Compressive Recovery of Sparse Precision Matrices

Titouan Vayer, Etienne Lasalle, Rémi Gribonval, Paulo Gonçalves

Compressive Recovery of Sparse Precision Matrices

Titouan Vayer*, Etienne Lasalle*, Rémi Gribonval*, Paulo Gonçalves*
*Univ Lyon, Inria, CNRS, ENS de Lyon, UCB Lyon 1,
LIP UMR 5668, F-69342, Lyon, France

Abstract

We consider the problem of learning a graph modeling the statistical relations of the d variables of a dataset with n samples $\mathbf{X} \in \mathbb{R}^{n\times d}$. Standard approaches amount to searching for a precision matrix $\mathbf{\Theta}$ representative of a Gaussian graphical model that adequately explains the data. However, most maximum likelihood-based estimators usually require storing the d^2 values of the empirical covariance matrix, which can become prohibitive in a high-dimensional setting. In this work, we adopt a "compressive" viewpoint and aim to estimate a sparse $\mathbf{\Theta}$ from a *sketch* of the data, *i.e.* a low-dimensional vector of size $m \ll d^2$ carefully designed from $\mathbf{X}$ using nonlinear random features. Under certain assumptions on the spectrum of $\mathbf{\Theta}$ (or its condition number), we show that it is possible to estimate it from a sketch of size $m = \Omega\left((d+2k)\log(d)\right)$ where k is the maximal number of edges of the underlying graph. These information-theoretic guarantees are inspired by compressed sensing theory and involve restricted isometry properties and instance optimal decoders. We investigate the possibility of achieving practical recovery with an iterative algorithm based on the graphical lasso, viewed as a specific denoiser. We compare our approach and graphical lasso on synthetic datasets, demonstrating its favorable performance even when the dataset is compressed.

I. Introduction

Inferring a complex network from data is used in many applications, such as in neuroscience for the treatment of epilepsy [1], for biological networks [2] or in genomics to identify gene interactions [3, 4]. We consider in this paper the problem of estimating a certain graph representing the statistical relations between d variables from n observed signals $\mathbf{x}_1, \cdots, \mathbf{x}_n$ where $\mathbf{x}_i \in \mathbb{R}^d$ follow a certain distribution μ, assumed to be centered for simplicity and with covariance matrix $\mathbf{\Sigma}$. This graph is often associated with the so-called *precision matrix* $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ which is sparse in many practical situations due to limited conditional dependencies [5]. However, simply inverting the empirical covariance matrix $\widehat{\mathbf{\Sigma}}$ usually does not yield a sparse estimation of the precision matrix estimation. The challenge is thus to infer a sparse precision matrix from $\widehat{\mathbf{\Sigma}}$. One of the most popular methods to do this graph estimation is probably the graphical lasso [5, 6]. Given the empirical covariance matrix

$$\widehat{\mathbf{\Sigma}} \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top, \tag{1}$$

the graphical lasso estimator searches for a certain matrix $\widehat{\mathbf{\Theta}}_{\mathrm{GL}} \in \mathbb{R}^{d\times d}$, representative of the graph edges, that solves the optimization problem

$$\widehat{\mathbf{\Theta}}_{\mathrm{GL}} \triangleq \underset{\mathbf{\Theta}\succ 0}{\arg\min} - \log\det(\mathbf{\Theta}) + \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Theta}\rangle + \lambda \|\mathbf{\Theta}\|_{1,\mathrm{off}}\,, \tag{2}$$

where the $\arg\min$ is taken over the set of symmetric positive definite matrices and $\|\mathbf{\Theta}\|_{1,\mathrm{off}} \triangleq \sum_{i<j}|\Theta_{ij}|$ promotes sparsity. An interpretation is as follows: for a Gaussian model $\mu = \mathcal{N}(0, \mathbf{\Sigma} = \mathbf{\Theta}^{-1})$, equation (2) corresponds to an ℓ_1-penalized maximum likehood estimator [7]. In other words, the graphical lasso estimates a sparse graph that best fits the data. More precisely, in the Gaussian case $\mu = \mathcal{N}(0, \mathbf{\Sigma} = \mathbf{\Theta}^{-1})$, the pattern of zeros of the precision matrix $\mathbf{\Theta}$ corresponds to conditional independencies among the variables, hence $\mathbf{\Theta}$ encodes the statistical relations between them[1] [8, 9]. Despite its many good properties, the graphical lasso suffers from a scaling problem with respect to the dimension. Indeed computing (2) requires to store in memory the entire empirical covariance matrix $\widehat{\mathbf{\Sigma}}$ and thus has a space complexity of $\mathcal{O}(d^2)$; this is problematic for applications where d is very large, such as high-resolution fMRI datasets or gene-microarray data where the number of genes d is typically around tens of thousands.

Graphical models estimation is a very active field of research and many alternatives to graphical lasso have been proposed in the literature. Numerous strategies aim to address the computational scalability of the optimization problem (2) through the use of approximation techniques, *e.g.* QUIC [10], BIG & QUIC [11], SQUIC [12]. Others are looking for estimators that have either better statistical guarantees or better algorithmic properties. We can mention some estimators that rely on non-convex penalties [13–16], ℓ_2 penalties [17] or other estimators like CLIME (and

[1] If $\Theta_{ij} = 0$ the variables i and j are conditionally independent, given the other variables

Dantzig-like estimators) [18, 19], RobustCLIME [20], D-trace [21] or Elem-GGM [22]. Finally, some approaches make assumptions on the underlying graph (*e.g.* that it is chordal) in order to find fast algorithms or try to constrain the structure of the graph sought in the estimation [23–25].

A. Compressive learning of graphical models

In this work we take another path: that of compression. Inspired by the theories of compressed sensing [26] and sketching [27], we try to estimate a sparse precision matrix $\mathbf{\Theta}$ associated to a covariance matrix $\mathbf{\Sigma} = \mathbf{\Theta}^{-1}$, not from the whole dataset, nor from the empirical covariance matrix $\widehat{\mathbf{\Sigma}}$, but from a *compressed version* of the data called *sketch*. More precisely, and given a well-chosen function $\Phi : \mathbb{R}^d \to \mathbb{R}^m$, we seek to estimate $\mathbf{\Theta}$ from the vector $\mathbf{s} \in \mathbb{R}^m$ defined by (see Figure 1)

$$\mathbf{s} \triangleq \frac{1}{n} \sum_{i=1}^{n} \Phi(\mathbf{x}_i) \in \mathbb{R}^m. \tag{3}$$

In sketching theory the Φ function is typically an adequately chosen *random function* and the reconstruction guarantees are usually based on ideas from compressed sensing [28]. The advantages of this type of approach are multiple: 1) when $m \ll nd$ the dataset is massively compressed, with benefits for storage and transfer 2) as averages, the sketches can be computed online, in one pass on the data, or in a distributed manner 3) as the data get aggregated, sketches have also good properties regarding differential privacy [29]. The main objective of this paper is to use the sketching approach and to answer the following question:

Objective. *Can we find Φ and a sketch dimension $m \ll d^2$ such that the sparse precision matrix $\mathbf{\Theta}$ can be well estimated from a sketch of the data $\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{x}_i) \in \mathbb{R}^m$?*

In other words, can we obtain an estimate of the graph structure using a small-sized vector compared to the size of the empirical covariance matrix? We will shortly furnish a precise definition of what we consider to be "well estimated". However, the underlying intuition guiding this inquiry is that, similar to the graphical lasso model, graphs estimated from data typically demonstrate sparsity. Therefore, there is hope that good approximations of these graphs can be obtained by accessing only a limited number of measurements compared to $\mathcal{O}(d^2)$, the total number of elements in the covariance matrix.

B. Contributions

The article presents the following key contributions:

1) We investigate, from an information-theoretic perspective, a sketching mechanism where we project against *i.i.d.* rank-one matrices. We show that when $m \gtrsim \|\mathbf{\Theta}\|_0 \log(d)$, the relevant information is preserved during the sketching phase[2]. More precisely, we prove that the corresponding sketching operator satisfies a *Restricted Isometry Property* (RIP), ensuring that the precision matrix can be reconstructed accurately using a so-called robust decoder.
2) This robust decoder being *a priori* not tractable, we propose a practical decoding scheme to estimate the sparse precision matrix from the sketch of the data. The algorithm relies on an iterative procedure that alternates between a gradient descent step associated to a sketch fidelity term and a "denoising step" performed by the graphical lasso.
3) For practical applications, we propose to deviate from the theoretical framework of 1) and leverage structured matrices (*e.g.,* Walsh-Hadamard matrices) to define a similar sketching mechanism but with boosted algorithmic efficiency. We demonstrate in the experiments that the combination of this sketching mechanism and our practical decoding scheme is able to significantly compress the data (*i.e.,* $m \ll d^2$) while still retaining the ability to recover the covariance and precision matrices.

C. Organization

The paper is organized as follows. In Section II, we explain how we sketch the data (*i.e.,* encoding phase) by using rank-one projections. Two variations are presented. The first one, for theory, consists in projecting against independent rank-one matrices. The second variation, for practical use, proposes to use random *structured* matrices for better memory and computational efficiency. Section III outlines the main information-theoretic results for the first variation. The main theorem (Theorem 3) proves that the corresponding sketching operator satisfies a certain RIP with high probability. For an easier read, the technical part required to prove this theorem is deferred to the end of the paper, Section VI, where we obtain the control of covering numbers and a concentration inequality for the sketching operator. In Section IV, we propose a practical algorithm to retrieve the precision matrix (*i.e.,* decoding phase) from the sketch. Section V presents the experiments on the encoding and decoding phases.

[2] $\|\mathbf{\Theta}\|_0$ denotes the number non non-zero coefficients of $\mathbf{\Theta}$.

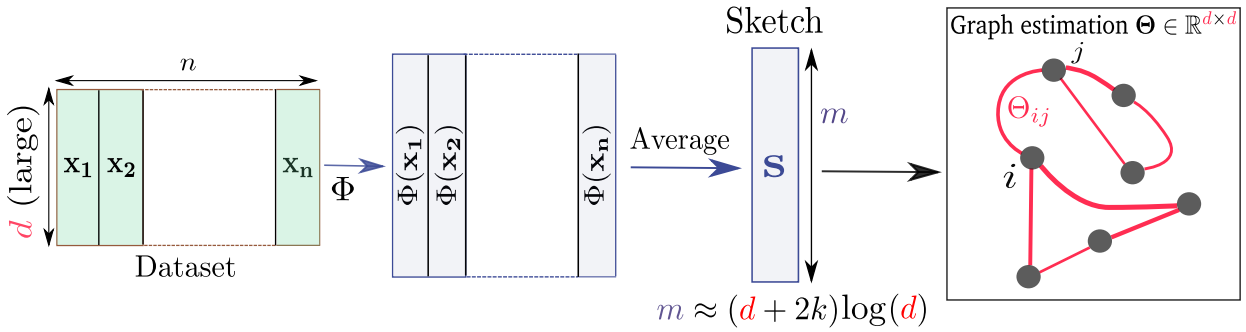

Fig. 1: Summary of our approach. The objective is to estimate a graph $\mathbf{\Theta} \in \mathbb{R}^{d\times d}$ between the d^2 variables (i,j) of the data set from a sketch of the data, $\mathbf{s} \triangleq \frac{1}{n}\sum_{i=1}^n \Phi(\mathbf{x}_i) \in \mathbb{R}^m$. We will show in the following sections that, under certain graph sparsity assumptions, it is theoretically possible to estimate $\mathbf{\Theta}$ using a sketch of size $m \approx (d+2k)\log(d) \ll d^2$ where k is the maximal number of edges of the underlying graph.

D. Notations and definitions.

For any integer m, $[\![m]\!] \triangleq \{1,\dots,m\}$. $S_d(\mathbb{R})$ is the linear space of symmetric matrices of size $d\times d$, while $S_d^{++}(\mathbb{R})$ (*resp.* $S_d^{+}(\mathbb{R})$) is the set of symmetric positive definite matrices (*resp.* positive semi-definite). We often write $\mathbf{A} \in S_d^{++}(\mathbb{R})$ as $\mathbf{A} \succ 0$ (*resp* $\mathbf{A} \succeq 0$ for $S_d^{+}(\mathbb{R})$). We denote by $\|\cdot\|_{2\to 2}$ the 2-operator norm or spectral norm for matrices *i.e.* $\|\mathbf{A}\|_{2\to 2} \triangleq \sup_{\|\mathbf{x}\|_2\leq 1}\|\mathbf{A}\mathbf{x}\|_2 = \boldsymbol{\sigma}_{\max}(\mathbf{A})$ where $\boldsymbol{\sigma}_{\max}(\mathbf{A})$ denotes the largest singular value of $\mathbf{A}$. The spectrum of $\mathbf{A} \in S_d(\mathbb{R})$ is denoted by $\mathrm{spec}(\mathbf{A})$. The standard euclidean scalar product on matrices is denoted by $\langle\cdot,\cdot\rangle$ and its associated norm (the Frobenius norm) is denoted by $\|\cdot\|_{\mathrm{Fro}}$. We denote by $\|\mathbf{A}\|_0$ the number of non-zeros components of $\mathbf{A}$, that is to say the ℓ_0 "norm" of $\mathbf{A}$. We also introduce the function $\mathrm{inv}(\mathbf{A}) \triangleq \mathbf{A}^{-1}$.

II. The sketching mechanism

The sketching procedure proposed in this article is based on *rank-one projections* and (well-chosen) *randomness*. Recall that we are given a sample of d-dimensional vectors $(\mathbf{x}_1,\dots,\mathbf{x}_n)$, and we want to define a function $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ such that the sketch is the average of the m-dimensional vectors $\Phi(\mathbf{x}_i)$ as in (3). Given a collection of d-dimensional random vectors $(\mathbf{a}_1,\dots\mathbf{a}_m)$, *not necessarily independent*, we consider in this work[3]

$$\Phi(\mathbf{x}) \triangleq \frac{1}{m}\left(|\mathbf{a}_1^\top\mathbf{x}|^2, |\mathbf{a}_2^\top\mathbf{x}|^2, \cdots, |\mathbf{a}_m^\top\mathbf{x}|^2\right)^\top. \tag{4}$$

Before proceeding, it is important to note that Φ can also be computed as $\Phi(\mathbf{x}) = \frac{1}{m}\{\mathbf{a}_j^\top\mathbf{x}\ \mathbf{x}^\top\mathbf{a}_j\}_{j\in[\![m]\!]} = \frac{1}{m}\{\langle\mathbf{a}_j\mathbf{a}_j^\top, \mathbf{x}\mathbf{x}^\top\rangle\}_{j\in[\![m]\!]}$. Therefore, by linearity of the inner product and equations (1) and (3), the induced sketch $\mathbf{s}$ is the vector whose coordinates are the projections of the empirical covariance matrix $\widehat{\mathbf{\Sigma}}$ against the rank-one matrices $\mathbf{a}_j\mathbf{a}_j^\top$: $\mathbf{s} = \frac{1}{m}\{\langle\mathbf{a}_j\mathbf{a}_j^\top, \widehat{\mathbf{\Sigma}}\rangle\}_{j\in[\![m]\!]}$.

This sketching mechanism appears in various works on compressed sensing, in particular for low-rank matrix completion [30–34], estimation of low-rank & semi-positive definite (SDP) matrices [35, 36], low-rank & sparse matrices [37] or in statistical learning for PCA [28] and in phase retrieval [38]. Here we propose to study it for the *recovery of sparse precision matrices*. The information-theoretic study will be conducted with independent vectors $\mathbf{a}_j$, while for practical experiments we consider a construction of $(\mathbf{a}_1,\dots\mathbf{a}_m)$ based on *structured matrices*, aiming to accelerate sketch computations and reduce storage requirements.

A. Independent case

The simplest case to consider is when the random vectors $\mathbf{a}_1,\dots,\mathbf{a}_m$ are independent and identically distributed according to some distribution Λ on $\mathbb{R}^d$. In Section III, we establish information-theoretic results demonstrating that, for some distributions Λ (standard multivariate distribution or uniform distribution on a sphere), and under appropriate sparsity assumptions on $\mathbf{\Theta}$, this sketching mechanism preserves enough information to retrieve the precision matrix.

Nevertheless, to retrieve the signal from the sketch, it is essential to have knowledge of the function Φ, which implies that the m d-dimensional vectors $(\mathbf{a}_j)_{j\in[\![m]\!]}$ must be stored as well. Hence, the overall memory cost comprises

[3]Readers familiar with sketching methods might be puzzled by the factor $1/m$ rather than the usual $1/\sqrt{m}$. In our case, as we will consider the ℓ_1-norm on $\mathbb{R}^m$, it is in fact the right scaling to obtain averages.

a reasonable $\mathcal{O}(m)$ expense to store the sketch and an additional $\mathcal{O}(md)$ cost to store the m d-dimensional vectors $(\mathbf{a}_j)_{j\in[\![m]\!]}$. Unfortunately, according to Theorem 3, theoretical recovery guarantees are obtained when the sketch is of size at least $m \gtrsim d\log(d)$, yielding a total memory cost that is larger than $\mathcal{O}(d^2)$, which is the cost of storing the empirical covariance matrix. In practice, we propose to use structured matrices leading to *dependent* random vectors $\mathbf{a}_1, \cdots, \mathbf{a}_m$ for the projections.

B. Structured matrices case

Using structured random matrices is a recurrent idea in compressive learning [39–41]. It reduces the degrees of freedom while mimicking the behavior of random matrices with *i.i.d.* columns. In the following, we assume that $d = 2^K$ is a power of 2 and that $m = B \times d$ is a multiple of d. Padding strategies can be implemented when these requirements are not met, but we leave this technicality out of the scope of this paper for the sake of simplicity and refer the interested reader to [29]. Here, we adopt the same approach as [40, 41]. We take $(\mathbf{a}_1, \dots \mathbf{a}_m)$ as the columns of the random matrix $\mathbf{A} = (\mathbf{B}_1|\dots|\mathbf{B}_B) \in \mathbb{R}^{d\times m}$ made of B independent structured blocks $\mathbf{B}_l \in \mathbb{R}^{d\times d}$ defined as triple-Rademacher matrices:

$$\mathbf{B}_l \triangleq \frac{1}{d^{3/2}}\mathbf{H}\mathbf{D}_l^{(1)}\mathbf{H}\mathbf{D}_l^{(2)}\mathbf{H}\mathbf{D}_l^{(3)}.$$

The matrix $\mathbf{H}$ denotes the Walsh-Hadamard matrix with entries in $\{\pm 1\}$ and $\mathbf{D}_l^{(k)}$ are independent random diagonal "sign-flipping" matrices, *i.e.,* random diagonal matrices with independent Rademacher entries. The scaling factors $d^{3/2}$ yields $\|\mathbf{a}_j\|_2 = 1$. We emphasize that this strategy results in the use of *dependent* random vectors $\mathbf{a}_1, \cdots, \mathbf{a}_m$.

