E, \delta) + \mathcal{N}\left(\mathcal{S}_\varepsilon, d_E, \delta'\right)\right] \cdot e^{-c(t/2)}$$

Then the operator Ψ satisfies the non-uniform LRIP for the model $\mathfrak{S}$ with constant $\alpha = (1-t)^{-1}$, metric d_E, probability $1 - \rho$ and error $\eta = 0$.

4.3. Illustration

In this section we illustrate the non-linear LRIP on a simple example; that of recovering a vector from a *random features embedding*, which is a random map initially designed for kernel approximation, see [23, 24]. Such a random embedding can be seen as a one-layer neural network with random weights, for which invertibility and preservation of information have recently been topics of interest [16, 17].

Consider $E = \mathbb{R}^d$ and define $\mathfrak{S}$ to be a *Union of Subspaces*, which is a popular model in compressed sensing [5], with controlled norm: $\mathfrak{S} = \bigcup_{i=1}^{N} (S_i \cap B_M)$ where $B_M = \{\mathbf{x},\ \|\mathbf{x}\|_2 \leq M\}$ and each S_i is an s-dimensional subspace of $\mathbb{R}^d$. We choose a sampling that is a reweighted version of the original Fourier sampling for kernel approximation [23] with increased stability, as in [18], for the Gaussian kernel with bandwidth $\sigma > 0$. It is defined as follows: for a number of measurements m, the measurements space is $F = \mathbb{C}^m$, the random map is defined as $\Psi(\mathbf{x}) = \frac{1}{\sqrt{m}} \left[\varphi_{\boldsymbol{\omega}_j}(\mathbf{x})\right]_{j=1}^{m}$ where $\varphi_{\boldsymbol{\omega}}(\mathbf{x}) = \frac{e^{\mathrm{i}\boldsymbol{\omega}_j^\top \mathbf{x}}}{f(\boldsymbol{\omega})}$ and $\boldsymbol{\omega}_j \in \mathbb{R}^d$ are drawn *i.i.d.* from $\Lambda = f(\boldsymbol{\omega})^2 \mathcal{N}(0, \sigma^{-2}\mathbf{I})$ (where $\mathcal{N}$ is a Gaussian), with $f(\boldsymbol{\omega})^2 = \frac{1}{3}\sum_{\ell=0}^{2} \frac{\|\boldsymbol{\omega}\|_2^{2\ell}}{\gamma_{2\ell}}$ and $\gamma_\ell = \mathbb{E}_{\boldsymbol{\omega}\sim\mathcal{N}(0,\sigma^{-2}\mathbf{I})} \|\boldsymbol{\omega}\|_2^\ell = \mathcal{O}\left(\frac{d^{\ell/2}}{\sigma^\ell}\right)$. One can verify that $\int d\Lambda(\boldsymbol{\omega}) = 1$, *i.e.* Λ is a valid probability distribution. The metric d_E is here the *kernel metric* associated to the Gaussian kernel with bandwidth σ:

$$d_E(\mathbf{x}, \mathbf{x}') := 2\left(1 - \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\sigma^2}\right)\right)$$

We have the following result, which proof is in Appendix E.

Proposition 3. *If the number of measurements is such that*

$$m \gtrsim t^{-2} \cdot \left(s \cdot \log\left(\tfrac{Md}{\sigma t}\right) + \log(N) + \log\left(\tfrac{1}{\rho}\right)\right)$$

Then the operator Ψ satisfies the non-uniform LRIP for the model $\mathfrak{S}$ with constant $\alpha = (1-t)^{-1}$, metric d_E, probability $1-\rho$ and error $\eta = 0$, as well as the BP (Def. 5) with constant $\beta = 1+t$, metric d_E and probability $1-\rho$.

Hence, using Theorem 2, we have shown that a reduced number of random features preserves all information when encoding signals that are (in this case) well-modeled by a union of subspaces, *with respect to the associated kernel metric*. This preliminary analysis may have consequences for classical random feature bounds in a learning context [1, 25].

5. Conclusion

In this paper we generalized a classical property, the equivalence between the existence of an instance-optimal decoder and the LRIP, to non-linear inverse problems with possible quantization error. We also formulated a version of the result for non-uniform recovery, by introducing a non-uniform version of the LRIP. To further illustrate this principle, we provided a typical proof strategy for the LRIP that one might use in practical contexts, and gave an example of non-linear LRIP on random features for kernel approximation.