The benefits of this approach are twofold: improved memory and computational efficiency. Indeed, notice that $\Phi(\mathbf{x})$ can be expressed as $\Phi(\mathbf{x}) = \{(\mathbf{A}^\top\mathbf{x})_j^2\}_{j\in[\![m]\!]} = (\mathbf{A}^\top\mathbf{x}) \odot (\mathbf{A}^\top\mathbf{x})$ where $\odot$ denotes the Hadamard product. Thus the memory and computational bottleneck *a priori* lies in the storage of $\mathbf{B}_l$ and the computations of $\mathbf{B}_l^\top\mathbf{x}$ for $l \in [\![B]\!]$. Fortunately, to reduce these costs, we can benefit from the properties of the fast Walsh-Hadamard transform: 1) it computes $\mathbf{H}^\top\mathbf{y}$ for any vector $\mathbf{y}$ without storing $\mathbf{H}$ as it is hard-coded into the algorithm, 2) based on a divide-and-conquer algorithm the matrix-vector multiplication $\mathbf{H}^\top\mathbf{y}$ requires only $\mathcal{O}(d\log d)$ operations. As a consequence for our sketching mechanism, only the diagonal matrices $\mathbf{D}_l^{(k)}$ must be stored. This reduces the space complexity of storing the dense matrix $\mathbf{A} \in \mathbb{R}^{d\times m}$ from $\mathcal{O}(md)$ to $\mathcal{O}(m)$ [40, 42]. Moreover, point 2) implies that $\Phi(\mathbf{x})$ can be computed with a time complexity of $\mathcal{O}(m\log d)$ as each of the B matrix-vector multiplications $\mathbf{B}_l^\top\mathbf{x}$ requires $\mathcal{O}(d\log d)$ operations with the fast Walsh-Hadamard transform. This compressive scheme will be applied in the experiments of Section V, along with a practical algorithm to recover the covariance from the sketch. The theoretical examination of recovery guarantees employing these structured sketching operators, especially the derivation of concentration results, is deferred for future research. Nevertheless, we anticipate that comparable guarantees to those in the independent case can be achieved.

Remark 1 (Dense Gaussian projections)**.** *We want to mention that another compressive scheme, maybe more natural for theoretical analysis, could have been considered. Instead of projecting against rank-one matrices, one could project against symmetric matrices with independent Gaussian entries via* $\Phi(\mathbf{x}) = \{\mathbf{x}^\top\mathbf{A}_j\mathbf{x}\}_{j\in[\![m]\!]}$ *where* $\mathbf{A}_j \in \mathbb{R}^{d\times d}$ *are dense Gaussian matrices. The theoretical analysis of this approach would be quite straightforward following classical compressed sensing works [26, 43]. However, this sketching mechanism would be far from efficient as the computation of* $\Phi(\mathbf{x})$ *would have an* $\mathcal{O}(md^2)$ *time and space complexity, which is even greater than directly storing* $\widehat{\mathbf{\Sigma}}$*. Nevertheless, it is worth noting that* optical computing*, which relies on optical processing units, could be leveraged to compute these random features in constant time in any dimension [44], albeit constrained by the current hardware capabilities.*

III. Recovery guarantees

In this section, we analyze the random function Φ defined in (4) in the *independent case*. We consider two probability distributions for the vectors $\mathbf{a}_j \sim \Lambda$. The first is $\Lambda = \Lambda_{\mathrm{G}} = \mathcal{N}(0, \mathbf{I}_d)$ the standard multivariate Gaussian distribution and the second $\Lambda = \Lambda_{\mathrm{U}} = \mathcal{U}(\sqrt{d}\mathbb{S}^{d-1})$ the uniform distribution on the hypersphere of radius $\sqrt{d}$. This choice of radius allows the $\mathbf{a}_j$ to have similar magnitudes under both distributions[4].

The entry point of our theoretical analysis is to notice that the random function Φ defined in (4) involves a *linear* operator on symmetric matrices [45]. Indeed, our objective can be reformulated as finding $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ from $\mathbf{s} = \mathcal{A}(\widehat{\mathbf{\Sigma}})$ where $\mathcal{A} : S_d(\mathbb{R}) \to \mathbb{R}^m$ is defined by

$$\mathcal{A}(\mathbf{M}) \triangleq \frac{1}{m}\left\{\langle \mathbf{a}_j\mathbf{a}_j^\top, \mathbf{M}\rangle\right\}_{j\in[\![m]\!]}, \ \mathbf{a}_j \overset{i.i.d}{\sim} \Lambda \in \{\Lambda_{\mathrm{G}}, \Lambda_{\mathrm{U}}\}. \tag{5}$$

We can therefore reformulate our problem as a compressed sensing problem: given a noisy linear measurement of the true covariance matrix $\mathbf{s} = \mathcal{A}(\mathbf{\Sigma}) + \mathbf{e} \in \mathbb{R}^m$, where $\mathbf{e} = \mathcal{A}(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})$ is the noise, we want to recover the precision

[4] $\mathbb{E}_{\Lambda_{\mathrm{G}}}\left[\|\mathbf{a}\|_2^2\right] = \mathbb{E}_{\Lambda_{\mathrm{U}}}\left[\|\mathbf{a}\|_2^2\right] = d.$

matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ that underlies certain sparsity assumptions. There are two difficulties here. First $m < d^2$, thus the problem is *a priori* ill-posed. Second, the sparsity assumption does not involve the "signal" $\mathbf{\Sigma}$ itself but a non-linear transformation of it, given by its inverse $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$.

A. Recipe for recovery guarantees

This work can be put in perspective with works on matrix completion (*e.g.* low-rank [28, 46], sparse + low-rank [37, 47–49] or low-rank SDP [35, 36]) where structural assumptions are satisfied *directly by the signal undergoing noisy linear measurements*. However, despite this non-linearity $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, the general idea to obtain guarantees is the same as in compressed sensing: we will assume that $\mathbf{\Theta}$ is in a "low dimensional" set, so that it is possible to estimate it from a very small number of measurements compared to the ambient dimension d^2 [50]. This hypothesis is formalized here by assuming that the true covariance matrix $\mathbf{\Sigma}$ belongs to a certain subset of $S_d^{++}(\mathbb{R})$, called *model set*. More precisely, and given $k \in \mathbb{N}, 0 < a \leq b$, we will study here the model set

$$\mathfrak{S}_{k,a,b} \triangleq \left\{ \mathbf{\Sigma} \in S_d^{++}(\mathbb{R}) : \mathbf{\Theta} = \mathbf{\Sigma}^{-1} \succ 0, \|\mathbf{\Theta}\|_0 \leq d + 2k, \mathrm{spec}(\mathbf{\Theta}) \subseteq [a, b] \right\} . \tag{6}$$

This set corresponds to covariance matrices associated to the $d + 2k$-sparse precision matrices whose spectra are well localized in $[a, b]$. This constraint on the spectrum of the precision matrix could be relaxed to a constraint on its condition number[5] as we will see later on. The number k represents the maximal number of non-zero components of $\mathbf{\Theta}$ on its upper triangular part[6]. In other words k is the maximal number of edges of the graph corresponding to $\mathbf{\Theta}$, without counting self-loops.

In practice a precision matrix $\mathbf{\Theta}$ such that $\mathbf{\Sigma} = \mathbf{\Theta}^{-1} \in \mathfrak{S}_{k,a,b}$ defines a graph with few connections and some conditions on its spectrum. Similar spectral conditions are usually assumed in the case of graphical lasso to derive sample complexities of the estimators [18, 51, 52]. The sparsity of the graph is also quite natural in many application where one wants to have a simple graph thus with few connections. We will see later that $\mathfrak{S}_{k,a,b}$ is the image by inv of a "low dimensional" set and, by using a certain notion of stability of the inverse, we will be able to recover $\mathbf{\Theta}$ from $\mathbf{s}$. We emphasize that while we have presented this specific model set, the majority of the results in this paper are general and can be extended to accommodate other model sets. Consequently, these results can apply to alternative assumptions concerning the underlying graph structure (*e.g.,* chordal graphs).

The central property for our guarantees is the *Restricted Isometric Property* (RIP) [43, 53], adapted to our context. In the following, we consider two general norms $\|\cdot\|_{S_d}$ and $\|\cdot\|_{\mathbb{R}^m}$ on S_d and $\mathbb{R}^m$, respectively.

Definition 1 (RIP)**.** *A linear operator $\mathcal{A} : (S_d(\mathbb{R}), \|\cdot\|_{S_d}) \to (\mathbb{R}^m, \|\cdot\|_{\mathbb{R}^m})$ satisfies the RIP_δ for some $\delta \in [0, 1[$ on a set $\mathfrak{S} \subset S_d^{++}(\mathbb{R})$ if for every $(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \in \mathfrak{S}^2$, it satisfies*

$$(1 - \delta)\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_{S_d} \leq \|\mathcal{A}(\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2)\|_{\mathbb{R}^m} \leq (1 + \delta)\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_{S_d} .$$

We can show that, when the linear operator $\mathcal{A}$ satisfies the RIP_δ, we can find a *decoder* that is robust to noise and that will allow us to recover $\mathbf{\Sigma}$ from $\mathbf{s}$. More precisely, following standard results in compressed sensing [26], we have the following result (the proof is given in Appendix A1 for completeness):

Theorem 1. *Let $\mathcal{A} : (S_d(\mathbb{R}), \|\cdot\|_{S_d}) \to (\mathbb{R}^m, \|\cdot\|_{\mathbb{R}^m})$ be a linear operator. Suppose that $\mathcal{A}$ satisfies RIP_δ on a model set $\mathfrak{S} \subset S_d(\mathbb{R})$ and consider the decoder $\Delta : \mathbb{R}^m \to \mathfrak{S}$ defined by*[7]

$$\Delta[\mathbf{y}] \in \underset{\mathbf{\Sigma} \in \mathfrak{S}}{\arg\min} \ \|\mathcal{A}(\mathbf{\Sigma}) - \mathbf{y}\|_{\mathbb{R}^m} . \tag{7}$$

Suppose that $\mathbf{x}_1, \cdots, \mathbf{x}_n \overset{i.i.d}{\sim} \mu$ where μ is a centered probability distribution with covariance $\mathbf{\Sigma} = \mathbb{E}_{\mathbf{x}\sim\mu}\left[\mathbf{x}\mathbf{x}^T\right] \succ 0$. Consider $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$ the empirical covariance matrix and $\mathbf{s} = \mathcal{A}(\widehat{\mathbf{\Sigma}})$. If $\mathbf{\Sigma} \in \mathfrak{S}$, then $\mathbf{\Sigma}^\star \triangleq \Delta[\mathbf{s}]$ satisfies

$$\|\mathbf{\Sigma}^\star - \mathbf{\Sigma}\|_{S_d} \leq \frac{2}{1 - \delta}\|\mathcal{A}(\widehat{\mathbf{\Sigma}}) - \mathcal{A}(\mathbf{\Sigma})\|_{\mathbb{R}^m} . \tag{8}$$

When $\mathbf{\Sigma} \notin \mathfrak{S}$, the bound (8) still holds up to an additional "modeling error" term $d^\circ(\mathbf{\Sigma}, \mathfrak{S})$ which quantifies the distance from $\mathbf{\Sigma}$ to $\mathfrak{S}$ (see Appendix A1).

The estimator given by (7) has many desirable properties: 1) it is robust to noise as shown in [28, 54] 2) it allows a recovery of $\mathbf{\Sigma}$ from $\mathbf{s}$ as the number of samples n grows. Indeed the term $\|\mathcal{A}(\widehat{\mathbf{\Sigma}}) - \mathcal{A}(\mathbf{\Sigma})\|_2 \leq |||\mathcal{A}|||\,\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2$, where

[5] The condition number is the ratio between the largest and smallest eigenvalues.

[6] As $\mathbf{\Theta} \succ 0$ it has necessarily d strictly positive coefficients on the main diagonal and thus $\|\mathbf{\Theta}\|_0 \geq d$.

[7] We always assume that the minimization problem (7) has at least one solution. The decoder can be adjusted as in [50], to handle the case where the argmin is only approximated to a certain accuracy.

$|||\mathcal{A}|||$ is the operator norm of $\mathcal{A}$, typically behaves as $\mathcal{O}(n^{-1/2})$ [55, Section 6]. Thus, solutions of the optimization problem (7) theoretically yield good approximations of $\mathbf{\Sigma}$. As an additional outcome, by leveraging the regularity of the inverse mapping for matrices with bounded spectra, this estimation of $\mathbf{\Sigma}$ also results in a reliable estimate of $\mathbf{\Theta}$. We refer to the discussion below Theorem 3 for a more precise statement.

With these remarks in mind, the strategy is now to obtain this RIP assumption, which enables the information-theoretic decoder and recovery guarantees. By the linearity of $\mathcal{A}$, we can observe that obtaining RIP_δ is equivalent to showing that, for $0 \leq \delta < 1$,

$$\sup_{\mathbf{U} \in S[\mathfrak{S}]} \left| \|\mathcal{A}(\mathbf{U})\|_{\mathbb{R}^m} - 1 \right| < \delta \,, \tag{9}$$

where we set

$$S[\mathfrak{S}] \triangleq \{ \frac{\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2}{\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_{S_d}} \ : (\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \in \mathfrak{S}^2, \|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_{S_d} > 0 \} \,. \tag{10}$$

This set is called the *normalized secant set* of $\mathfrak{S}$ [43]. We will provide a more detailed description of the space $S[\mathfrak{S}]$ later on. For now, we present the classical framework that will enable us to prove the RIP_δ property of $\mathcal{A}$.

The operator $\mathcal{A}$ being random we will show that, given a sufficient (but reasonable) dimension m, we can have a control of type (9) with high probability. In the following we denote by $\mathcal{N}(S[\mathfrak{S}], \|\cdot\|_{S_d}, \varepsilon)$ the *covering number* of $S[\mathfrak{S}]$ which, informally, quantifies the effective "dimension" of this set (see Section VI-A for more details). The following theorem (whose proof is deferred to Appendix A2) describes the main ingredients for establishing RIP_δ.

Theorem 2. *Consider a random sketching operator $\mathcal{A} : S_d(\mathbb{R}) \to \mathbb{R}^m$ and denote its operator norm by $|||\mathcal{A}||| \triangleq \sup_{\|\mathbf{U}\|_{S_d}=1} \|\mathcal{A}\mathbf{U}\|_{\mathbb{R}^m}$. Suppose that we are given two functions $C_1, C_2 : \mathbb{R}_+ \to \mathbb{R}_+$ such that*

$$\forall t > 0, \ \mathbb{P}\left(|||\mathcal{A}||| > t\right) \leq C_1(t) \,, \tag{11}$$

$$\forall \mathbf{U} \in S[\mathfrak{S}], \ \forall t > 0, \ \mathbb{P}\big(\left| \|\mathcal{A}(\mathbf{U})\|_{\mathbb{R}^m} - 1 \right| > t \big) \leq C_2(t) \,. \tag{12}$$

Then, for any $\varepsilon > 0$ and $\delta \in [0, 1[$,

$$\sup_{\mathbf{U} \in S[\mathfrak{S}]} \left| \|\mathcal{A}(\mathbf{U})\|_{\mathbb{R}^m} - 1 \right| < \delta \,, \tag{13}$$

with probability at least $1 - \mathcal{N}(S[\mathfrak{S}], \|\cdot\|_{S_d}, \varepsilon) C_2(\frac{\delta}{2}) - C_1(\frac{\delta}{2\varepsilon})$. Consequently with the same probability the operator $\mathcal{A}$ satisfies RIP_δ on $\mathfrak{S}$.

Remark 2. *The above theorem is presented in its most usual form, with both controls (11) and (12). However, (12) is sometimes easier to obtain (by leveraging classical concentration inequalities). Fortunately, we can obtain (11) from (12)[8], as for $\varepsilon' > 0$, defining C_1 by*

$$C_1(t) = \mathcal{N}(B_{S_d}, \|\cdot\|_{S_d}, \varepsilon') \ C_2((1-\varepsilon')t - 1) \quad \forall t > 1/(1-\varepsilon') \,, \tag{14}$$

where $\mathcal{N}(B_{S_d}, \|\cdot\|_{S_d}, \varepsilon')$ is the covering number of the unit sphere $B_{S_d} = \{\mathbf{U} \in S_d : \ \|\mathbf{U}\|_{S_d} = 1\}$, yields a valid upper-bound of $\mathbb{P}\left(|||\mathcal{A}||| > t\right)$. See Appendix A3 for the proof.

Guided by Theorem 2, to obtain the desired RIP_δ with high probability, we need control the covering number of $S[\mathfrak{S}_{k,a,b}]$ as well as the concentration of the random operator $\mathcal{A}$. The former will be done in Section VI-A, the latter in Section VI-B. To do so, specific norms will be chosen for $\|\cdot\|_{S_d}$ and $\|\cdot\|_{\mathbb{R}^m}$. For the former, we choose a norm that is *specific* to how our random operator $\mathcal{A}$ is drawn. Indeed, the probability distribution $\Lambda \in \{\Lambda_{\mathrm{G}}, \Lambda_{\mathrm{U}}\}$ induces a norm on $S_d(\mathbb{R})$ defined by

$$\forall \mathbf{M} \in S_d(\mathbb{R}), \ \|\mathbf{M}\|_\Lambda \triangleq \mathbb{E}_{\mathbf{a} \sim \Lambda} \left[\left| \langle \mathbf{a}\mathbf{a}^\top, \mathbf{M} \rangle \right| \right] = \mathbb{E}_{\mathbf{a} \sim \Lambda} \left[\left| \mathbf{a}^\top \mathbf{M} \mathbf{a} \right| \right] . \tag{15}$$

This choice is adapted to our problem when $\|\cdot\|_{\mathbb{R}^m} = \|\cdot\|_1$ since $\mathbb{E}_{\mathbf{a}_j \sim \Lambda} \left[\|\mathcal{A}(\mathbf{M})\|_1 \right] = \|\mathbf{M}\|_\Lambda$. Therefore, if for every $\mathbf{U} \in S[\mathfrak{S}]$, $\mathcal{A}(\mathbf{U})$ well concentrates around its expectation, then it has a high probability of being close to its expectation $\|\mathbf{U}\|_\Lambda = 1$, and thus (13) will hold.

From these choices for the norms and the approach suggested by Theorem 2 (*i.e.,* controlling the covering number and the concentration), we show that with a reasonable value of m the rank-one operators $\mathcal{A}$ satisfy the RIP on $\mathfrak{S}_{k,a,b}$. It is formalized in the theorem below which is the main theoretical result of the paper. We advise readers that are interested in the proof to refer to Section VI where the results on the covering numbers and the concentration of $\mathcal{A}$ are derived.

[8] At the expense of a restricted range for t that will be of no consequence for the rest of the paper.

Theorem 3. *Let $\mathcal{A} : (S_d(\mathbb{R}), \|\cdot\|_\Lambda) \to (\mathbb{R}^m, \|\cdot\|_1)$ be a rank-one projection operator as defined in (5), with $(\mathbf{a}_j)_j$ either Gaussian or uniform. For all $\delta, \rho \in]0,1[$, there exists $C = C(\delta, \rho, b/a)$, independent of m, k and d, such that, whenever*

$$m \geq C(d+2k)\log d\,, \tag{16}$$

the operator $\mathcal{A}$ satisfies RIP_δ on $\mathfrak{S}_{k,a,b}$ with probability at least $1-\rho$. In particular the following holds uniformly on $\mathbf{\Sigma} \in \mathfrak{S}_{k,a,b}$ with probability at least $1-\rho$: Let $\mathbf{x}_1, \cdots, \mathbf{x}_n \sim \mu$ with μ a centered probability distribution with covariance $\mathbf{\Sigma}$, $\widehat{\mathbf{\Sigma}}$ the empirical covariance matrix and $\mathbf{s} = \mathcal{A}(\widehat{\mathbf{\Sigma}})$ a sketch of the data. The estimator $\mathbf{\Sigma}^\star = \Delta[\mathbf{s}]$ defined in (7) satisfies

$$\|\mathbf{\Sigma}^\star - \mathbf{\Sigma}\|_\Lambda \leq \frac{2}{1-\delta}\|\mathcal{A}(\widehat{\mathbf{\Sigma}}) - \mathcal{A}(\mathbf{\Sigma})\|_1\,. \tag{17}$$

A more precise condition on m can be found in the proof of this theorem in Appendix C2 (in particular see (79)). This theorem shows that our sketching operators can satisfy the RIP with a sketching dimension $m \gtrsim (d+2k)\log d$, which is much smaller than the d^2 required to recover a matrix of size $d \times d$. Let us mention that although in (17) the error between the estimator $\mathbf{\Sigma}^\star$ and the true covariance matrix $\mathbf{\Sigma}$ is measured with the unusual Λ-norm, we can have the same type of control for $\|\mathbf{\Sigma}^\star - \mathbf{\Sigma}\|_{\mathrm{Fro}}$, at the expense of a multiplicative absolute constant $1/c_{\mathrm{Fro}} = \frac{9\sqrt{15}}{2}$ (see Proposition 4). In addition, (17) also provides guarantees for the recovery of the precision matrix $\mathbf{\Theta}$. By making use of the bounded spectra of $(\mathbf{\Sigma}^\star)^{-1}$ and $\mathbf{\Theta}$, we can exploit the regularity of the inverse map to obtain[9] $\|(\mathbf{\Sigma}^\star)^{-1} - \mathbf{\Theta}\|_{\mathrm{Fro}} \leq b^2\|\mathbf{\Sigma}^\star - \mathbf{\Sigma}\|_{\mathrm{Fro}}$. Therefore, whenever (17) is verified, we also have

$$\|(\mathbf{\Sigma}^\star)^{-1} - \mathbf{\Theta}\|_{\mathrm{Fro}} \leq \frac{9\sqrt{15}\;b^2}{(1-\delta)}\|\mathcal{A}(\widehat{\mathbf{\Sigma}}) - \mathcal{A}(\mathbf{\Sigma})\|_1\,. \tag{18}$$

Therefore, we obtain guarantees for the recovery of precision matrices from a sketch of the data based on *i.i.d.* rank-one projections.

Remark 3 (A similar result with an unbounded model set.)**.** *We want to point out that in (79), the condition on m depends on a and b only through the ratio b/a. This might suggest that the fundamental quantity to consider is the ratio between the largest and smallest eigenvalues, i.e., the condition number of the matrix, instead of the eigenvalues themselves.*

In Appendix F, we introduce a model set $\mathfrak{S}_{k,\kappa_0}$, where precision matrices have a condition number bounded by some $\kappa_0 \geq 1$. Even though $\mathfrak{S}_{k,\kappa_0}$ is much "bigger" than $\mathfrak{S}_{k,a,b}$, as the latter is bounded and the former is not, we obtain an upper-bound on the covering number of its normalized secant $S[\mathfrak{S}_{k,\kappa_0}]$ (see Theorem 5 in Appendix F). This is sufficient to derive a result similar to Theorem 3 with the same $m \gtrsim (d+2k)\log d$ condition, yielding guarantees on the recovery of $\mathbf{\Sigma}$ with only bounded condition numbers rather than prescribed spectra. However, recovery guarantees for $\mathbf{\Theta}$ as in (18) whould probably still requires spectral assumptions, as indicated by the fact that (18) does not onlyd depends on b/a.

Remark 4 (On the memory complexity)**.** *The previous theorem is valid* uniformly *on all covariance matrices $\mathbf{\Sigma} \in \mathfrak{S}_{k,a,b}$. Consequently, a single operator $\mathcal{A}$ allows us to compute the sketches of several datasets and to find the respective precision matrices. On the contrary, the graphical lasso requires to store the empirical covariance* for each dataset. *Consequently, if the goal is to recover N precision matrices, the presented approach only necessitates a cumulative memory expense of $\mathcal{O}(Nm)$ compared to $\mathcal{O}(Nd^2)$.*

B. Connection to prior works

Theorem 3 indicates that it is theoretically possible to recover $\mathbf{\Theta}$, with a $\mathcal{O}(n^{-1/2})$ error, by keeping in memory only a single sketch of size $m \gtrsim (d+2k)\log(d)$. To the best of our knowledge this is the first result for compressive recovery of precision matrices. It can however be put in perspective with [56] which tries to find the support of $\mathbf{\Theta}$ by observing, in an adaptive manner, only a small fraction of the entries of the *true* covariance $\mathbf{\Sigma}$. The authors show that it is possible to find $\mathbf{\Theta}$ from $\mathcal{O}(d\,\mathrm{polylog}(d))$ elements of $\mathbf{\Sigma}$ with some assumptions on the graph (small treewidth). Interestingly enough, we are able to obtain the same type of guarantees but with the big difference that, in our case, we observe a non-adaptive compressed version of the empirical covariance that takes the whole matrix into account, which is more in line with concrete applications.

IV. Towards practical recovery

In this section, we present a heuristic algorithmic approach to obtain the precision matrix from a sketch of the data. The previous section provides theoretical guarantees that indicate that the sketch contains the necessary information

[9] It results from the application of Lemma 1.

Algorithm 1 Algorithm for solving the decoding problem

Input: Sketching operator $\mathcal{A}$ and sketch of the data $\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{x}_i) = \mathcal{A}(\widehat{\boldsymbol{\Sigma}}) \in \mathbb{R}^m$.
Initial guess $\boldsymbol{\Sigma}_0 \succ 0$, regularization parameter $\lambda > 0$, step size $\gamma > 0$, and maximum number of iterations $t_{\max}$.
for $t \in 0, 1 \cdots, t_{\max} - 1$ **do**
    $\boldsymbol{\Sigma}_{t+\frac{1}{2}} \leftarrow \boldsymbol{\Sigma}_t - \gamma \mathcal{A}^\star(\mathcal{A}(\boldsymbol{\Sigma}_t) - \mathbf{s})$
    $\boldsymbol{\Sigma}_{t+1} \leftarrow \mathrm{GLASSO}_{\lambda\gamma}[\boldsymbol{\Sigma}_{t+\frac{1}{2}}]$ // with standard graphical lasso solver
end for
return estimated covariance $\boldsymbol{\Sigma}_{t_{\max}}$ and precision $\boldsymbol{\Sigma}_{t_{\max}}^{-1}$.

to recover the true covariance matrix, as long as its inverse (the precision matrix) is sparse. To recover this covariance, the decoder

$$\Delta[\mathbf{s}] \in \operatorname*{arg\,min}_{\boldsymbol{\Sigma} \in \mathfrak{S}_{k,a,b}} \|\mathcal{A}(\boldsymbol{\Sigma}) - \mathbf{s}\|_{\mathbb{R}^m}\,, \tag{19}$$

has many interesting recovery guarantees described in Theorem 3. However, solving the optimization problem (19) is difficult as the constraint $\boldsymbol{\Sigma} \in \mathfrak{S}_{k,a,b}$ implies to search among the matrices $\boldsymbol{\Sigma}$ such that $\|\boldsymbol{\Sigma}^{-1}\|_0 \leq (d+2k)$ which is a highly non-convex constraint. In this context, we will instead demonstrate that a computationally simpler alternative decoder works effectively in practice.

a) The graphical lasso as a denoiser: Our algorithmic solution is inspired by the connections between proximal operators and denoisers in the context of inverse problems. Numerous studies have indeed demonstrated that proximal operators can be regarded as efficient denoisers [57–60]. The core concept behind our approach thus revolves around utilizing a decoder that relies on the graphical lasso (2), which not only benefits from efficient algorithms but also carries the interpretation of a proximal operator [61]. Subsequently, we introduce

$$\mathrm{GLASSO}_\lambda[\mathbf{Z}] \stackrel{\Delta}{=} \operatorname*{arg\,min}_{\boldsymbol{\Sigma} \succ 0} \; -\log\det \boldsymbol{\Sigma}^{-1} + \langle \boldsymbol{\Sigma}^{-1}, \mathbf{Z}\rangle + \lambda \|\boldsymbol{\Sigma}^{-1}\|_{1,\mathrm{off}}\,, \tag{20}$$

where we recall that $\|\mathbf{M}\|_{1,\mathrm{off}} = \sum_{i<j} |M_{ij}|$. This operator is equivalent to the one introduced in (2), which computes the precision matrix instead of the covariance matrix. Indeed, although (20) is non-convex, a solution can be computed by solving the convex graphical lasso problem (2) with the change of variable $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. We also emphasize that most graphical lasso solvers such as [5, 6, 62] compute both the covariance $\boldsymbol{\Sigma}$ and the precision $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, without calculating the inverse, by relying instead on duality theory. We argue that the operator (20), when applied to the empirical covariance of the data $\widehat{\boldsymbol{\Sigma}}$, can be interpreted as a "denoiser" of $\widehat{\boldsymbol{\Sigma}}$ in the sense that it returns a covariance matrix whose inverse is sparse. This interpretation is motivated by the fact that the graphical lasso can be seen as a specific proximal operator in the framework of *Bregman divergences* [63, 64] which we now briefly describe (we refer to [65] for a more precise discussion). Let $\mathcal{H}$ be a Hilbert space and $h : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be proper, convex and differentiable on its open domain $\mathrm{dom}(h)$. The Bregman divergence associated to h is given by

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{H} \times \mathcal{H}, D_h(\mathbf{x}|\mathbf{y}) = \begin{cases} h(\mathbf{x}) - h(\mathbf{y}) - \langle \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \text{ if } \mathbf{y} \in \mathrm{dom}(h)\,, \\ +\infty \text{ otherwise.} \end{cases} \tag{21}$$

The function D_h measures a similarity or "distance" between the points in $\mathcal{H}$. For instance if $\mathcal{H} = \mathbb{R}^d$ and $h = \frac{1}{2}\|\cdot\|_2^2$ then D_h is simply $D_h(\mathbf{x}|\mathbf{y}) = \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$. For a function $\varphi : \mathcal{H} \to \mathbb{R}$, the so-called *(left) Bregman proximal operator* of φ is defined as $\mathrm{prox}_\varphi^h(\mathbf{z}) \stackrel{\Delta}{=} \arg\min_{\mathbf{x}}\; \varphi(\mathbf{x}) + D_h(\mathbf{x}|\mathbf{z})$ [65, Definition 2.3]. It generalizes the standard Euclidean proximal operator that can be computed with $h = \frac{1}{2}\|\cdot\|_2^2$. To relate with the context of the graphical lasso, we introduce the function $h(\mathbf{X}) = -\log\det \mathbf{X}$ if $\mathbf{X} \succ 0$ and $h(\mathbf{X}) = +\infty$ otherwise. It is strictly convex, continuously differentiable over its domain, and $\nabla h(\mathbf{X}) = -\mathbf{X}^{-1}$ [66, Appendix A.4.1]. The associated Bregman divergence writes (see *e.g.* [67, 68])

$$\forall \mathbf{X}, \mathbf{Y} \in S_d^{++}(\mathbb{R}),\;\; D_h(\mathbf{X}|\mathbf{Y}) \stackrel{\Delta}{=} -\log\det(\mathbf{X}) + \log\det \mathbf{Y} + \langle \mathbf{Y}^{-1}, \mathbf{X}\rangle - d\,. \tag{22}$$

Based on the definition of D_h we can rewrite the operator (20) as $\mathrm{GLASSO}_\lambda[\mathbf{Z}] = \operatorname*{arg\,min}_{\boldsymbol{\Sigma} \succ 0} \lambda\|\boldsymbol{\Sigma}^{-1}\|_{1,\mathrm{off}} + D_h(\mathbf{Z}|\boldsymbol{\Sigma})$. The graphical lasso can thus be interpreted as a proximal Bregman operator of the function $\varphi : \boldsymbol{\Sigma} \mapsto \lambda\|\boldsymbol{\Sigma}^{-1}\|_{1,\mathrm{off}}$ but by operating on the *right variable* of the Bregman divergence. Although not previously studied with the graphical lasso, this type of operators, known as *right proximity operator* [69–71], has nevertheless been considered in the context of Poisson inverse problems [72], for image restoration problems [73, Section 4] or in [74] where it was shown to admit a characterization in terms of gradient of a convex function. This discussion highlights that $\mathrm{GLASSO}_\lambda[\widehat{\boldsymbol{\Sigma}}]$ computes a covariance matrix candidate which is both close to $\widehat{\boldsymbol{\Sigma}}$ in the sense of D_h and whose inverse tends to be sparse.

b) Algorithm: Inspired by this interpretation of the graphical lasso as a denoiser, we present an iterative algorithm for estimating the true covariance from the sketch. In the following we consider the data fidelity term $f(\mathbf{\Sigma}) \triangleq \frac{1}{2}\|\mathcal{A}(\mathbf{\Sigma}) - \mathbf{s}\|_2^2$ whose gradient is $\nabla f(\mathbf{\Sigma}) = \mathcal{A}^\star(\mathcal{A}(\mathbf{\Sigma}) - \mathbf{s})$ where

$$\mathcal{A}^\star : \mathbf{y} \in \mathbb{R}^m \mapsto \sum_{j=1}^{m} y_i \mathbf{a}_j \mathbf{a}_j^\top \in S_d \tag{23}$$

is the adjoint operator of $\mathcal{A}$. Given a initial estimate $\mathbf{\Sigma}_0 \succ 0$ and a step-size $\gamma > 0$, our proposed algorithmic solution computes

$$\forall t \in \{0, \cdots, t_{\max}\},\ \mathbf{\Sigma}_{t+1} = \mathrm{GLASSO}_{\gamma\lambda}[\mathbf{\Sigma}_t - \gamma \nabla f(\mathbf{\Sigma}_t)]\,. \tag{24}$$

In other words, we alternate between a gradient step $\mathbf{\Sigma}_{t+\frac{1}{2}} = \mathbf{\Sigma}_t - \gamma \nabla f(\mathbf{\Sigma}_t)$ in the direction of minimizing f and a denoising/proximal Bregman step $\mathbf{\Sigma}_{k+1} = \mathrm{GLASSO}_{\gamma\lambda}[\mathbf{\Sigma}_{t+\frac{1}{2}}]$ (with parameter $\gamma\lambda$) in the vein of standard forward-backward algorithms. The overall procedure is summarized in Algorithm 1. The computational bottleneck of this algorithm is the graphical lasso step that has a computational $\mathcal{O}(d^3)$ complexity with standard graphical lasso solvers such as [5, 6, 62], which rely on the dual formulation of the graphical lasso, or [75] which rely on an iterative thresholding procedure. Although more efficient solvers exist [10–12] we choose to solve graphical lasso with the `scikit-learn` implementation [76]. It relies on the block coordinate procedure described in [62] and finds the solution of the graphical lasso with cubic complexity. It is noteworthy that the iterations of our algorithm, as described in (24), can be linked to Bregman Proximal Gradient (BPG) descent [72] with the Bregman divergence D_h. In Appendix D, we demonstrate that the iterations in (24) correspond to those of a BPG algorithm, albeit with a Riemannian gradient instead of the usual Euclidean gradient in (24). In practice, we observe that this algorithm always converges when $\gamma > 0$ is sufficiently small to ensure that the iterates $(\mathbf{\Sigma}_{t+\frac{1}{2}})_t$ remain positive definite (we give a safe step-size strategy ensuring this condition in Appendix E). The intriguing questions regarding the convergence rate of this algorithm and the minimizers associated with it are left for future research. We envision the use of guarantees inspired by Plug-and-Play literature, as described *e.g.* in [77]. However, we experimentally demonstrate in the next section that the associated estimator effectively recovers a precision matrix from a sketch of the data.

V. Experiments

In this section we provide experiments for assessing the efficiency of Algorithm 1. The following experiments are conducted using *structured sketching* as described in Section II-B. The objectives of these experiments is to answer the following questions:

(i) Does the decoder described in Algorithm 1 give qualitatively coherent results?
(ii) What is the impact of the sketch size m on the final estimation? More precisely, can we obtain a good estimate of the true precision matrix with a number of measurements m close to the theoretical $m_0 = C(d + 2k)\log(d)$ obtained in Theorem 3?
(iii) How does the sketching approach combined with the decoder described in Algorithm 1 compare to classical methods such as the graphical lasso in terms of performance?

In all experiments, the setting is as follows: we generate a true sparse precision matrix $\mathbf{\Theta} \in S_d^{++}(\mathbb{R})$ and we consider $\mathbf{x}_1, \cdots, \mathbf{x}_n \sim \mathcal{N}(0, \mathbf{\Theta}^{-1})$ *i.e.* n i.i.d. samples of a multivariate Gaussian with covariance $\mathbf{\Sigma} = \mathbf{\Theta}^{-1}$. In the experiments, we explore two methods for generating sparse precision matrices $\mathbf{\Theta}$. For each method, the matrix $\mathbf{\Theta}$ first consists of L blocks, each having a size of $M \times M$ with $L \times M = d$. The generation of each block follows the distribution of a random graph. In the first method, referred to as `Erdos`, each block follows the distribution of an Erdös-Rényi graph [78] where the probability of connection is set to $p = 0.2$. In the second method, referred to as `PowerLaw`, the random graph is a tree with a power-law degree distribution. Both methods utilize the `Networkx` library [79] for generating these distributions. After finding the support with the previous methods, the value of each coefficient is set as εu where $u \sim \mathrm{Unif}[1, 4]$ and $\varepsilon = 1$ with probability 0.5 and -1 with probability 0.5. The matrix $\mathbf{\Theta}$ is then symmetrized and made positive definite by adding a sufficiently large diagonal $(0.1 + \lambda_{\min})\mathbf{I}$. Finally a random permutation permutes the rows and columns of the matrix. Two examples of matrices $\mathbf{\Theta}$ generated according to these procedures are shown on the left side of Figure 2 (for $d = 64$).

In the experiments, we consider two performance measures. The first one is the relative error computed as

$$\mathrm{RE} \triangleq \frac{\|\mathbf{\Theta}_{\text{true}} - \mathbf{\Theta}_{\text{esti}}\|_{\text{Fro}}}{\|\mathbf{\Theta}_{\text{true}}\|_{\text{Fro}}} \tag{25}$$

and the second one is the $F_1 \in [0, 1]$ score between the true matrix and the estimated one. It is calculated as $F_1 = \frac{2\,\mathrm{tp}}{2\,\mathrm{tp} + \mathrm{fp} + \mathrm{fn}}$ where true positive tp stands for the case when there is an actual edge and the algorithm detects it; false positive fp stands for the case when there is no actual edge but the algorithm detects one, and false negative fn stands for the case when the algorithm failed to detect an actual edge.