Although relatively simple in their proofs, these results may have important consequences for a large class of linear or non-linear inverse problems, where one seeks stable and robust recovery. Naturally, once the LRIP guarantees (or disproves) the existence of an instance optimal decoder, an outstanding question is the existence of *efficient* algorithms that provide equivalent guarantees, as in classical compressive sensing [15] or some of its recent extensions [26].

References

[1] Francis Bach. On the Equivalence between Kernel Quadrature Rules and Random Feature Expansions. *Journal of Machine Learning Research*, pages 1–38, 2017.

[2] Richard Baraniuk, Mark Davenport, Ronald A Devore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[3] Amir Beck and Yonina C. Eldar. Sparsity Constrained Nonlinear Optimization: Optimality Conditions and Algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

[4] Thomas Blumensath. Compressed sensing with nonlinear observations and related nonlinear optimization problems. *Information Theory, IEEE Transactions on*, pages 1–19, 2013.

[5] Thomas Blumensath and Mike E. Davies. Sampling theorems for signals from the union of finite-dimensional linear subspaces. *IEEE Transactions on Information Theory*, 55(4):1872–1882, 2009.

[6] Petros T Boufounos and Richard G Baraniuk. 1-Bit Compressive Sensing. In *Information Sciences and Systems*, pages 16–21, 2008.

[7] Petros T. Boufounos, Shantanu Rane, and Hassan Mansour. Representation and Coding of Signal Geometry. *Information and Inference: a Journal of the IMA*, 6(4):349–388, 2017.

[8] Anthony Bourrier, Mike E. Davies, Tomer Peleg, and Rémi Gribonval. Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems. *IEEE Transactions on Information Theory*, 60(12):7928–7946, 2014.

[9] Emmanuel J Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9-10):1–4, 2008.

[10] Emmanuel J. Candès, Justin K. Romberg, and Terrence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):480–509, 2006.

[11] Emmanuel J. Candès and Terrence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[12] Albert Cohen, Wolfgang Dahmen, and Ronald A Devore. Compressed sensing and best k-term approximation. *Journal of the American mathematical Society*, 22(1):211–231, 2009.

[13] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[14] Heinz W Engl and Philipp Kügler. Nonlinear inverse problems: theoretical aspects and some industrial applications. In *Multidisciplinary Methods for Analysis Optimization and Control of Complex Systems*, volume 06, pages 3–47. 2005.

[15] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Springer New York, NY, 2013.

[16] Anna C. Gilbert, Yi Zhang, Kibok Lee, Yuting Zhang, and Honglak Lee. Towards understanding the invertibility of convolutional neural networks. *IJCAI International Joint Conference on Artificial Intelligence*, pages 1703–1710, 2017.

[17] Raja Giryes, Guillermo Sapiro, and Alex M Bronstein. Deep Neural Networks with Random Gaussian Weights : A Universal Classification Strategy ? *IEEE Transactions on Signal Processing*, 64(13):3444–3457, 2016.

[18] Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, and Yann Traonmilin. Compressive Statistical Learning with Random Feature Moments. *arXiv:1706.07180*, 2017.

[19] Laurent Jacques. Small width, low distortions: quasi-isometric embeddings with quantized sub-Gaussian random projections. pages 1–26, 2015.

[20] Nicolas Keriven, Anthony Bourrier, Rémi Gribonval, and Patrick Pérèz. Sketching for Large-Scale Learning of Mixture Models. *Information and Inference: A Journal of the IMA*, pages 1–62, 2017.

[21] Tomer Peleg, Rémi Gribonval, and Mike E. Davies. Compressed Sensing and Best Approximation from Unions of Subspaces: Beyond Dictionaries. *European Signal Processing Conference (EUSIPCO)*, 2013.

[22] Gilles Puy, Mike E. Davies, and Rémi Gribonval. Linear embeddings of low-dimensional subsets of a Hilbert space to R m. In *European Signal Processing Conference (EUSIPCO)*, pages 469–473, 2015.

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

[24] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[25] Alessandro Rudi and Lorenzo Rosasco. Generalization Properties of Learning with Random Features. In *Advances in Neural Information Processing System (NIPS)*, 2017.

[26] Yann Traonmilin and Rémi Gribonval. Stable recovery of low-dimensional cones in Hilbert spaces : One RIP to rule them all. *Applied and Computational Harmonic Analysis*, 2016.