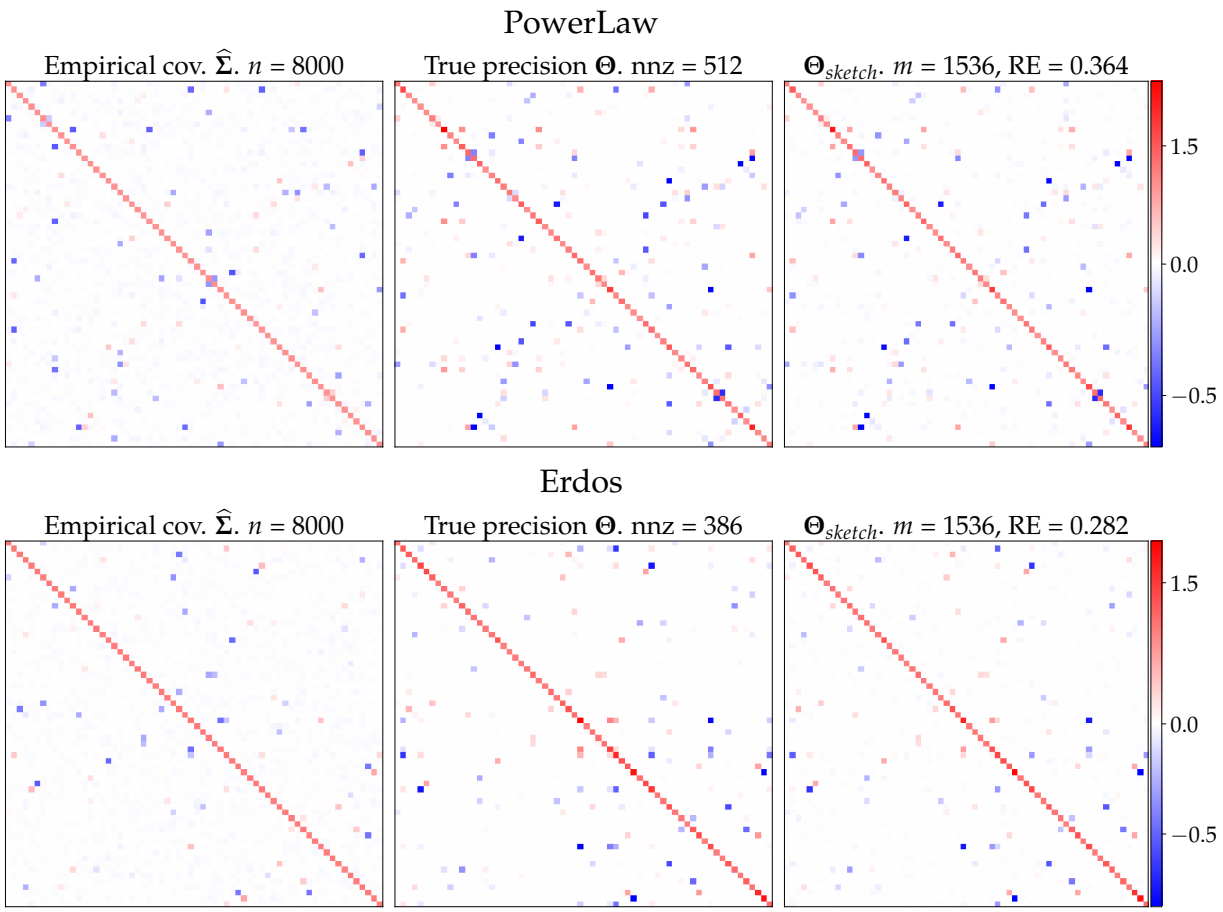

Fig. 2: Illustrative example of estimation using the decoding procedure described in Algorithm 1. The dimension is $d = 64$ and the sketch size $m = 1536$. We set $\lambda = 0.008$ for the ℓ_1 penalty parameter for all solvers. RE stands for the relative error between the true precision matrix and the estimated one (25). **(First row)** With a precision matrix $\mathbf{\Theta}$ generated from the `Erdos` process. **(Second row)** With a matrix $\mathbf{\Theta}$ generated from the `Powerlaw` process. The columns are (from left to right): the empirical covariance $\widehat{\mathbf{\Sigma}}$ for $n = 8000$, the true precision matrix $\mathbf{\Theta}$ and the decoder Algorithm 1 based on a sketch of the data with structured rank-one projections.

A. First illustration

First, we qualitatively illustrate the behavior of our Algorithm 1. In this experiment, we set $d = 64$ and generate $n = 8000$ samples from $\mathbf{\Theta}$ using the `Erdos` and `Powerlaw` procedures. The number of non-zero elements are respectively 512 for `PowerLaw` and 386 for `Erdos`. We consider one draw of structured sketching operator as described in Section II-B. We compute a sketch of the dataset $\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{x}_i) = \mathcal{A}(\widehat{\mathbf{\Sigma}}) \in \mathbb{R}^m$ and set the number of measurements for the sketch to $m = 1536$. In other words, the final sketch has a number of measurements that is approximately 75% the degrees of freedom of $d \times d$ symmetric matrices, which is $d(d+1)/2 = 2080$. We fix the number of iterations to $t_{\max} = 3500$, the step size to $\gamma = 0.005$ and $\lambda = 0.008$. The results are depicted in Figure 2. We can observe that in both cases, our method provides a visually consistent estimation of the precision matrix.

B. Impact of the number of measurements on the estimation - asymptotic regime

In order to quantitatively evaluate the impact of the number of measurements m on the final estimation we consider the asymptotic regime where $n = +\infty$ or equivalently we sketch the true covariance matrix as $\mathbf{s} = \mathcal{A}(\mathbf{\Sigma})$ instead of the empirical covariance matrix. We take $d = 256$ and we draw three precision matrices from the settings `PowerLaw` and `Erdos`. For each setting `PowerLaw` and `Erdos`, each $m \in \{256, 512, 1024, 2048, 4096, 8192, 16384, 32768 = d^2/2\}$, and each score function (RE and F_1-score) we choose the parameter $\lambda \in \{1\mathrm{e}-4, 5\mathrm{e}-4, 1\mathrm{e}-3, \cdots, 5\mathrm{e}-1\}$ that gives the best average score. We report the results in Figure 3.

We observe that for `Erdos` the estimator based on the sketching procedure and Algorithm 1 gives a recovery with a relative error that is below 10% starting from $m \approx 4096$ that is $\frac{m}{d^2/2} \approx 12\%$. This corresponds to a compression rate

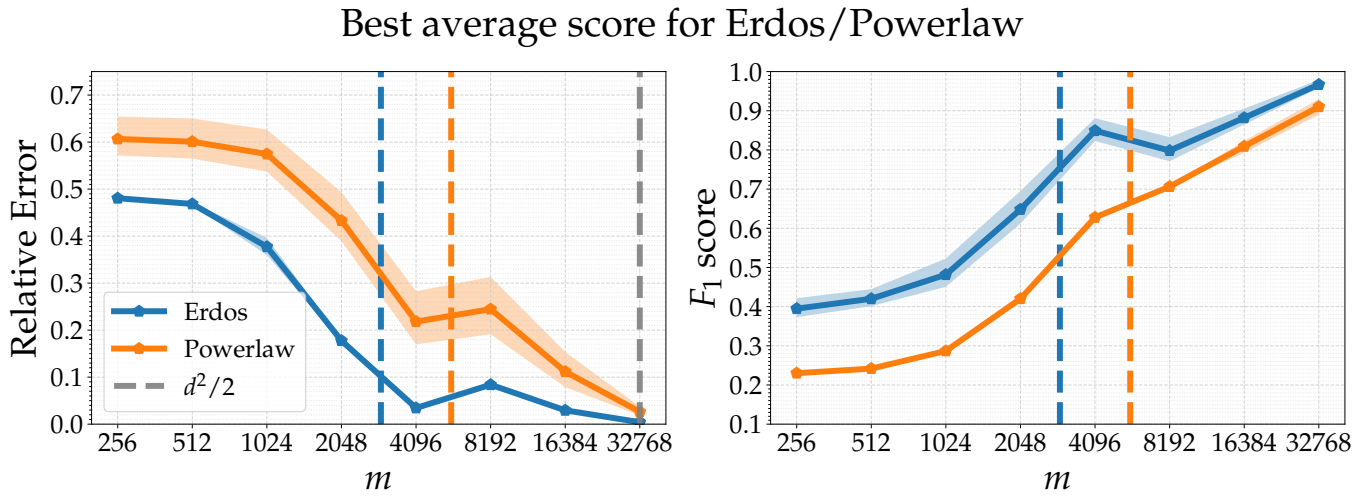

Fig. 3: Impact the of number of measurements on the final estimation for the datasets `Erdos` and `PowerLaw`. **(Left)** the best average relative relative error on the three draws of the true precision matrix for the `PowerLaw` and `Erdos` settings. **(Right)** the best average F_1 score. The 10-th percentile for these scores are reported in shaded line. Vertical colored dashed lines are located at $\hat{k}\log(d)$ where $\hat{k}$ is the average number of non-zero elements $\|\mathbf{\Theta}\|_0$ of the precision matrices for each setting.

of approximately 88%. This result is also consistent with the theoretical bound $m \approx \|\mathbf{\Theta}\|_0 \log(d)$ found in Theorem 3: we can see that the best average relative error is below 10% starting from this limit (the vertical dashed blue line). From $\frac{m}{d^2/2} = 50\%$, the recovery is also nearly perfect. The F_1 score is in agreement with the relative error: starting from $m \approx 4096$, the F_1 score is above 0.8, indicating that our estimator captures the correct statistical relationships between the variables.

For the `PowerLaw` setting, the results are slightly less favorable, and a relative error below 10% is only achieved for $\frac{m}{d^2/2} \approx 50\%$, resulting in a compression rate of approximately 50%. This can be explained by the fact that the precision matrices in this case are less sparse (vertical dashed orange line), and the theory also involves constants (*e.g.,* the constant $C(\delta, \rho, b/a)$ in Theorem 3) that can have a significant impact on the actual number of measurements required for reconstruction.

In all cases, this experiment indicates that the proposed sketching method preserves information, and the decoding method allows for a good estimation in an optimistic scenario with a large number of samples and a optimally-chosen regularization parameter.

C. Comparison with graphical lasso type estimators - finite sample regime

In a last experiment we compare quantitatively our approach with the graphical lasso estimator (computed with the `Scikit-learn` implementation [76]) and we investigate the influence of the sample size on the final estimation. We consider $d = 256$ and n samples $\mathbf{x}_1, \cdots, \mathbf{x}_n \sim \mathcal{N}(0, \mathbf{\Theta}^{-1})$ with three draws of precision matrices for each `Erdos` and `Powerlaw` settings. We use a sketch of the data for a number of samples n varying from 50 (in order to have the small sample regime) to 500000. We compute the sketch of the data for sketch size $m \in \{1024, 2048, 4096, 8192, 16384, 32768 = d^2/2\}$. As in the previous experiment, for each method, each n and each score function we pick the regularization parameter $\lambda \in \{1\mathrm{e}-4, 5\mathrm{e}-4, 1\mathrm{e}-3, \cdots, 5\mathrm{e}-1)\}$ that leads to the best average score.

We also consider a simple baseline, namely the pseudo-inverse of the empirical covariance matrix $(\widehat{\mathbf{\Sigma}})^\dagger$ as an estimate for $\mathbf{\Theta}_{\text{true}}$. Note that when $n > d$ the matrix $\widehat{\mathbf{\Sigma}}$ is almost surely invertible to that this baseline corresponds to the maximum likelihood estimator of $\mathbf{\Theta}_{\text{true}}$ without ℓ_1 regularization (same as graphical lasso with $\lambda = 0$ in this case). We report these results in Figure 4.

In both settings, the results show that our estimator improves as m and n increase, as expected. Moreover, for $m = 32768 = d^2/2$, the results of the graphical lasso and our estimator are nearly identical, demonstrating that our estimator is consistent with the graphical lasso when the number of measurements reaches the number of degrees of freedom of the empirical covariance matrix. Furthermore, we notice that for the `Erdos` setting, the performances are very similar to those of the graphical lasso even for $m \approx 8192$. This result indicates that even in the case of a finite number of samples, we are able to accurately estimate the precision matrices with a limited number of

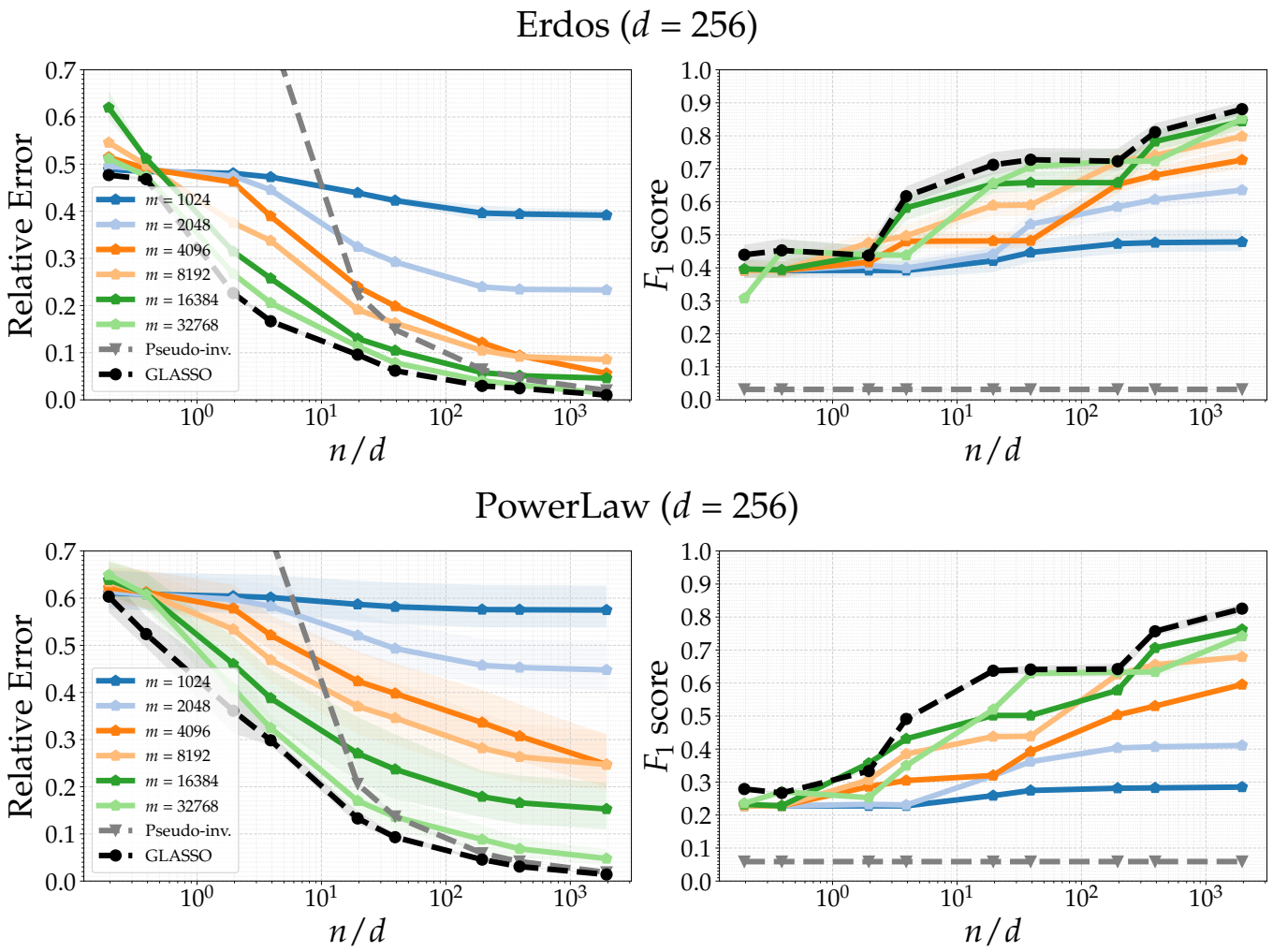

Fig. 4: Comparison between our estimator and graphical lasso type estimators with the `Erdos` **(top row)** and `PowerLaw` settings **(bottom row)**. The 10-th percentile for these scores are reported in shaded line.

measurements (8192 corresponds to a compression rate of $\approx 75\%$). For the `PowerLaw` setting, the results in terms of relative error are more mixed. However, the F_1 score remains comparable to that of the graphical lasso, indicating that our approach captures the correct statistical dependencies but may struggle with accurately estimating the intensities of these dependencies. In a rather reassuring way, in the small sample regime, our estimator outperforms the MLE estimator $(\widehat{\mathbf{\Sigma}})^\dagger$ in terms of relative error and consistently outperforms it in terms of the F_1 score (this is reasonable because $(\widehat{\mathbf{\Sigma}})^\dagger$ unlike $\mathbf{\Theta}_{\text{true}}$).

VI. Covering numbers bounds and concentration inequalities for Theorem 3

In this section, we provide the necessary ingredients to prove Theorem 3. Guided by Theorem 2, we start by giving general results to control covering numbers before applying them to the one considered in this article. Then, we provide the concentration inequality needed on the sketching operator $\mathcal{A}$.

A. Controlling covering numbers

We take the general point of view of normed vector spaces E, F with norms $\|\cdot\|_E, \|\cdot\|_F$ and a function $f : \Omega \subseteq E \to F$ defined on some domain $\Omega = \mathrm{dom}(f)$ of E. For a subset $\mathfrak{X} \subseteq E$ the covering number of $\mathfrak{X}$ *w.r.t.* $\|\cdot\|_E$ with radius $\varepsilon > 0$, denoted by $\mathcal{N}(\mathfrak{X}, \|\cdot\|_E, \varepsilon)$, is the minimal number of closed balls of radius ε (*w.r.t.* $\|\cdot\|_E$) required to entirely cover $\mathfrak{X}$ and whose centers are in $\mathfrak{X}$ (see Appendix B1 for a formal definition). The core of our reasoning is based on the following sets, which generalize definition (10).

Definition 2 (Normalized secant sets)**.** *We define the normalized secant set of some* $\mathfrak{X} \subseteq E$ *as*

$$S[\mathfrak{X}] \triangleq \{\frac{x-y}{\|x-y\|_E} : \ (x,y) \in \mathfrak{X}^2, \|x-y\|_E > 0\}\,.$$

We introduce also the normalized secant set of $\mathfrak{X} \subseteq \Omega = \mathrm{dom}(f)$ *embedded by* f *as*

$$S[f(\mathfrak{X})] \triangleq \{\frac{f(x)-f(y)}{\|f(x)-f(y)\|_F} : \ (x,y) \in \mathfrak{X}^2, \|f(x)-f(y)\|_F > 0\}\,.$$

The main contribution of this section is the control of the covering number of a normalized secant $S[f(\mathfrak{X})] \subseteq F$ by the covering number of $\mathfrak{X}$ and $S[\mathfrak{X}]$, for a sufficiently smooth f. The intuition is the following: if we consider a signal $x \in \mathfrak{X}$ in a "low-dimensional" space, then its image by f is also "low-dimensional" if f is sufficiently regular on $\mathfrak{X}$.

Remark 5 (Link with the precision matrix estimation problem)**.** *This section is presented in a general case, but its results will ultimately be applied to control the covering number of the normalized secant of our model set* $\mathfrak{S}_{k,a,b}$ *defined in* (6)*. Therefore, one should keep in mind that our final application framework will consider* $E = F = S_d$*,* $f = \mathrm{inv}$ *with* Ω *the set of symmetric and invertible matrices and* $\mathfrak{X} = \mathfrak{S}^{-1}_{k,a,b} = \{\mathbf{\Theta} = \mathbf{\Sigma}^{-1} : \ \mathbf{\Sigma} \in \mathfrak{S}_{k,a,b}\}$.