Appendix A. Proof of Theorem 1

(i) Consider $\mathbf{x}, \mathbf{x}' \in \mathfrak{S}$. By triangular inequality we have

$$d_E(\mathbf{x}, \mathbf{x}') \leq d_E(\mathbf{x}, \Delta(\Psi, \Psi(\mathbf{x}')) + d_E(\Delta(\Psi, \Psi(\mathbf{x}')), \mathbf{x}').$$

Then, by applying the Instance Optimality Property with noise $\mathbf{e} := \Psi(\mathbf{x}') - \Psi(\mathbf{x})$ we get $d_E(\mathbf{x}, \Delta(\Psi, \Psi(\mathbf{x}')) \leq B \|\Psi(\mathbf{x}') - \Psi(\mathbf{x})\|_F + \lambda$, and by applying again the Instance Optimality Property it holds that $d_E(\Delta(\Psi, \Psi(\mathbf{x}')), \mathbf{x}') \leq \lambda$, hence the result.

(ii) Consider any signal $\mathbf{x}^\star \in E$ and noise $\mathbf{e} \in F$, denote $\mathbf{y} = \Psi(\mathbf{x}^\star) + \mathbf{e}$ and $\widetilde{\mathbf{x}} = \Delta(\Psi, \mathbf{y})$. Let $\mathbf{x}_{\mathfrak{S}} \in \mathfrak{S}$ be any element of the model. We have:

$$\begin{aligned} d_E(\mathbf{x}^\star, \widetilde{\mathbf{x}}) &\leq d_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + d_E(\mathbf{x}_{\mathfrak{S}}, \widetilde{\mathbf{x}}) \\ &\overset{\text{LRIP}}{\leq} d_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + \alpha \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\widetilde{\mathbf{x}})\|_F + \eta \\ &\leq d_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + \alpha \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \mathbf{y}\|_F + \alpha \|\mathbf{y} - \Psi(\widetilde{\mathbf{x}})\|_F + \eta \,. \end{aligned}$$

By definition of the decoder (4) we have $\|\Psi(\widetilde{\mathbf{x}}) - \mathbf{y}\|_F \leq \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \mathbf{y}\|_F$ and therefore

$$\begin{aligned} d_E(\mathbf{x}^\star, \widetilde{\mathbf{x}}) &\leq d_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + 2\alpha \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \mathbf{y}\|_F + \eta \\ &\leq d_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + 2\alpha \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\mathbf{x}^\star)\|_F + 2\alpha \|\Psi(\mathbf{x}^\star) - \mathbf{y}\|_F + \eta \\ &\leq d'_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + 2\alpha \|\mathbf{e}\|_F + \eta \end{aligned}$$

where $d'_E(\mathbf{x}, \mathbf{x}') = d_E(\mathbf{x}, \mathbf{x}') + 2\alpha \|\Psi(\mathbf{x}) - \Psi(\mathbf{x}')\|_F$. Since the result is valid for all $\mathbf{x}_{\mathfrak{S}} \in \mathfrak{S}$, we can take the infimum of $d'_E(\mathbf{x}, \mathbf{x}_{\mathfrak{S}})$ with respect to $\mathbf{x}_{\mathfrak{S}} \in \mathfrak{S}$ and obtain the result.

Appendix B. Proof of Theorem 2

Let $\mathbf{x}^\star \in E$ be a fixed signal and denote by $\mathbf{x}_{\mathfrak{S}} = \mathrm{P}_{\mathfrak{S}}(\mathbf{x}^\star) \in \mathfrak{S}$.

Applying the non-uniform LRIP, with probability at least $1 - \rho_1$ on the draw of the operator Ψ we have

$$\forall \mathbf{x}' \in \mathfrak{S},\ d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}') \leq \alpha \left\|\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi\left(\mathbf{x}'_{\mathfrak{S}}\right)\right\|_F + \eta \,. \tag{B.1}$$

In the same fashion, applying the non-uniform Boundedness Property, with probability at least $1 - \rho_2$ on the draw of the operator Ψ we have

$$\|\Psi(\mathbf{x}^\star) - \Psi(\mathbf{x}_{\mathfrak{S}})\|_F \leq \beta d_G(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) \,. \tag{B.2}$$

Therefore, by a union bound, with probability at least $1 - \rho_1 - \rho_2$, both (B.1) and (B.2) are satisfied. If this is the case, for all noise vector $\mathbf{e} \in F$, denoting by $\mathbf{y} = \Psi(\mathbf{x}^\star) + \mathbf{e}$ and $\widetilde{\mathbf{x}} = \Delta(\Psi, \mathbf{y})$, it holds that:

$$\begin{aligned} d_E(\mathbf{x}^\star, \widetilde{\mathbf{x}}) &\leq d_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + d_E(\mathbf{x}_{\mathfrak{S}}, \widetilde{\mathbf{x}}) \\ &\overset{\text{(B.1)}}{\leq} d_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + \alpha \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\widetilde{\mathbf{x}})\|_F + \eta \\ &\leq d_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + \alpha \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \mathbf{y}\|_F + \alpha \|\mathbf{y} - \Psi(\widetilde{\mathbf{x}})\|_F + \eta \,. \end{aligned}$$

Once again by definition of the decoder (4) we have $\|\Psi(\widetilde{\mathbf{x}}) - \mathbf{y}\|_F \leq \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \mathbf{y}\|_F$ and therefore

$$\begin{aligned} d_E(\mathbf{x}^\star, \widetilde{\mathbf{x}}) &\leq d_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + 2\alpha \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \mathbf{y}\|_F + \eta \\ &\leq d_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + 2\alpha \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\mathbf{x}^\star)\|_F + 2\alpha \|\Psi(\mathbf{x}^\star) - \mathbf{y}\|_F + \eta \\ &\leq d'_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + 2\alpha \|\mathbf{e}\|_F + \eta. \end{aligned}$$

by applying (B.2), which is the desired result.

Appendix C. Proof of Proposition 1

From the definition of the normalized secant set, our goal is to prove that with high probability on the draw of Ψ, for all $\mathbf{s} \in \mathcal{S}$ we have $\|\Psi \mathbf{s}\|_F \geq 1 - t$.

Let $\delta, \varepsilon > 0$ be small constants that we shall define later and $N := \mathcal{N}(\mathcal{S}, d_E, \delta)$ be the covering numbers of the normalized secant set. Let $\mathbf{s}_1, \ldots, \mathbf{s}_N$ be an δ-covering of the normalized secant set $\mathcal{S}$.

By the concentration property, it holds that, with probability at least $1 - Ne^{-c(t/2)}$, for all $\mathbf{s}_i$ we have

$$\|\Psi \mathbf{s}_i\|_F \geq 1 - t/2 \tag{C.1}$$

Now, given any element of the normalized secant set $\mathbf{s} \in \mathcal{S}$, one can find an element of the covering $\mathbf{s}_i$ such that $\|\mathbf{s} - \mathbf{s}_i\|_E \leq \delta$. Assuming (C.1) holds, we have

$$\|\Psi \mathbf{s}\|_F \geq \|\Psi \mathbf{s_i}\|_F - \|\Psi(\mathbf{s} - \mathbf{s_i})\|_F \geq 1 - \frac{t}{2} - C\delta$$

and therefore by choosing $\delta = t/(2C)$ we obtain the desired result.

Appendix D. Proof of Proposition 2

Fix any $\mathbf{x}_{\mathfrak{S}} \in \mathfrak{S}$. Let $\delta, \delta', \varepsilon > 0$ be small constants that we shall define later such that $\varepsilon \leq \varepsilon_0$. Define $\mathfrak{S}_\varepsilon = \mathfrak{S} \backslash \{\mathbf{x} \in \mathfrak{S},\ d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}) \leq \varepsilon\}$ the model with a ball around $\mathbf{x}_{\mathfrak{S}}$ removed, and $N := \mathcal{N}(\mathfrak{S}_\varepsilon, d_E, \delta)$, $N' := \mathcal{N}(\mathcal{S}_\varepsilon, d_E, \delta')$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and $\mathbf{s}_1, \ldots, \mathbf{s}_{N'}$ be, respectively, an δ-covering of $\mathfrak{S}_\varepsilon$ and a δ'-covering of $\mathcal{S}_\varepsilon$, with $\mathbf{x}'_i \in \mathfrak{S}$ defined such that $\mathbf{s}_i = \frac{\mathbf{x}_{\mathfrak{S}} - \mathbf{x}'_i}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}'_i)}$ and $0 < d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}'_i) \leq \varepsilon$.