To carry out the analysis of $S[f(\mathfrak{X})]$ we introduce the notions of *long* and *short chords*. These objects are inspired by results in compressive independent component analysis and the theory of random projections on manifolds [80].

a) Long and short chords: The set $S[f(\mathfrak{X})]$ can be divided in two subsets that are more analytically tractable. First, we introduce, for $\eta > 0$, the following sets

$$\mathrm{C}^+_\eta(\mathfrak{X}, f) \triangleq \{(x,y) \in \mathfrak{X}^2 : \|f(x)-f(y)\|_F > \eta\}\,,$$
$$\mathrm{C}^-_\eta(\mathfrak{X}, f) \triangleq \{(x,y) \in \mathfrak{X}^2 : 0 < \|f(x)-f(y)\|_F \leq \eta\}\,.$$

With these notations, $S[f(\mathfrak{X})]$ can be decomposed into long and short chords $S[f(\mathfrak{X})] = S^+_\eta[f(\mathfrak{X})] \cup S^-_\eta[f(\mathfrak{X})]$, where the long and short chords are respectively defined by

$$\begin{aligned} S^+_\eta[f(\mathfrak{X})] &\triangleq \{\frac{f(x)-f(y)}{\|f(x)-f(y)\|_F} : \ (x,y) \in \mathrm{C}^+_\eta(\mathfrak{X}, f)\}\,, \\ S^-_\eta[f(\mathfrak{X})] &\triangleq \{\frac{f(x)-f(y)}{\|f(x)-f(y)\|_F} : \ (x,y) \in \mathrm{C}^-_\eta(\mathfrak{X}, f)\}\,. \end{aligned} \tag{26}$$

In order to control the covering number of $S[f(\mathfrak{X})]$ we will exploit its decomposition via the spaces $S^\pm_\eta[f(\mathfrak{X})]$ whose covering numbers are easier to obtain, assuming some regularity of the function f on $\mathfrak{X}$.

Definition 3 (Bi-Lipschitz assumption)**.** *Given positive constants* α *and* β*, we say that a function* f *is* (α, β)*-bi-Lipschitz if for all* $(x,y) \in \mathfrak{X}^2$ *it verifies*

$$\alpha\|x-y\|_E \leq \|f(x)-f(y)\|_F \tag{27}$$
$$\|f(x)-f(y)\|_F \leq \beta\|x-y\|_E\,. \tag{28}$$

In order to define the notion of differential we assume that the model set is contained in the interior of $\Omega = \mathrm{dom}(f)$, *i.e.,* $\mathfrak{X} \subseteq \mathrm{int}(\Omega)$. In particular when f is differentiable on $\mathfrak{X}$, (28) implies that $\sup_{x \in \mathfrak{X}} \|\mathrm{D}f_x\|_{\mathrm{op}} \leq \beta < +\infty$, where $\|\cdot\|_{\mathrm{op}}$ is the operator norm defined as $\|\mathrm{D}f_x\|_{\mathrm{op}} \triangleq \sup_{\|h\|_E \leq 1} \|\mathrm{D}f_x(h)\|_F$.

We further consider the following regularity assumptions:

Assumption A-1. *(Second order approximation)* f *is differentiable on* $\mathrm{int}(\Omega)$ *and there exists* $L > 0$ *such that*

$$\forall (x,y) \in \mathfrak{X}^2, \ \|f(x)-f(y)-\mathrm{D}f_y(x-y)\|_F \leq L\|x-y\|^2_E\,.$$

Assumption A-2. *(Bounded curvature)* f *is differentiable on* $\mathrm{int}(\Omega)$ *and* ζ*-smooth on* $\mathfrak{X}$ *i.e. there exists* $\zeta > 0$ *such that*

$$\forall (x,y) \in \mathfrak{X}^2, \ \|\mathrm{D}f_x - \mathrm{D}f_y\|_{\mathrm{op}} \leq \zeta\|x-y\|_E\,.$$

As described in Lemma 4 (see Appendix B2) the condition **A-2** implies **A-1** when $\mathfrak{X}$ is a *convex* model set.

b) Covering numbers of $S^+_\eta[f(\mathfrak{X})]$*:* controlling the covering with long chords is relatively easy under some assumptions on f as proven in the following proposition (proof in Appendix B3).

Proposition 1. *Assume that* f *is* β*-Lipschitz as in* (28)*. Then for any* $\eta > 0$ *and* $\varepsilon > 0$*,*

$$\mathcal{N}(S^+_\eta[f(\mathfrak{X})], \|\cdot\|_F, \varepsilon) \leq \mathcal{N}(\mathfrak{X}, \|\cdot\|_E, \frac{\eta}{16\beta}\varepsilon)^2\,. \tag{29}$$

In other words, the "dimension" of the long chords can be controlled by that of the model set itself, assuming only that f is Lipschitz-continuous.

c) Covering numbers of $S_\eta^-[f(\mathfrak{X})]$: controlling the covering number of short chords is a little more delicate and generally requires a fine analysis of $\mathfrak{X}$ [81]. However, by adding the hypothesis **A-1** and **A-2** we are able to prove the following result (proof in Appendix B4).

Proposition 2. *Assume that f is (α,β)-bi-Lipschitz and satisfies assumptions **A-1** and **A-2**. Then for any $\varepsilon>0$ and $\eta>0$,*

$$\mathcal{N}(S_\eta^-[f(\mathfrak{X})],\|\cdot\|_F,2(\varepsilon+\frac{L}{\alpha^2}\eta)) \leq \left(\frac{\zeta+\beta\alpha}{\alpha^2\varepsilon}\right)\times\mathcal{N}(\mathfrak{X},\|\cdot\|_E,\frac{\alpha\varepsilon}{\zeta+\beta\alpha})\times\mathcal{N}(S[\mathfrak{X}],\|\cdot\|_E,\frac{\alpha^2\varepsilon}{2(\zeta+\beta\alpha)}). \tag{30}$$

Intuition of the proof. The idea is to show that each element of $S_\eta^-[f(\mathfrak{X})]$ can be described by a "tangent" vector to $S[\mathfrak{X}]$ and that the set of tangent vectors has a covering number which can be controlled by those of $\mathfrak{X}$ and $S[\mathfrak{X}]$. □

By combining the two propositions we are now ready to state the main theorem of this section:

Theorem 4. *Assume that f is (α,β)-bi-Lipschitz and satisfies assumptions **A-1** and **A-2** Then for all $\eta>0,\varepsilon>0$, we have*

$$\mathcal{N}(S[f(\mathfrak{X})],\|\cdot\|_F,2\left[\varepsilon+\frac{L}{\alpha^2}\eta\right]) \leq \mathcal{N}(\mathfrak{X},\|\cdot\|_E,\frac{\eta}{8\beta}\left[\varepsilon+\frac{L}{\alpha^2}\eta\right])^2+\frac{\zeta+\beta\alpha}{\alpha^2\varepsilon}\mathcal{N}(\mathfrak{X},\|\cdot\|_E,\frac{\alpha\varepsilon}{\zeta+\beta\alpha})\times\mathcal{N}(S[\mathfrak{X}],\|\cdot\|_E,\frac{\alpha^2\varepsilon}{2(\zeta+\beta\alpha)}).$$

Proof. We use that for any $\eta>0$ $S[\mathfrak{X}]=S_\eta^+[f(\mathfrak{X})]\cup S_\eta^-[f(\mathfrak{X})]$ thus the covering number of $S[f(\mathfrak{X})]$ is less than the sum of the coverings of $S_\eta^+[f(\mathfrak{X})]$ and $S_\eta^-[f(\mathfrak{X})]$. Then we apply Proposition 1 and Proposition 2. □

d) Putting everything together in the case of sparse precision matrix: we apply the previous results to our framework, that is $E=F=S_d$, $f=\mathrm{inv}$ and $\mathfrak{X}=\mathfrak{S}_{k,a,b}^{-1}=\{\mathbf{\Theta}=\mathbf{\Sigma}^{-1}:\ \mathbf{\Sigma}\in\mathfrak{S}_{k,a,b}\}$. We set the norms as follow : $\|\cdot\|_E=\|\cdot\|_{\mathrm{Fro}}$ the Frobenius norm and $\|\cdot\|_F=\|\cdot\|_\Lambda$ as defined in (15). We have the following lemma which shows that $f=\mathrm{inv}$ satisfies all the necessary regularity assumptions (the proof can be found in Appendix B5).

Lemma 1. *Assume that there exist constants c_{Fro} and C_{Fro} such that $c_{\mathrm{Fro}}\|\mathbf{M}\|_{\mathrm{Fro}}\leq\|\mathbf{M}\|_\Lambda\leq C_{\mathrm{Fro}}\|\mathbf{M}\|_{\mathrm{Fro}}$. Then the function $f=\mathrm{inv}$ is (α,β)-bi-Lipschitz and satisfies assumptions **A-1** and **A-2** on $\mathfrak{S}_{k,a,b}^{-1}$ with $\alpha=\frac{c_{\mathrm{Fro}}}{b^2}$, $\beta=\frac{C_{\mathrm{Fro}}}{a^2}$, $L=\frac{C_{\mathrm{Fro}}}{a^3}$, $\zeta=2L$.*

Note that expressions for the constants c_{Fro} and C_{Fro} will be provided in Proposition 4. In order to control the covering number of $S[\mathfrak{S}_{k,a,b}]=S[\mathrm{inv}(\mathfrak{S}_{k,a,b}^{-1})]$ we only have to check those of $\mathfrak{S}_{k,a,b}^{-1}$ and $S[\mathfrak{S}_{k,a,b}^{-1}]$. It is done in the following lemma (the proof can be found in Appendix B6).

Lemma 2. *For any $\varepsilon>0$ we have*

$$\mathcal{N}(\mathfrak{S}_{k,a,b}^{-1},\|\cdot\|_{\mathrm{Fro}},\varepsilon)\leq(\frac{ed^2}{2k})^k(\frac{18\sqrt{d}\times b}{\varepsilon})^{d+k}\ \text{and}\ \mathcal{N}(S[\mathfrak{S}_{k,a,b}^{-1}],\|\cdot\|_{\mathrm{Fro}},\varepsilon)\leq(\frac{ed^2}{4k})^{2k}(\frac{18}{\varepsilon})^{d+2k}. \tag{31}$$

This result show that the *box counting dimensions* [82] (also called entropy dimensions) of $\mathfrak{S}_{k,a,b}^{-1}$ and $S[\mathfrak{S}_{k,a,b}^{-1}]$ are smaller than $(d+k)\log(d)$ and $(d+2k)\log(d)$ and thus much smaller than d^2, which will allow us to have the guarantees presented in the introduction.

Corollary 1. *Assuming the existence of the constant $C_{\mathrm{Fro}},c_{\mathrm{Fro}}$ as in Lemma 1, there exist absolute constants $c_0,c_1\geq1$, such that, for any $\varepsilon>0$, the covering number of $S[\mathfrak{S}_{k,a,b}]$ verifies*

$$\mathcal{N}\left(S[\mathfrak{S}_{k,a,b}],\|\cdot\|_\Lambda,\varepsilon\right)\leq\left(\frac{ed^2}{2k}\right)^{4k}\left[\left(c_0\frac{\sqrt{d}C_{\mathrm{Fro}}^2}{\varepsilon^2c_{\mathrm{Fro}}^2}\frac{b^5}{a^5}\right)^{2(d+k)}+\left(c_1\frac{\sqrt{d}C_{\mathrm{Fro}}^2}{\varepsilon^2c_{\mathrm{Fro}}}(\frac{2}{c_{\mathrm{Fro}}}+1)\frac{b^5}{a^5}\right)^{d+2k+1}\right].$$

This is a direct consequence of Theorem 4, combined with Lemma 1 and 2. See Appendix B7 for the proof. This corollary allows for a control of the covering number that is required to prove a RIP for rank-one projections.

B. Application to rank-one projections

The results of the previous section allowed us to control one of the three quantities of interest for establishing the RIP_δ : the covering number. We are left with studying the concentrations the $\mathcal{A}$ operator (Theorem 2).

A natural choice for the norm $\|\cdot\|_{\mathbb{R}^m}$ would be the standard euclidean norm $\|\cdot\|_2$. However, in order to hope for satisfying the RIP, one would have to choose $\|\mathbf{M}\|_{S_d}=\left(\mathbb{E}_{\mathbf{A}\sim\Lambda}\left[|\langle\mathbf{M}_j,\mathbf{M}\rangle|^2\right]\right)^{1/2}$. Unfortunately, with these choices, we were unable to find sufficiently tight concentration functions C_1,C_2 that lead to better guaranties than $m\gtrsim d^2$. This phenomenom had already been observed, as written in [31] the rank-one projections lead to loose RIP constants for low-rank matrix completion problems (unless $m\gtrsim d^2$), and we envision that similar results hold in our case. The remedy found in [31] is to consider instead ℓ_1 RIP (*i.e.* taking the ℓ_1 norm for $\|\cdot\|_{\mathbb{R}^m}$ in Definition 1 instead of the

ℓ_2 norm), leading to the choice of $\|\cdot\|_\Lambda$ defined in (15) for $\|\cdot\|_{S_d}$. This next part shows that this remedy helps for the recovery of covariance matrices and thus of precision matrices.

Recall that we consider here the operators $\mathcal{A}(\mathbf{\Sigma}) := \frac{1}{m}\left(\mathbf{a}_1^\top \mathbf{\Sigma}\mathbf{a}_1, \dots, \mathbf{a}_m^\top \mathbf{\Sigma}\mathbf{a}_m\right)^\top$, where the vectors $\mathbf{a}_i$ either follow $\mathcal{N}(0, \mathbf{I}_d)$ in the Gaussian case or $\mathcal{U}(\sqrt{d}\mathbb{S}^{d-1})$ in the uniform case, and their associated norms $\|\mathbf{\Sigma}\|_\Lambda = \mathbb{E}[|\mathbf{a}^\top \mathbf{\Sigma}\mathbf{a}|]$. In this setting, the following concentration result holds.

Proposition 3. *For any $\mathbf{U} \in S_d(\mathbb{R})$ satisfying $\|\mathbf{U}\|_\Lambda = 1$ and for any $t > 0$, we have*

$$\mathbb{P}\left(\left|\|\mathcal{A}(\mathbf{U})\|_1 - 1\right| > t\right) \leq 2\exp\left(-\frac{m}{8e^2}\min\left(\frac{t^2}{K_\Lambda^2}, \frac{t}{K_\Lambda}\right)\right), \tag{32}$$

where K_Λ is an absolute constant given by $K_\Lambda = \frac{76e^2\sqrt{15}}{\log 2}$ for the Gaussian case and $K_\Lambda = \frac{304e^3\sqrt{15}}{\log 2}$ for the uniform case.

Intuition of the proof. The proof is based on a Bernstein-type concentration inequality for sums of sub-exponential variables. The essential ingredient of the proof is thus to show that the centered random variable $|\mathbf{a}^\top \mathbf{U}\mathbf{a}| - 1$ involved in $\|\mathcal{A}\mathbf{U}\|_1 - 1$ is subexponential with a subexponential norm bounded by an absolute constant K_Λ. The full proof can be found in Appendix C1. □

This result provides the function C_2 required to obtain a RIP with Theorem 2. This is the only result that would need to be adapted if one wants to provide information-theoretic guarantees to the sketching operator defined from random structured matrices. Indeed, notice that all previous results are still valid in the structured case.

Finally to be able to control the covering number, the norm $\|\cdot\|_\Lambda$ needs to verify the hypothesis of Corollary 1. This is done in the following proposition (see Appendix C1 for the proof).

Proposition 4. *Let $\Lambda \in \{\Lambda_{\mathrm{G}}, \Lambda_{\mathrm{U}}\}$ be either the Gaussian or uniform distribution on $\mathbb{R}^d$ and consider the associated Λ-norm defined in (15). Then,*

$$\forall \mathbf{M} \in S_d(\mathbb{R}),\ \frac{2}{9\sqrt{15}}\|\mathbf{M}\|_{\mathrm{Fro}} \leq \|\mathbf{M}\|_\Lambda \leq \sqrt{d}\|\mathbf{M}\|_{\mathrm{Fro}}\,.$$

This gives $c_{\mathrm{Fro}} = 2/(9\sqrt{15})$ and $C_{\mathrm{Fro}} = \sqrt{d}$ in Lemma 1.

In this section, we have established control over the covering numbers via Corollary 1 and introduced the concentration inequality for the sketching operator through Proposition 3. These components provide the necessary foundation for proving Theorem 3. The comprehensive proof of this theorem can be found in Appendix C2.

VII. Conclusion & perspectives

In this work, we have presented a compressive approach based on sketching to estimate sparse precision matrices. We have shown that it is possible to estimate a $(d+2k)$-sparse precision matrix from a data sketch of the order $(d+2k)\log(d)$, which is significantly smaller than the typical memory complexity of d^2 associated with the graphical lasso. Our analysis is supported by information-theoretic guarantees, where we have established restricted isometries and instance optimality properties. Finally, we have proposed a practical algorithmic solution for computing an estimation of the precision matrix from the sketch.

Our work opens several new lines of research. Given the generality of the tools presented in this paper, it would be interesting to explore whether similar guarantees can be established for other model sets based on specific graph structures [24]. Also, as practitioners are not always interested in the graph associated with the precision matrix *per se* but rather in some of its properties (*e.g.,* a group structure among the nodes), it would be interesting to see how these properties can directly be inferred using a compressive learning approach and whether it can help further reduce the sketch's dimension.

From an algorithmic point of view, our work also raises several questions. The proposed estimator is costly as it requires to solve several graphical lasso. An interesting further work would be to design a more efficient decoder that is sufficiently close to the optimal decoder given by the theory. In this context, the choice proposed in this paper is a first step toward practical recovery, but other algorithms could be used based on different precision matrix estimators like [22]. Finally, from an application point of view, an interesting perspective would be to use the sketching approach to learn, in an online way, a dynamic graph, in the way of the time varying graphical lasso [83].

Acknowledgements

We gratefully acknowledge the support of the Centre Blaise Pascal's IT test platform at ENS de Lyon (Lyon, France) for Machine Learning facilities. The platform operates the SIDUS solution [84]. All the figures are generated using `Matplotlib` [85] and we also used `Scipy` and `Numpy` [86, 87] for computing our estimator. The authors are grateful to Mathurin Massias for insightful discussions.

REFERENCES

[1] G. Frusque, "Inférence et décomposition modale de réseaux dynamiques en neurosciences," Ph.D. dissertation, Université de Lyon, Dec. 2020.
[2] J. Ye and J. Liu, "Sparse methods for biomedical data," *SIGKDD Explor. Newsl.*, vol. 14, no. 1, p. 4–15, dec 2012.
[3] C. Lingjærde, T. G. Lien, Ø. Borgan, H. Bergholtz, and I. K. Glad, "Tailored graphical lasso for data integration in gene network reconstruction," *BMC bioinformatics*, vol. 22, no. 1, pp. 1–22, 2021.
[4] Y. Zuo, Y. Cui, G. Yu, R. Li, and H. W. Ressom, "Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical lasso," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–14, 2017.
[5] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, Jul. 2008.
[6] O. Banerjee, L. E. Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data," *Journal of Machine Learning Research*, vol. 9, no. 15, pp. 485–516, 2008.
[7] M. Yuan and Y. Lin, "Model selection and estimation in the gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.
[8] A. P. Dempster, "Covariance selection," *Biometrics*, pp. 157–175, 1972.
[9] S. L. Lauritzen, *Graphical models*. Clarendon Press, 1996, vol. 17.
[10] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar, "Quic: Quadratic approximation for sparse inverse covariance estimation," *Journal of Machine Learning Research*, vol. 15, no. 83, pp. 2911–2947, 2014.
[11] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, P. K. Ravikumar, and R. Poldrack, "Big & quic: Sparse inverse covariance estimation for a million variables," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013.
[12] M. Bollhöfer, A. Eftekhari, S. Scheidegger, and O. Schenk, "Large-scale sparse inverse covariance matrix estimation," *SIAM Journal on Scientific Computing*, vol. 41, no. 1, 2019.
[13] P. Xu, J. Ma, and Q. Gu, "Speeding up latent variable gaussian graphical model estimation via nonconvex optimization," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[14] Q. Sun, K. M. Tan, H. Liu, and T. Zhang, "Graphical nonconvex optimization via an adaptive convex relaxation," in *International Conference on Machine Learning*, vol. 80, 2018, pp. 4810–4817.
[15] K. Sagar, S. Banerjee, J. Datta, and A. Bhadra, "Precision matrix estimation under the horseshoe-like prior-penalty dual," 2021.
[16] J. Fan, Y. Feng, and Y. Wu, "Network exploration via the adaptive LASSO and SCAD penalties," *The Annals of Applied Statistics*, vol. 3, no. 2, pp. 521 – 541, 2009.
[17] M. O. Kuismin, J. T. Kemppainen, and M. J. Sillanpää, "Precision matrix estimation with rope," *Journal of Computational and Graphical Statistics*, vol. 26, no. 3, pp. 682–694, 2017.
[18] T. Cai, W. Liu, and X. Luo, "A constrained l1 minimization approach to sparse precision matrix estimation," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 594–607, 2011.
[19] M. Yuan, "High dimensional inverse covariance matrix estimation via linear programming," *Journal of Machine Learning Research*, vol. 11, no. 79, pp. 2261–2286, 2010.
[20] L. Wang and Q. Gu, "Robust Gaussian graphical model estimation with arbitrary corruption," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 3617–3626.
[21] T. Zhang and H. Zou, "Sparse precision matrix estimation via lasso penalized D-trace loss," *Biometrika*, vol. 101, no. 1, pp. 103–120, 02 2014.
[22] E. Yang, A. C. Lozano, and P. K. Ravikumar, "Elementary estimators for graphical models," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.
[23] J. Ying, J. de Miranda Cardoso, and D. Palomar, "Nonconvex sparse graph learning under laplacian constrained graphical model," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 7101–7113.
[24] S. Kumar, J. Ying, J. V. de M. Cardoso, and D. P. Palomar, "A unified framework for structured graph learning via spectral constraints," *Journal of Machine Learning Research*, vol. 21, no. 22, pp. 1–60, 2020.
[25] H. E. Egilmez, E. Pavez, and A. Ortega, "Graph learning from data under laplacian and structural constraints," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 6, pp. 825–841, 2017.
[26] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, 2013.
[27] R. Gribonval, A. Chatalic, N. Keriven, V. Schellekens, L. Jacques, and P. Schniter, "Sketching Data Sets for Large-Scale Learning: Keeping only what you need," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 12–36, Sep. 2021.
[28] R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin, "Compressive Statistical Learning with Random Feature Moments," *Mathematical Statistics and Learning*, vol. 3, 2021.
[29] A. Chatalic, "Efficient and privacy-preserving compressive learning," Theses, Université Rennes 1, Nov. 2020.
[30] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Commun. ACM*, vol. 55, no. 6, p. 111–119, jun 2012.
[31] T. T. Cai and A. Zhang, "Rop: Matrix recovery via rank-one projections," *The Annals of Statistics*, vol. 43, no. 1, Feb 2015.
[32] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 4034–4059, 2015.
[33] M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege, "Stable low-rank matrix recovery via null space properties," *Information and Inference: A Journal of the IMA*, vol. 5, no. 4, pp. 405–441, 2016.
[34] R. Kueng, H. Rauhut, and U. Terstiege, "Low rank matrix recovery from rank one measurements," *Applied and Computational Harmonic Analysis*, vol. 42, no. 1, pp. 88–116, 2017.
[35] M. Slawski, P. Li, and M. Hein, "Regularization-free estimation in trace regression with symmetric positive semidefinite matrices," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015.
[36] Y. Li, Y. Sun, and Y. Chi, "Low-rank positive semidefinite matrix recovery from corrupted rank-one measurements," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 397–408, 2017.
[37] S. Bahmani and J. Romberg, "Sketching for simultaneously sparse and low-rank covariance matrices," in *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2015, pp. 357–360.
[38] K. Jaganathan, Y. C. Eldar, and B. Hassibi, "Phase retrieval: An overview of recent developments," *Optical Compressive Imaging*, pp. 279–312, 2016.
[39] Q. Le, T. Sarlós, A. Smola *et al.*, "Fastfood-approximating kernel expansions in loglinear time," in *Proceedings of the international conference on machine learning*, vol. 85, 2013, p. 8.
[40] F. X. X. Yu, A. T. Suresh, K. M. Choromanski, D. N. Holtmann-Rice, and S. Kumar, "Orthogonal random features," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016.
[41] A. Chatalic, R. Gribonval, and N. Keriven, "Large-Scale High-Dimensional Clustering with Fast Sketching," in *ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing*. Calgary, Canada: IEEE, Apr. 2018, pp. 4714–4718.

[42] Fino and Algazi, "Unified matrix treatment of the fast walsh-hadamard transform," *IEEE Transactions on Computers*, vol. C-25, no. 11, pp. 1142–1146, 1976.

[43] G. Puy, M. E. Davies, and R. Gribonval, "Recipes for stable linear embeddings from Hilbert spaces to $\mathbb{R}^m$," *IEEE Transactions on Information Theory*, 2017, submitted in 2015.

[44] A. Saade, F. Caltagirone, I. Carron, L. Daudet, A. Drémeau, S. Gigan, and F. Krzakala, "Random projections through multiple optical scattering: Approximating kernels at the speed of light," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6215–6219.

[45] R. Delogne, V. Schellekens, and L. Jacques, "Rop inception: signal estimation with quadratic random sketching," in *30th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2022.

[46] Y.-k. Liu, "Universal low-rank matrix recovery from pauli measurements," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., vol. 24. Curran Associates, Inc., 2011.

[47] E. Richard, P.-A. Savalle, and N. Vayatis, "Estimation of simultaneously sparse and low rank matrices," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ser. ICML'12. Madison, WI, USA: Omnipress, 2012, p. 51–58.

[48] S. Foucart, R. Gribonval, L. Jacques, and H. Rauhut, "Jointly low-rank and bisparse recovery: Questions and partial answers," 2019.

[49] J. Tanner and S. Vary, "Compressed sensing of low-rank plus sparse matrices," *ArXiv*, vol. abs/2007.09457, 2020.

[50] A. Bourrier, M. E. Davies, T. Peleg, P. Pérez, and R. Gribonval, "Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems," *IEEE Transactions on Information Theory*, pp. 7928–7946, Dec. 2014.

[51] L. Wang, X. Ren, and Q. Gu, "Precision matrix estimation in high dimensional gaussian graphical models with faster rates," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Gretton and C. C. Robert, Eds., vol. 51. Cadiz, Spain: PMLR, 09–11 May 2016, pp. 177–185.

[52] W. Liu and X. Luo, "Fast and adaptive sparse precision matrix estimation in high dimensions," *Journal of Multivariate Analysis*, vol. 135, pp. 153–162, 2015.

[53] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[54] N. Keriven and R. Gribonval, "Instance optimal decoding and the restricted isometry property," in *Journal of Physics: Conference Series*, vol. 1131, no. 1. IOP Publishing, 2018, p. 012002.

[55] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[56] G. Lugosi, J. Truszkowski, V. Velona, and P. Zwiernik, "Learning partial correlation graphs and graphical models by covariance queries," *Journal of Machine Learning Research*, vol. 22, no. 203, pp. 1–41, 2021.

[57] S. Hurault, A. Leclaire, and N. Papadakis, "Proximal denoiser for convergent plug-and-play optimization with nonconvex regularization," in *International Conference on Machine Learning (ICML)*. PMLR, 2022, pp. 9483–9505.

[58] S. Hurault, U. Kamilov, A. Leclaire, and N. Papadakis, "Convergent bregman plug-and-play image restoration for poisson inverse problems," *arXiv preprint arXiv:2306.03466*, 2023.

[59] R. Cohen, Y. Blau, D. Freedman, and E. Rivlin, "It has potential: Gradient-driven denoisers for convergent solutions to inverse problems," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 18 152–18 164, 2021.

[60] R. Cohen, M. Elad, and P. Milanfar, "Regularization by denoising via fixed-point projection (red-pro)," *SIAM Journal on Imaging Sciences*, vol. 14, no. 3, pp. 1374–1406, 2021.

[61] A. Beck, *First-Order Methods in Optimization*. Philadelphia, PA, USA: SIAM-Society for Industrial and Applied Mathematics, 2017.

[62] R. Mazumder and T. Hastie, "The graphical lasso: New insights and alternatives," *Electronic Journal of Statistics*, vol. 6, no. none, pp. 2125 – 2149, 2012.

[63] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR computational mathematics and mathematical physics*, vol. 7, no. 3, pp. 200–217, 1967.

[64] Y. Censor and S. A. Zenios, "Proximal minimization algorithm with d-functions," *Journal of Optimization Theory and Applications*, vol. 73, no. 3, pp. 451–464, 1992.

[65] M. Teboulle, "A simplified view of first order methods for optimization," *Mathematical Programming*, vol. 170, no. 1, pp. 67–96, 2018.

[66] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, March 2004.

[67] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935 – 980, 2011.

[68] J.-C. Pesquet, "Proximal approaches for matrix optimization problems: Application to robust precision matrix estimation," *Signal Processing*, vol. 169, p. 107417, 2020.

[69] H. H. Bauschke, P. L. Combettes, and D. Noll, "Joint minimization with alternating bregman proximity operators," *Pacific Journal of Optimization*, 2006.

[70] H. H. Bauschke, M. N. Dao, and S. B. Lindstrom, "Regularizing with bregman–moreau envelopes," *SIAM Journal on Optimization*, vol. 28, no. 4, pp. 3208–3228, 2018.

[71] T. T.-K. Lau and H. Liu, "Bregman proximal langevin monte carlo via bregman-moreau envelopes," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 049–12 077.

[72] H. H. Bauschke, J. Bolte, and M. Teboulle, "A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications," *Mathematics of Operations Research*, vol. 42, no. 2, pp. 330–348, 2017.

[73] H. Woo, "A characterization of the domain of beta-divergence and its connection to bregman variational model," *Entropy*, vol. 19, no. 9, p. 482, 2017.

[74] R. Gribonval and M. Nikolova, "A characterization of proximity operators," *Journal of Mathematical Imaging and Vision*, vol. 62, no. 6-7, pp. 773–789, 2020.

[75] B. Rolfs, B. Rajaratnam, D. Guillot, I. Wong, and A. Maleki, "Iterative thresholding algorithm for sparse inverse covariance estimation," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[76] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[77] E. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin, "Plug-and-play methods provably converge with properly trained denoisers," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5546–5557.

[78] P. Erdös and A. Rényi, "On random graphs i," *Publicationes Mathematicae Debrecen*, vol. 6, 1959.

[79] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," Los Alamos National Lab.(LANL), Los Alamos, NM (United States), Tech. Rep., 2008.

[80] K. L. Clarkson, "Tighter bounds for random projections of manifolds," in *Proceedings of the Twenty-Fourth Annual Symposium on Computational Geometry*, ser. SCG '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 39–48.

[81] R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin, "Statistical Learning Guarantees for Compressive Clustering and Compressive Mixture Modeling," *Mathematical Statistics and Learning*, vol. 3, 2021.

[82] J. C. Robinson, *Dimensions, Embeddings, and Attractors*, ser. Cambridge Tracts in Mathematics. Cambridge University Press, 2010.

[83] D. Hallac, Y. Park, S. Boyd, and J. Leskovec, "Network inference via the time-varying graphical lasso," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 205–213.
[84] E. Quemener and M. Corvellec, "Sidus—the solution for extreme deduplication of an operating system," *Linux Journal*, vol. 2013, no. 235, p. 3, 2013.
[85] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.
[86] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.
[87] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, sep 2020.
[88] R. Coleman, *Calculus on Normed Vector Spaces*, ser. Universitext. Springer New York, 2012.
[89] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.
[90] ——, *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press, 2012, p. 210–268.
[91] A. Han, B. Mishra, P. K. Jawanpuria, and J. Gao, "On riemannian optimization over positive definite matrices with the bures-wasserstein geometry," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 8940–8953, 2021.
[92] N. Boumal, *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

APPENDIX

A. Proofs of Section III-A

1) Proof of Theorem 1:

Theorem 1. *Let $\mathcal{A} : (S_d(\mathbb{R}), \|\cdot\|_{S_d}) \to (\mathbb{R}^m, \|\cdot\|_{\mathbb{R}^m})$ be a linear operator. Suppose that $\mathcal{A}$ satisfies RIP_δ on a model set $\mathfrak{S} \subset S_d(\mathbb{R})$ and consider the decoder $\Delta : \mathbb{R}^m \to \mathfrak{S}$ defined by*[10]

$$\Delta[\mathbf{y}] \in \operatorname*{arg\,min}_{\boldsymbol{\Sigma}\in\mathfrak{S}} \|\mathcal{A}(\boldsymbol{\Sigma}) - \mathbf{y}\|_{\mathbb{R}^m} . \tag{7}$$

Suppose that $\mathbf{x}_1, \cdots, \mathbf{x}_n \overset{i.i.d}{\sim} \mu$ where μ is a centered probability distribution with covariance $\boldsymbol{\Sigma} = \mathbb{E}_{\mathbf{x}\sim\mu}\left[\mathbf{x}\mathbf{x}^T\right] \succ 0$. Consider $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$ the empirical covariance matrix and $\mathbf{s} = \mathcal{A}(\widehat{\boldsymbol{\Sigma}})$. If $\boldsymbol{\Sigma} \in \mathfrak{S}$, then $\boldsymbol{\Sigma}^\star \triangleq \Delta[\mathbf{s}]$ satisfies

$$\|\boldsymbol{\Sigma}^\star - \boldsymbol{\Sigma}\|_{S_d} \leq \frac{2}{1-\delta}\|\mathcal{A}(\widehat{\boldsymbol{\Sigma}}) - \mathcal{A}(\boldsymbol{\Sigma})\|_{\mathbb{R}^m} . \tag{8}$$

When $\boldsymbol{\Sigma} \notin \mathfrak{S}$, the bound (8) still holds up to an additional "modeling error" term $d^\circ(\boldsymbol{\Sigma}, \mathfrak{S})$ which quantifies the distance from $\boldsymbol{\Sigma}$ to $\mathfrak{S}$ (see Appendix A1).

Proof. With the notations of the theorem $\boldsymbol{\Sigma}^* = \Delta[\mathcal{A}(\widehat{\boldsymbol{\Sigma}})] \in \arg\min_{\boldsymbol{\Sigma}\in\mathfrak{S}} \|\mathcal{A}(\boldsymbol{\Sigma}) - \mathcal{A}(\widehat{\boldsymbol{\Sigma}})\|_{\mathbb{R}^m}$. Then for any $\boldsymbol{\Sigma}_\mathfrak{S} \in \mathfrak{S}$:

$$\begin{aligned}
\|\boldsymbol{\Sigma}^* - \boldsymbol{\Sigma}\|_{S_d} &\leq \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_\mathfrak{S}\|_{S_d} + \|\boldsymbol{\Sigma}_\mathfrak{S} - \boldsymbol{\Sigma}^*\|_{S_d} \\
&\leq \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_\mathfrak{S}\|_{S_d} + \frac{1}{1-\delta}\|\mathcal{A}(\boldsymbol{\Sigma}^*) - \mathcal{A}(\boldsymbol{\Sigma}_\mathfrak{S})\|_{\mathbb{R}^m} \\
&\leq \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_\mathfrak{S}\|_{S_d} + \frac{1}{1-\delta}\left(\|\mathcal{A}(\boldsymbol{\Sigma}^*) - \mathcal{A}(\widehat{\boldsymbol{\Sigma}})\|_{\mathbb{R}^m} + \|\mathcal{A}(\widehat{\boldsymbol{\Sigma}}) - \mathcal{A}(\boldsymbol{\Sigma}_\mathfrak{S})\|_{\mathbb{R}^d}\right) \\
&\leq \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_\mathfrak{S}\|_{S_d} + \frac{2}{1-\delta}\|\mathcal{A}(\widehat{\boldsymbol{\Sigma}}) - \mathcal{A}(\boldsymbol{\Sigma}_\mathfrak{S})\|_{\mathbb{R}^m} \\
&\leq \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_\mathfrak{S}\|_{S_d} + \frac{2}{1-\delta}\|\mathcal{A}(\widehat{\boldsymbol{\Sigma}}) - \mathcal{A}(\boldsymbol{\Sigma})\|_{\mathbb{R}^d} + \frac{2}{1-\delta}\|\mathcal{A}(\boldsymbol{\Sigma}) - \mathcal{A}(\boldsymbol{\Sigma}_\mathfrak{S})\|_{\mathbb{R}^m} .
\end{aligned} \tag{33}$$

We introduce the following "distance" to the model set $\mathfrak{S}$:

$$d^\circ(\boldsymbol{\Sigma}, \mathfrak{S}) = \inf_{\mathbf{M}\in\mathfrak{S}} \|\boldsymbol{\Sigma} - \mathbf{M}\|_{S_d} + \frac{2}{1-\delta}\|\mathcal{A}(\boldsymbol{\Sigma}) - \mathcal{A}(\mathbf{M})\|_{\mathbb{R}^m} . \tag{34}$$

Then we have $d^\circ(\boldsymbol{\Sigma}, \mathfrak{S}) = 0$ if $\boldsymbol{\Sigma} \in \mathfrak{S}$ and

$$\|\boldsymbol{\Sigma}^* - \boldsymbol{\Sigma}\|_{S_d} \leq d^\circ(\boldsymbol{\Sigma}, \mathfrak{S}) + \frac{2}{1-\delta}\|\mathcal{A}(\boldsymbol{\Sigma}) - \mathcal{A}(\mathbf{M})\|_{\mathbb{R}^m} . \tag{35}$$

□

[10] We always assume that the minimization problem (7) has at least one solution. The decoder can be adjusted as in [50], to handle the case where the argmin is only approximated to a certain accuracy.

2) Proof of Theorem 2:

Theorem 2. *Consider a random sketching operator $\mathcal{A} : S_d(\mathbb{R}) \to \mathbb{R}^m$ and denote its operator norm by $|||\mathcal{A}||| \triangleq \sup_{\|\mathbf{U}\|_{S_d}=1} \|\mathcal{A}\mathbf{U}\|_{\mathbb{R}^m}$. Suppose that we are given two functions $C_1, C_2 : \mathbb{R}_+ \to \mathbb{R}_+$ such that*

$$\forall t > 0,\ \mathbb{P}\left(|||\mathcal{A}||| > t\right) \leq C_1(t)\,, \tag{11}$$

$$\forall \mathbf{U} \in S[\mathfrak{S}],\ \forall t > 0,\ \mathbb{P}\big(\,|\|\mathcal{A}(\mathbf{U})\|_{\mathbb{R}^m} - 1| > t\big) \leq C_2(t)\,. \tag{12}$$

Then, for any $\varepsilon > 0$ and $\delta \in [0, 1[$,

$$\sup_{\mathbf{U}\in S[\mathfrak{S}]} |\|\mathcal{A}(\mathbf{U})\|_{\mathbb{R}^m} - 1| < \delta\,, \tag{13}$$

with probability at least $1 - \mathcal{N}(S[\mathfrak{S}], \|\cdot\|_{S_d}, \varepsilon)C_2(\frac{\delta}{2}) - C_1(\frac{\delta}{2\varepsilon})$. Consequently with the same probability the operator $\mathcal{A}$ satisfies RIP_δ on $\mathfrak{S}$.