By the concentration property, it holds that, with probability at least $1 - (N + N')e^{-c(t/2)}$, for all $\tilde{\mathbf{x}} = \mathbf{x}_i$ or $\tilde{\mathbf{x}} = \mathbf{x}'_i$ for all indices i, we have

$$\|\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\tilde{\mathbf{x}})\|_F \geq (1 - t/2) \cdot d_E(\mathbf{x}_{\mathfrak{S}}, \tilde{\mathbf{x}}). \tag{D.1}$$

Our goal is to extend this property to any element $\tilde{\mathbf{x}} = \mathbf{x}$ in the model $\mathfrak{S}$. Let $\mathbf{x} \in \mathfrak{S}$ be any element of the model. We distinguish two cases. If $\mathbf{x} \in \mathfrak{S}_\varepsilon$, *i.e.* $d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}) \geq \varepsilon$, we consider $\mathbf{x}_i \in \mathfrak{S}_\varepsilon$ an element of the covering of $\mathfrak{S}_\varepsilon$ such that $d_E(\mathbf{x}, \mathbf{x}_i) \leq \delta$. We have:

$$\begin{aligned}
\left\| \frac{\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\mathbf{x})}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x})} - \frac{\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\mathbf{x}_i)}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}_i)} \right\|_F &\leq \left\| \frac{\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\mathbf{x})}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x})} - \frac{\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\mathbf{x}_i)}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x})} \right\|_F + \left\| \frac{\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\mathbf{x}_i)}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x})} - \frac{\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\mathbf{x}_i)}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}_i)} \right\|_F \\
&= \frac{\|\Psi(\mathbf{x}) - \Psi(\mathbf{x}_i)\|_F}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x})} + \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\mathbf{x}_i)\|_F \left| \frac{1}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x})} - \frac{1}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}_i)} \right| \\
&\overset{(i),(ii)}{\leq} \frac{C_1 \delta}{\varepsilon} + C_1 M_{\mathfrak{S}} \frac{d_E(\mathbf{x}, \mathbf{x}_i)}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}) d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}_i)} \\
&\leq \frac{C_1 \delta}{\varepsilon} \left(1 + \frac{M_{\mathfrak{S}}}{\varepsilon}\right)
\end{aligned}$$

and therefore

$$\left\| \frac{\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\mathbf{x})}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x})} \right\|_F \geq 1 - t/2 - \frac{C_1 \delta}{\varepsilon} \left(1 + \frac{M_{\mathfrak{S}}}{\varepsilon}\right) \tag{D.2}$$

Now, when $d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}) \leq \varepsilon$, we define $\mathbf{z} := \frac{\mathbf{x}_{\mathfrak{S}} - \mathbf{x}}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x})}$ and note that $\mathbf{z} \in \mathcal{S}_\varepsilon$. We approximate it by an element of the covering of the normalized secant set $\mathbf{s}_i = \frac{\mathbf{x}_{\mathfrak{S}} - \mathbf{x}'_i}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}'_i)} \in \mathcal{S}_\varepsilon$ (meaning that $d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}'_i) \leq \varepsilon$) that verifies $d_E(\mathbf{z}, \mathbf{s}_i) \leq \delta'$. Then, we have

$$\begin{aligned}
\left\| \frac{\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\mathbf{x})}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x})} - \frac{\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\mathbf{x}'_i)}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}'_i)} \right\|_F &\leq \left\| \frac{\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\mathbf{x})}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x})} - D_\Psi \mathbf{z} \right\|_F + \|D_\Psi(\mathbf{z} - \mathbf{s}_i)\|_F + \left\| D_\Psi \mathbf{s}_i - \frac{\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\mathbf{x}_i)}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}_i)} \right\|_F \\
&\overset{(iv)}{\leq} C_2 d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}) + C_3 \delta' + C_2 d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}'_i) \leq 2C_2 \varepsilon + C_3 \delta'
\end{aligned}$$

and therefore

$$\left\| \frac{\Psi\left(\mathbf{x}_{\mathfrak{S}}\right) - \Psi\left(\mathbf{x}\right)}{d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x})} \right\|_F \geq 1 - t/2 - (2C_2\varepsilon + C_3\delta') \tag{D.3}$$

To conclude, we set $\varepsilon \leq \min\left(\varepsilon_0, \frac{t}{8C_2}\right)$, $\delta' \leq \frac{t}{4C_3}$, $\delta \leq \frac{t\varepsilon^2/(4C_1)}{\varepsilon + M_{\mathfrak{S}}}$ to obtain the desired result.