Proof. Let us start by considering $\overline{S}_\varepsilon$, an ε-net of $S[\mathfrak{S}]$ with respect to $\|\cdot\|_{S_d}$, for some $\varepsilon > 0$. Then, for every $\mathbf{U} \in S[\mathfrak{S}]$, there exists $\overline{\mathbf{U}} \in \overline{S}_\varepsilon$ such that $\|\mathbf{U} - \overline{\mathbf{U}}\|_{S_d} \leq \varepsilon$. From the triangular inequality we have

$$|\|\mathcal{A}(\mathbf{U})\|_{\mathbb{R}^m} - 1| \leq \big|\|\mathcal{A}(\mathbf{U})\|_{\mathbb{R}^m} - \|\mathcal{A}(\overline{\mathbf{U}})\|_{\mathbb{R}^m}\big| + \big|\|\mathcal{A}(\overline{\mathbf{U}})\|_{\mathbb{R}^m} - 1\big|\,.$$

Focusing on the first term of the right-hand side, we obtain

$$\big|\|\mathcal{A}(\mathbf{U})\|_{\mathbb{R}^m} - \|\mathcal{A}(\overline{\mathbf{U}})\|_{\mathbb{R}^m}\big| \leq \|\mathcal{A}(\mathbf{U} - \overline{\mathbf{U}})\|_{\mathbb{R}^m} \leq |||\mathcal{A}||| \cdot \|\mathbf{U} - \overline{\mathbf{U}}\|_{S_d} \leq \varepsilon|||\mathcal{A}|||\,. \tag{36}$$

Hence for $\mathbf{U} \in S[\mathrm{inv}(\mathfrak{S})]$:

$$|\|\mathcal{A}(\mathbf{U})\|_{\mathbb{R}^m} - 1| \leq \max_{\overline{U}\in\overline{S}_\varepsilon} \big|\|\mathcal{A}(\overline{\mathbf{U}})\|_{\mathbb{R}^m} - 1\big| + \varepsilon|||\mathcal{A}|||. \tag{37}$$

So we have for any $0 < \delta < 1$:

$$\mathbb{P}\left(\sup_{\mathbf{U}\in S[\mathfrak{S}]} |\|\mathcal{A}(\mathbf{U})\|_{\mathbb{R}^m} - 1| \leq \delta\right) \geq 1 - \mathbb{P}(\varepsilon|||\mathcal{A}||| > \frac{\delta}{2}) - \mathbb{P}(\max_{\overline{U}\in\overline{S}_\varepsilon} \big|\|\mathcal{A}(\overline{\mathbf{U}})\|_{\mathbb{R}^m} - 1\big| > \frac{\delta}{2})\,. \tag{38}$$

We will control these two terms. For the first one we have the concentration with C_1. For the second one, using the union bound yields

$$\mathbb{P}(\max_{\overline{U}\in\overline{S}_\varepsilon} \big|\|\mathcal{A}(\overline{\mathbf{U}})\|_{\mathbb{R}^m} - 1\big| > \frac{\delta}{2}) \leq \sum_{\overline{\mathbf{U}}\in\overline{S}_\varepsilon} \mathbb{P}(\big|\|\mathcal{A}(\overline{\mathbf{U}})\|_{\mathbb{R}^m} - 1\big| > \frac{\delta}{2}) \tag{39}$$

Using the concentration given by C_2, for $\overline{\mathbf{U}} \in \overline{S}_\varepsilon$ and $t \in]0, 1[$, we have $\mathbb{P}(\big|\|\mathcal{A}\overline{\mathbf{U}}\|_{\mathbb{R}^m} - 1\big| > t) \leq C_2(t)$. Applying this with $t = \delta/2$ and using (39) gives

$$\mathbb{P}(\max_{\overline{U}\in\overline{S}_\varepsilon} \big|\|\mathcal{A}(\overline{\mathbf{U}})\|_2 - 1\big| > \frac{\delta}{2}) \leq |\overline{S}_\varepsilon| C_2(\delta/2)\,. \tag{40}$$

As a result, we obtain for $\delta \in]0, 1[$ and $\varepsilon > 0$

$$\mathbb{P}\left(\sup_{\mathbf{U}\in S[\mathfrak{S}]} |\|\mathcal{A}(\mathbf{U})\|_2 - 1| \leq \delta\right) \geq 1 - C_1(\frac{\delta}{2\varepsilon}) - |\overline{S}_\varepsilon| C_2(\delta/2)\,. \tag{41}$$

□

3) Proof of Remark 2 :

Remark 2. *The above theorem is presented in its most usual form, with both controls (11) and (12). However, (12) is sometimes easier to obtain (by leveraging classical concentration inequalities). Fortunately, we can obtain (11) from (12)[11], as for $\varepsilon' > 0$, defining C_1 by*

$$C_1(t) = \mathcal{N}(B_{S_d}, \|\cdot\|_{S_d}, \varepsilon')\ C_2((1-\varepsilon')t - 1) \quad \forall t > 1/(1-\varepsilon')\,, \tag{14}$$

where $\mathcal{N}(B_{S_d}, \|\cdot\|_{S_d}, \varepsilon')$ is the covering number of the unit sphere $B_{S_d} = \{\mathbf{U} \in S_d : \|\mathbf{U}\|_{S_d} = 1\}$, yields a valid upper-bound of $\mathbb{P}\left(|||\mathcal{A}||| > t\right)$. See Appendix A3 for the proof.

Proof. For some $\varepsilon' > 0$, consider $T_{\varepsilon'}$ an ε'-net of the sphere B_{S_d}. Then, for all $\mathbf{U} \in B_{S_d}$, there exists $\overline{\mathbf{U}} \in T_{\varepsilon'}$ such that

$$\|\mathcal{A}(\mathbf{U})\|_{\mathbb{R}^m} \leq \varepsilon'|||\mathcal{A}||| + \|\mathcal{A}(\overline{\mathbf{U}})\|_{\mathbb{R}^m}\,.$$

[11]At the expense of a restricted range for t that will be of no consequence for the rest of the paper.

Thus, taking the supremum over $\mathbf{U}$ yields

$$\sup_{\mathbf{U}\in B_{S_d}} \|\mathcal{A}(\mathbf{U})\|_{\mathbb{R}^m} \leq \varepsilon' |||\mathcal{A}||| + \max_{\overline{\mathbf{U}}\in T_{\varepsilon'}} \|\mathcal{A}(\overline{\mathbf{U}})\|_{\mathbb{R}^m} ,$$

therefore $|||\mathcal{A}||| \leq (1-\varepsilon')^{-1} \max_{\overline{\mathbf{U}}\in T_{\varepsilon'}} \|\mathcal{A}(\overline{\mathbf{U}})\|_{\mathbb{R}^m}$. As a consequence, for $t > 1/(1-\varepsilon')$,

$$\begin{aligned}\mathbb{P}\left(|||\mathcal{A}||| > t\right) &\leq \mathbb{P}\left(\max_{\overline{\mathbf{U}}\in T_{\varepsilon'}} \|\mathcal{A}(\overline{\mathbf{U}})\|_{\mathbb{R}^m} > (1-\varepsilon')t\right) \leq \sum_{\overline{\mathbf{U}}\in T_{\varepsilon'}} \mathbb{P}\left(|\|\mathcal{A}(\overline{\mathbf{U}})\|_{\mathbb{R}^m} - 1| > (1-\varepsilon')t - 1\right) \\ &\leq \mathcal{N}(B_{S_d}, \|\cdot\|_{S_d}, \varepsilon') C_2((1-\varepsilon')t - 1, m) .\end{aligned}$$

□

B. Proofs of Section VI-A

1) General results on covering numbers : Let (E,d) be a semi-metric space. The covering number of $\mathfrak{S} \subseteq E$ with radius ε with respect to d is defined as:

$$\mathcal{N}(\mathfrak{S}, d, \varepsilon) \triangleq \min\left\{ N \in \mathbb{N} : \exists x_1, \cdots, x_N \in \mathfrak{S}, \mathfrak{S} \subseteq \bigcup_{i=1}^{N} B_d(x_i, \varepsilon) \right\} . \tag{42}$$

If $N = \mathcal{N}(\mathfrak{S}, d, \varepsilon)$, then for any $x \in \mathfrak{S}$ there exists $i \in [\![N]\!]$ such that $d(x, x_i) \leq \varepsilon$. We recall the following Lemma regarding covering numbers that can be found in [81, Lemma A.3.].

Lemma 3. *Let Y, Z we two subset of a pseudo metric space (X,d) such that the following holds:*

$$\forall z \in Z, \exists y \in Y, d(z,y) \leq \delta , \tag{43}$$

where $\delta \geq 0$. Then for all $\varepsilon > 0$

$$\mathcal{N}(Z, d, 2(\delta+\varepsilon)) \leq \mathcal{N}(Y, d, \varepsilon) . \tag{44}$$

2) Descent Lemma:

Lemma 4 (Descent Lemma)**.** *Let E, F be two normed vector spaces and Ω a subset of E and $\mathfrak{X} \subseteq \mathrm{int}(\Omega)$. Consider $f : \Omega \to F$ a L-smooth function on $\mathfrak{X}$ i.e.*

$$\forall (x,y) \in \mathfrak{X}^2, \ \|\mathrm{D}f_x - \mathrm{D}f_y\|_{\mathrm{op}} \leq L\|x-y\|_E . \tag{45}$$

Let $(x,y) \in \mathfrak{X}^2$ such that the segment $[\![x,y]\!]$ lies in $\mathfrak{X}$. Then,

$$\|f(x) - f(y) - \mathrm{D}f_y(x-y)\|_F \leq L\|x-y\|_E^2 . \tag{46}$$

*In particular if $\mathfrak{X}$ is convex then **A-2** implies **A-1**.*

Proof. From [88, Corollary 3.3] f verifies

$$\|f(x) - f(y) - \mathrm{D}f_y(x-y)\|_F \leq \sup_{z\in[\![x,y]\!]} \|\mathrm{D}f_z - \mathrm{D}f_x\|_{\mathrm{op}} \|x-y\|_E .$$

Now take $z \in [\![x,y]\!]$ and write it as $z = (1-t)x + ty \in \mathfrak{X}$ for some $t \in [0,1]$. Since f is L-smooth on $\mathfrak{X}$ we have:

$$\|\mathrm{D}f_z - \mathrm{D}f_x\|_{\mathrm{op}} = \|\mathrm{D}f_{(1-t)x+ty} - \mathrm{D}f_x\|_{\mathrm{op}} \leq L\|(1-t)x + ty - x\|_E = Lt\|x-y\|_E \leq L\|x-y\|_E$$

Hence $\sup_{z\in[\![x,y]\!]} \|\mathrm{D}f_z - \mathrm{D}f_x\|_{\mathrm{op}} \leq L\|x-y\|_E$ and thus $\|f(x) - f(y) - \mathrm{D}f_y(x-y)\|_F \leq L\|x-y\|_E^2$ □

3) Proof of Proposition 1 :

Proof. In the following we note $\|(x_1,y_1) - (x_2,y_2)\|_{\otimes 2} = \|x_1 - x_2\|_E + \|y_1 - y_2\|_E$. We introduce, for $\eta \geq 0$, the set

$$\mathfrak{X}_\eta^2 = \{(x,y) \in \mathfrak{X}^2 : \ \|f(x) - f(y)\|_F > \eta\} . \tag{47}$$

First note that $\mathfrak{X}_\eta^2 \subset \mathfrak{X}^2$, and consequently

$$\mathcal{N}(\mathfrak{X}_\eta^2, \|\cdot\|_{\otimes 2}, \varepsilon) \leq \mathcal{N}(\mathfrak{X}^2, \|\cdot\|_{\otimes 2}, \frac{\varepsilon}{2}) \leq \mathcal{N}(\mathfrak{X}, \|\cdot\|_E, \frac{\varepsilon}{4})^2 . \tag{48}$$

We consider $\eta > 0$ and define $g : \mathfrak{X}_\eta^2 \to S_\eta^+(f(\mathfrak{X}))$ by $g(x,y) = \frac{f(x)-f(y)}{\|f(x)-f(y)\|_F}$ for $(x,y) \in \mathfrak{X}_\eta^2$. By definition, g is surjective. We will show that it is also Lipschitz. With $(x_1,y_1),(x_2,y_2) \in \mathfrak{X}_\eta^2 \times \mathfrak{X}_\eta^2$ we obtain

$$
\begin{aligned}
&\|\frac{f(x_1)-f(y_1)}{\|f(x_1)-f(y_1)\|_F} - \frac{f(x_2)-f(y_2)}{\|f(x_2)-f(y_2)\|_F}\|_F \\
\leq &\|\frac{f(x_1)-f(y_1)}{\|f(x_1)-f(y_1)\|_F} - \frac{f(x_1)-f(y_1)}{\|f(x_2)-f(y_2)\|_F}\|_F + \|\frac{f(x_1)-f(y_1)}{\|f(x_2)-f(y_2)\|_F} - \frac{f(x_2)-f(y_2)}{\|f(x_2)-f(y_2)\|_F}\|_F \\
\leq &\frac{1}{\|f(x_2)-f(y_2)\|_F}\left(\|f(x_1)-f(x_2)\|_F + \|f(y_1)-f(y_2)\|_F\right) \\
&\quad + \left|\frac{1}{\|f(x_1)-f(y_1)\|_F} - \frac{1}{\|f(x_2)-f(y_2)\|_F}\right| \|f(x_1)-f(y_1)\|_F \\
\leq &\frac{2\beta}{\|f(x_2)-f(y_2)\|_F}\|(x_1,y_1)-(x_2,y_2)\|_{\otimes 2} \\
&\quad + \quad \left|\|f(x_2)-f(y_2)\|_F - \|f(x_1)-f(y_1)\|_F\right| \times \frac{\|f(x_1)-f(y_1)\|_F}{\|f(x_1)-f(y_1)\|_F\|f(x_2)-f(y_2)\|_F} \\
\leq &\frac{2\beta}{\eta}\|(x_1,y_1)-(x_2,y_2)\|_{\otimes 2} + \|f(x_2)-f(x_1)-f(y_2)+f(y_1)\|_F \times \frac{1}{\|f(x_2)-f(y_2)\|_F} \\
\leq &\frac{2\beta}{\eta}\|(x_1,y_1)-(x_2,y_2)\|_{\otimes 2} + 2\beta\|(x_1,y_1)-(x_2,y_2)\|_{\otimes 2} \times \frac{1}{\|f(x_2)-f(y_2)\|_F} \\
\leq &\frac{4\beta}{\eta}\|(x_1,y_1)-(x_2,y_2)\|_{\otimes 2}\,.
\end{aligned}
$$

Consequently,

$$
\|g(x_1,y_1)-g(x_2,y_2)\|_F \leq \frac{4\beta}{\eta}\|(x_1,y_1)-(x_2,y_2)\|_{\otimes 2}\,. \tag{49}
$$

So g is a surjective $\frac{4\beta}{\eta}$-Lipschitz function from $(\mathfrak{X}_\eta^2, \|\cdot\|_{\otimes 2})$ to $(S_\eta^+(f(\mathfrak{X})), \|\cdot\|_F)$. Hence using [81, Lemma A.2.] we obtain

$$
\mathcal{N}(S_\eta^+(f(\mathfrak{X})), \|\cdot\|_F, \varepsilon) \leq \mathcal{N}(\mathfrak{X}_\eta^2, \|\cdot\|_{\otimes 2}, \frac{\eta}{4\beta}\varepsilon) \leq \mathcal{N}(\mathfrak{X}, \|\cdot\|_E, \frac{\eta}{16\beta}\varepsilon)^2\,. \tag{50}
$$

□

4) Proof of Proposition 2 : We recall the proposition:

Proposition 2. *Assume that f is (α,β)-bi-Lipschitz and satisfies assumptions **A-1** and **A-2**. Then for any $\varepsilon > 0$ and $\eta > 0$,*

$$
\mathcal{N}(S_\eta^-[f(\mathfrak{X})], \|\cdot\|_F, 2(\varepsilon + \frac{L}{\alpha^2}\eta)) \leq \left(\frac{\zeta+\beta\alpha}{\alpha^2\varepsilon}\right) \times \mathcal{N}(\mathfrak{X}, \|\cdot\|_E, \frac{\alpha\varepsilon}{\zeta+\beta\alpha}) \times \mathcal{N}(S[\mathfrak{X}], \|\cdot\|_E, \frac{\alpha^2\varepsilon}{2(\zeta+\beta\alpha)})\,. \tag{30}
$$

In order to prove this result, the reasoning will be the following: 1) we will show that any element of $S_\eta^-[f(\mathfrak{X})]$ is close to an element of a certain "tangent space" $\mathrm{D}\frac{x-y}{\|f(x)-f(y)\|_F}$ 2) we will control the covering number of this space. We begin with the following result.

Lemma 5. *Assume that f is (α,β)-bi-Lipschitz and satisfies assumptions **A-1**. For $\varepsilon_0 > 0$, consider the following subset of $\mathfrak{X}^2$:*

$$
I_{\varepsilon_0} \triangleq \{(x,y) \in \mathfrak{X}^2 :\ 0 < \|x-y\|_E \leq \varepsilon_0\}\,. \tag{51}
$$

Then, for any $\varepsilon_0 > 0$ and $(x,y) \in I_{\varepsilon_0}$,

$$
\|\frac{f(x)-f(y)}{\|f(x)-f(y)\|_F} - \mathrm{D}f_y\frac{x-y}{\|f(x)-f(y)\|_F}\|_F \leq \frac{L}{\alpha}\varepsilon_0\,.
$$

In particular, for any $\varepsilon_0 > 0$,

$$
\forall (x,y) \in I_{\varepsilon_0},\ \exists h \in \mathfrak{X},\ \|\frac{f(x)-f(y)}{\|f(x)-f(y)\|_F} - \mathrm{D}f_h\frac{x-y}{\|f(x)-f(y)\|_F}\|_F \leq \frac{L}{\alpha}\varepsilon_0\,.
$$

Proof. First, assumption **A-1** gives

$$
\|f(x)-f(y)-\mathrm{D}f_y(x-y)\|_F \overset{\text{A-1}}{\leq} L\|x-y\|_E^2 \leq L\varepsilon_0\|x-y\|_E\,. \tag{52}
$$

Now since $\|x-y\|_E > 0$, we have $\|f(x)-f(y)\|_F > 0$ using (27) (from the inverse Lipschitz property). Thus by dividing by $\|f(x)-f(y)\|_F$ in (52) we obtain

$$\|\frac{f(x)-f(y)}{\|f(x)-f(y)\|_F} - \mathrm{D}f_y \frac{x-y}{\|f(x)-f(y)\|_F}\|_F \leq L\varepsilon_0 \frac{\|x-y\|_E}{\|f(x)-f(y)\|_F} \overset{(27)}{\leq} \frac{L}{\alpha}\varepsilon_0 .$$

□

The above result induces the following corollary.