Appendix E. Proof of Proposition 3

Concentration property. The concentration result is based on the fact that $\|\Psi\left(\mathbf{x}\right) - \Psi\left(\mathbf{x}\right)\|_2^2 \approx \mathbb{E}_{\boldsymbol{\omega}} \left|\varphi_{\boldsymbol{\omega}}(\mathbf{x} - \varphi_{\boldsymbol{\omega}}(\mathbf{x}')\right|^2 = d_E(\mathbf{x}, \mathbf{x}')^2$ by definition of the random features. Using simple function studies and a Bernstein concentration inequality with a control on moments of all orders, it is possible to show (see [18], eq. (160) then Prop. 6.11) that the concentration result (8) is valid with $c(t) \propto \frac{mt^2}{1+t}$ (we do not reproduce the detailed proof here for brevity).

Since this Berstein inequality is in fact valid for all vectors (not necessarily in the model), as a consequence the Boundedness Property is also satisfied with constant $\beta = 1 + t$, metric d_E and probability $1 - e^{-c(t)}$.

We now check hypotheses (i)–(iv) in Proposition 2. We are going to repeatedly use the fact that for $\mathbf{x}, \mathbf{x}' \in \mathfrak{S}$, since $\|\mathbf{x}\|_2 \leq M$, we have

$$\ell \left\|\mathbf{x} - \mathbf{x}'\right\|_2 \leq d_E(\mathbf{x}, \mathbf{x}') \leq L \left\|\mathbf{x} - \mathbf{x}'\right\|_2 \tag{E.1}$$

where $\ell := 2 \cdot \frac{1 - \exp\left(-M^2/(2\sigma^2)\right)}{M}$ and $L = \frac{M}{\sigma^2}$.

(i) Using a first-order Taylor expansion we have $\|\Psi\left(\mathbf{x}\right) - \Psi\left(\mathbf{x}'\right)\|_2 \leq \left(\sum_j \frac{\|\boldsymbol{\omega}_j\|_2}{\sqrt{m} f(\boldsymbol{\omega}_j)}\right) \|\mathbf{x} - \mathbf{x}'\|_2 \lesssim \sqrt{\gamma_2} \|\mathbf{x} - \mathbf{x}'\|_2$ hence by (E.1) hypothesis (i) is valid with $C_1 \propto \frac{\sqrt{\gamma_2}}{\ell}$.

(ii) It is immediate that $M_{\mathfrak{S}} = ML$. Then, using the well-known fact that for each S_j we have $\mathcal{N}\left(S_j \cap B_M, \|\cdot\|_2, \delta\right) = \mathcal{O}\left(\left(\frac{M}{\delta}\right)^s\right)$, by a union bound and (E.1) we have $\mathcal{N}\left(\mathfrak{S}, d_E, \delta\right) \leq \mathcal{O}\left(N\left(\frac{LM}{\delta}\right)^s\right)$.

(iii) Similar to the model, for all ε the normalized secant set is included in a union of subspace $\mathcal{S}_\varepsilon \subset \bigcup_{i,j=1}^N S_{ij} \cap B_{M/\ell}$, with norm controlled by M/ℓ by (E.1), and where $S_{ij} = S_i + S_j$ (sum of subspaces). Hence, since $\mathcal{N}\left(S_{ij} \cap B_{M/\ell}, \|\cdot\|_2, \delta\right) = \mathcal{O}\left(\left(\frac{M}{\ell\delta}\right)^{2s}\right)$, by a union bound and (E.1) we have $\mathcal{N}\left(\mathcal{S}_\varepsilon, d_E, \delta\right) \leq \mathcal{O}\left(N^2\left(\frac{LM}{\ell\delta}\right)^{2s}\right)$.

(iv) Finally, by a Taylor expansion we have

$$\varphi_{\boldsymbol{\omega}}(\mathbf{x}) - \varphi_{\boldsymbol{\omega}}(\mathbf{x}') = \frac{\mathrm{i}\boldsymbol{\omega}^\top}{f(\boldsymbol{\omega})}(\mathbf{x} - \mathbf{x}')e^{\mathrm{i}\boldsymbol{\omega}^\top \mathbf{x}} - \frac{1}{2} \cdot \frac{(\boldsymbol{\omega}^\top(\mathbf{x} - \mathbf{x}'))^2}{f(\boldsymbol{\omega})} e^{\mathrm{i}\boldsymbol{\omega}^\top \mathbf{x}''}$$

where $\mathbf{x}'' = \theta\mathbf{x} + (1-\theta)\mathbf{x}'$. Hence hypothesis (iv) is satisfied with $C_2 = C_1$ and $C_3 = \frac{\sqrt{\gamma_4}}{\ell^2}$, which concludes the proof.