Corollary 2. *Assume that f is (α,β)-bi-Lipschitz and satisfies assumptions **A-1**. For $\eta > 0$ consider $S_\eta^-[f(\mathfrak{X})]$ as defined in (26) and the normalized secant set $S[\mathfrak{X}]$ (see Definition 2). Let*

$$C \triangleq \{\lambda v : \ \lambda \in]0, \frac{1}{\alpha}], v \in S[\mathfrak{X}]\} , \tag{53}$$

and

$$T_C \triangleq \{\mathrm{D}f_h(c) : (h,c) \in \mathfrak{X} \times C\} . \tag{54}$$

Then

$$\forall u \in S_\eta^-[f(\mathfrak{X})], \exists t \in T_C, \ \|u-t\|_F \leq \frac{L}{\alpha^2}\eta . \tag{55}$$

Proof. Take $u = \frac{f(x)-f(y)}{\|f(x)-f(y)\|_F} \in S_\eta^-[f(\mathfrak{X})]$ (thus with $0 < \|f(x)-f(y)\|_F \leq \eta$). By using that f is (α,β)-bi-Lipschitz, we have $0 < \|x-y\|_E \leq \eta/\alpha$. So we can apply the Lemma 5 with $\varepsilon_0 = \eta/\alpha$ to prove that there exists $h \in \mathfrak{X}$ such that

$$\|\frac{f(x)-f(y)}{\|f(x)-f(y)\|_F} - \mathrm{D}f_h \frac{x-y}{\|f(x)-f(y)\|_F}\|_F \leq \frac{L}{\alpha^2}\eta .$$

Rewrite $\frac{x-y}{\|f(x)-f(y)\|_F} = \frac{x-y}{\|x-y\|_F}\frac{\|x-y\|_E}{\|f(x)-f(y)\|_F}$ and define $\lambda = \frac{\|x-y\|_E}{\|f(x)-f(y)\|_F}$. Therefore, we have $\lambda > 0$ and $|\lambda| \leq \frac{1}{\alpha}$. It proves that there exists $\lambda \in]0, 1/\alpha]$ and $h \in \mathfrak{X}$, such that

$$\|\frac{f(x)-f(y)}{\|f(x)-f(y)\|_F} - \mathrm{D}f_h \lambda \frac{x-y}{\|x-y\|_E}\|_F \leq \frac{L}{\alpha^2}\eta .$$

Considering $v = \frac{x-y}{\|x-y\|_E} \in S[\mathfrak{X}]$ concludes the proof.

□

The last thing to do is to control the covering number of T_C in the previous result. This will be done using the following lemma.

Lemma 6. *Let $C \subseteq E$ be any set such that $\forall c \in C, \|c\|_E \leq \delta$ for some $\delta > 0$. Assume that f is (α,β)-bi-Lipschitz and satisfies **A-2**. Consider $T_C \triangleq \{\mathrm{D}f_h(c) : (h,c) \in \mathfrak{X} \times C\}$. Then for any $\varepsilon > 0$,*

$$\mathcal{N}(T_C, \|\cdot\|_F, \varepsilon) \leq \mathcal{N}(\mathfrak{X}, \|\cdot\|_E, \frac{\varepsilon}{\zeta\delta+\beta}) \times \mathcal{N}(C, \|\cdot\|_E, \frac{\varepsilon}{\zeta\delta+\beta}) .$$

Proof. Take $\overline{\mathfrak{X}}$ a ε-net of $\mathfrak{X}$ and $\overline{C}$ a ε-net of C. Then take $t = \mathrm{D}f_h(u) \in T_C$ and consider $\overline{u}, \overline{h} \in \overline{\mathfrak{X}} \times \overline{C}$ such that $\|\overline{c}-c\|_E \leq \varepsilon$ and $\|\overline{h}-h\|_E \leq \varepsilon$. Then with $\overline{t} = \mathrm{D}f_{\overline{h}}(\overline{c}) \in T_C$

$$\begin{aligned}
\|t-\overline{t}\|_F &= \|\mathrm{D}f_h(c) - \mathrm{D}f_{\overline{h}}(\overline{c})\|_F \leq \|\mathrm{D}f_h(c) - \mathrm{D}f_{\overline{h}}(c)\|_F + \|\mathrm{D}f_{\overline{h}}(c) - \mathrm{D}f_{\overline{h}}(\overline{c})\|_F \\
&\leq \|\mathrm{D}f_h - \mathrm{D}f_{\overline{h}}\|_{\mathrm{op}}\|c\|_E + \|\mathrm{D}f_{\overline{h}}\|_{\mathrm{op}}\|c-\overline{c}\|_E \\
&\overset{\text{A-2}}{\leq} \zeta\|h-\overline{h}\|_E\delta + \sup_{h\in\mathfrak{X}}\|\mathrm{D}f_h\|_{\mathrm{op}}\varepsilon \\
&\leq \varepsilon(\zeta\delta + \sup_{h\in\mathfrak{X}}\|\mathrm{D}f_h\|_{\mathrm{op}}) \overset{(28)}{\leq} \varepsilon(\zeta\delta+\beta) .
\end{aligned} \tag{56}$$

This gives $\mathcal{N}(T_C, \|\cdot\|_F, \varepsilon(\zeta\delta+\beta)) \leq |\overline{\mathfrak{X}}| \times |\overline{C}| \leq \mathcal{N}(\mathfrak{X}, \|\cdot\|_E, \varepsilon) \times \mathcal{N}(C, \|\cdot\|_E, \varepsilon)$.

□

We can now prove Proposition 2 as follows.

Proof. Proof of Proposition 2. Consider C and T_C as defined in Corollary 2. We have for any $c \in C, \|c\|_E \leq \frac{1}{\alpha}$ since $\forall u \in S[\mathfrak{X}], \|u\|_E = 1$. So by applying Lemma 6 with $\delta = \frac{1}{\alpha}$ we obtain

$$\mathcal{N}(T_C, \|\cdot\|_F, \varepsilon) \leq \mathcal{N}(\mathfrak{X}, \|\cdot\|_E, \frac{\varepsilon}{\frac{\zeta}{\alpha}+\beta}) \times \mathcal{N}(C, \|\cdot\|_E, \frac{\varepsilon}{\frac{\zeta}{\alpha}+\beta}) .$$

Also, by Corollary 2,

$$\forall u \in S_\eta^-[f(\mathfrak{X})], \exists t \in T_C,\ \|u-t\|_F \leq \frac{L}{\alpha^2}\eta\,. \tag{57}$$

We then apply Lemma 3 (with $\delta = \frac{L}{\alpha^2}\eta$) to prove that, for any $\varepsilon > 0$,

$$\mathcal{N}(S_\eta^-[f(\mathfrak{X})], \|\cdot\|_F, 2(\varepsilon + \frac{L}{\alpha^2}\eta)) \leq \mathcal{N}(T_C, \|\cdot\|_F, \varepsilon)\,. \tag{58}$$

Consequently,

$$\mathcal{N}(S_\eta^-[f(\mathfrak{X})], \|\cdot\|_F, 2(\varepsilon + \frac{L}{\alpha^2}\eta)) \leq \mathcal{N}(\mathfrak{X}, \|\cdot\|_E, \frac{\varepsilon}{\frac{\zeta}{\alpha}+\beta}) \times \mathcal{N}(C, \|\cdot\|_E, \frac{\varepsilon}{\frac{\zeta}{\alpha}+\beta})\,. \tag{59}$$

All we need now is to control $\mathcal{N}(C, \|\cdot\|_F, \frac{\varepsilon}{\frac{\zeta}{\alpha}+\beta})$. Take $\overline{S[\mathfrak{X}]}$ a ε-net of $S[\mathfrak{X}]$ with respect to $\|\cdot\|_E$ and $\overline{(0,\frac{1}{\alpha}]}$ a (ε/α)-net of $(0,\frac{1}{\alpha}]$ with respect to $|\cdot|$. Consider $c = \lambda u \in C$ with $\lambda \in (0,\frac{1}{\alpha}]$ and $u \in S[\mathfrak{X}]$. Then there exists $\overline{u} \in \overline{S[\mathfrak{X}]}$ such that $\|u-\overline{u}\|_E \leq \varepsilon$ and there exists $\overline{\lambda} \in \overline{(0,\frac{1}{\alpha}]}$ such that $|\lambda - \overline{\lambda}| \leq \varepsilon/\alpha$. We consider $\overline{c} = \overline{\lambda}\overline{u}$ which belongs to C. Then $\|c-\overline{c}\|_E = \|\lambda u - \overline{\lambda}\overline{u}\|_E \leq |\lambda-\overline{\lambda}|\|\overline{u}\|_E + |\lambda|\|\overline{u}-u\|_E \leq 2\varepsilon/\alpha$. Thus for any $\varepsilon > 0$ we have

$$\mathcal{N}(C, \|\cdot\|_E, 2\varepsilon/\alpha) \leq \mathcal{N}(S[\mathfrak{X}], \|\cdot\|_E, \varepsilon) \times \mathcal{N}((0,\frac{1}{\alpha}], |\cdot|, \varepsilon/\alpha) = \mathcal{N}(S[\mathfrak{X}], \|\cdot\|_E, \varepsilon) \times \mathcal{N}((0,1], |\cdot|, \varepsilon)\,. \tag{60}$$

Equivalently with a change of variable $\varepsilon \leftarrow 2\varepsilon/\alpha$ we have $\forall \varepsilon > 0, \mathcal{N}(C, \|\cdot\|_E, \varepsilon) \leq \mathcal{N}(S[\mathfrak{X}], \|\cdot\|_E, \frac{\varepsilon\alpha}{2}) \times \mathcal{N}((0,1], |\cdot|, \frac{\varepsilon\alpha}{2})$. Overall,

$$\mathcal{N}(S_\eta^-[f(\mathfrak{X})], \|\cdot\|_F, 2(\varepsilon + \frac{L}{\alpha^2}\eta)) \leq \mathcal{N}(\mathfrak{X}, \|\cdot\|_E, \frac{\varepsilon}{\frac{\zeta}{\alpha}+\beta}) \times \mathcal{N}(S[\mathfrak{X}], \|\cdot\|_E, \frac{\alpha\varepsilon}{2(\frac{\zeta}{\alpha}+\beta)}) \times \mathcal{N}((0,1], |\cdot|, \frac{\alpha\varepsilon}{2(\frac{\zeta}{\alpha}+\beta)})\,. \tag{61}$$

But for any $\varepsilon > 0, \mathcal{N}((0,1], |\cdot|, \varepsilon) \leq \frac{1}{2\varepsilon}$ thus

$$\mathcal{N}(S_\eta^-[f(\mathfrak{X})], \|\cdot\|_F, 2(\varepsilon + \frac{L}{\alpha^2}\eta)) \leq \left(\frac{\zeta+\beta\alpha}{\alpha^2\varepsilon}\right) \times \mathcal{N}(\mathfrak{X}, \|\cdot\|_E, \frac{\varepsilon}{\frac{\zeta}{\alpha}+\beta}) \times \mathcal{N}(S[\mathfrak{X}], \|\cdot\|_E, \frac{\alpha\varepsilon}{2(\frac{\zeta}{\alpha}+\beta)})\,, \tag{62}$$

which concludes the proof. □

5) Proof of Lemma 1 :

Proof. The proof is based on various computations and the following identity:

$$\boldsymbol{\Theta}_1^{-1} + \boldsymbol{\Theta}_2^{-1} = \boldsymbol{\Theta}_1^{-1}(\boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_2)\boldsymbol{\Theta}_2^{-1}, \qquad \forall (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \in (\mathfrak{S}_{k,a,b}^{-1})^2. \tag{63}$$

We will also use

$$\|\mathbf{AB}\|_{\mathrm{Fro}} \leq \|\mathbf{A}\|_{2\to2}\|\mathbf{B}\|_{\mathrm{Fro}} \quad \text{and} \quad \|\mathbf{AB}\|_{\mathrm{Fro}} \leq \|\mathbf{A}\|_{\mathrm{Fro}}\|\mathbf{B}\|_{2\to2}\,, \quad \forall \mathbf{A}, \mathbf{B} \in S_d(\mathbb{R})\,.$$

For the rest of the proof, we take $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \in \mathfrak{S}_{k,a,b}^{-1}$. Hence we have $\|\boldsymbol{\Theta}_1\|_{2\to2} \leq b$ and $\|\boldsymbol{\Theta}_1^{-1}\|_{2\to2} \leq \frac{1}{a}$ (the same inequalities hold for $\boldsymbol{\Theta}_2$).

Now we can prove that

$$\begin{aligned}\|\boldsymbol{\Theta}_1^{-1} - \boldsymbol{\Theta}_2^{-1}\|_\Lambda &\leq C_{\mathrm{Fro}}\|\boldsymbol{\Theta}_1^{-1}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)(-\boldsymbol{\Theta}_2)^{-1}\|_{\mathrm{Fro}} \leq C_{\mathrm{Fro}}\|\boldsymbol{\Theta}_1^{-1}\|_{2\to2}\|\boldsymbol{\Theta}_2^{-1}\|_{2\to2}\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|_{\mathrm{Fro}} \\ &\leq C_{\mathrm{Fro}}\frac{1}{a^2}\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|_{\mathrm{Fro}}\,.\end{aligned}$$

This proves (28) with $\beta = \frac{C_{\mathrm{Fro}}}{a^2}$. The reverse inequality for (27) can be proven by using this time

$$\begin{aligned}\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|_{\mathrm{Fro}} &= \|\boldsymbol{\Theta}_1(\boldsymbol{\Theta}_1^{-1} - \boldsymbol{\Theta}_2^{-1})\boldsymbol{\Theta}_2\|_{\mathrm{Fro}} \leq \|\boldsymbol{\Theta}_1\|_{2\to2}\|\boldsymbol{\Theta}_2\|_{2\to2}\|\boldsymbol{\Theta}_1^{-1} - \boldsymbol{\Theta}_2^{-1}\|_{\mathrm{Fro}} \\ &\leq b^2\frac{1}{c_{\mathrm{Fro}}}\|\boldsymbol{\Theta}_1^{-1} - \boldsymbol{\Theta}_2^{-1}\|_\Lambda\,.\end{aligned}$$

Thus we have (27) with $\alpha = \frac{c_{\mathrm{Fro}}}{b^2}$.

For assumption **A-1**, we use the formula of the differential of the inverse of a matrix $\mathrm{D}\,\mathrm{inv}_{\boldsymbol{\Theta}}[\mathbf{H}] = -\boldsymbol{\Theta}^{-1}\mathbf{H}\boldsymbol{\Theta}^{-1}$. We have:

$$\begin{aligned}\|\boldsymbol{\Theta}_1^{-1} - \boldsymbol{\Theta}_2^{-1} - \mathrm{D}\,\mathrm{inv}_{\boldsymbol{\Theta}_2}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\|_\Lambda &= \|\boldsymbol{\Theta}_1^{-1} - \boldsymbol{\Theta}_2^{-1} - [-\boldsymbol{\Theta}_2^{-1}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\boldsymbol{\Theta}_2^{-1}]\|_\Lambda \\ &= \|\boldsymbol{\Theta}_1^{-1}(\boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_1)\boldsymbol{\Theta}_2^{-1} - \boldsymbol{\Theta}_2^{-1}(\boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_1)\boldsymbol{\Theta}_2^{-1}\|_\Lambda \\ &= \|\left(\boldsymbol{\Theta}_1^{-1}(\boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_1) - \boldsymbol{\Theta}_2^{-1}(\boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_1)\right)\boldsymbol{\Theta}_2^{-1}\|_\Lambda \\ &= \|\left(\boldsymbol{\Theta}_1^{-1} - \boldsymbol{\Theta}_2^{-1}\right)(\boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_1)\boldsymbol{\Theta}_2^{-1}\|_\Lambda \\ &\leq C_{\mathrm{Fro}}\|\boldsymbol{\Theta}_2^{-1}\|_{2\to2}\|\boldsymbol{\Theta}_1^{-1} - \boldsymbol{\Theta}_2^{-1}\|_{\mathrm{Fro}}\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|_{\mathrm{Fro}} \\ &\leq C_{\mathrm{Fro}}\|\boldsymbol{\Theta}_2^{-1}\|_{2\to2}\|\boldsymbol{\Theta}_1^{-1}\|_{2\to2}\|\boldsymbol{\Theta}_2^{-1}\|_{2\to2}\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|_{\mathrm{Fro}}^2 \\ &\leq C_{\mathrm{Fro}}\frac{1}{a^3}\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|_{\mathrm{Fro}}^2.\end{aligned}$$

This gives **A-1** with $L = \frac{C_{\text{Fro}}}{a^3}$.

Now take $\mathbf{M}$, such that $\|\mathbf{M}\|_{\text{Fro}} \leq 1$. We have

$$\begin{aligned}
\|\mathrm{D}\,\mathrm{inv}_{\mathbf{\Theta}_1}(\mathbf{M}) - \mathrm{D}\,\mathrm{inv}_{\mathbf{\Theta}_2}(\mathbf{M})\|_\Lambda &= \|\mathbf{\Theta}_1^{-1}\mathbf{M}\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1}\mathbf{M}\mathbf{\Theta}_2^{-1}\|_\Lambda \\
&\leq \|\mathbf{\Theta}_1^{-1}\mathbf{M}\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_1^{-1}\mathbf{M}\mathbf{\Theta}_2^{-1}\|_\Lambda + \|\mathbf{\Theta}_1^{-1}\mathbf{M}\mathbf{\Theta}_2^{-1} - \mathbf{\Theta}_2^{-1}\mathbf{M}\mathbf{\Theta}_2^{-1}\|_\Lambda \\
&= \|\mathbf{\Theta}_1^{-1}\mathbf{M}(\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1})\|_\Lambda + \|(\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1})\mathbf{M}\mathbf{\Theta}_2^{-1}\|_\Lambda \\
&\leq C_{\text{Fro}}\|\mathbf{\Theta}_1^{-1}\mathbf{M}\|_{\text{Fro}}\|\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1}\|_{\text{Fro}} + C_{\text{Fro}}\|\mathbf{M}\mathbf{\Theta}_2^{-1}\|_{\text{Fro}}\|\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1}\|_{\text{Fro}} \\
&\leq C_{\text{Fro}}\left(\|\mathbf{\Theta}_1^{-1}\|_{2\to 2} + \|\mathbf{\Theta}_2^{-1}\|_{2\to 2}\right)\|\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1}\|_{\text{Fro}} \\
&\leq 2C_{\text{Fro}}(\frac{1}{a})\|\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1}\|_{\text{Fro}} \\
&\leq 2C_{\text{Fro}}(\frac{1}{a^3})\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_{\text{Fro}}\,.
\end{aligned}$$

This gives **A-2** with $\zeta = 2\frac{C_{\text{Fro}}}{a^3} = 2L$. □

6) Proof of Lemma 2 : In order to prove the result we will use the following lemma.

Lemma 7. *Consider*

$$\mathfrak{W}_k = \{\mathbf{\Theta} \in S_d(\mathbb{R}) : \ \|\mathbf{\Theta}\|_0 \leq d + 2k, \|\mathbf{\Theta}\|_{\text{Fro}} \leq 1\}\,. \tag{64}$$

Then

$$\mathcal{N}(\mathfrak{W}_k, \|\cdot\|_{\text{Fro}}, \varepsilon) \leq (\frac{ed^2}{2k})^k(\frac{9}{\varepsilon})^{d+k}\,. \tag{65}$$

Proof. Take $\mathbf{\Theta} \in \mathfrak{W}_k$, it can be written as $\mathbf{\Theta} = \mathbf{D} + \mathbf{T} + \mathbf{T}^\top$ where $\mathbf{D}$ is diagonal with d positive elements, $\mathbf{T}$ is a stricly upper triangular matrix with at most k non zero elements. We have also that $\|\mathbf{D}\|_{\text{Fro}} \leq \|\mathbf{\Theta}\|_{\text{Fro}} \leq 1$ and same for $\mathbf{T}, \mathbf{T}^\top$. Consider $\overline{D}$ a $\varepsilon/3$-net for the diagonal and $\overline{T}$ a $\varepsilon/3$-net for the upper triangle, both with respect to the $\|\cdot\|_{\text{Fro}}$ norm. Then with standard covering arguments $|\overline{D}| \leq (9/\varepsilon)^d$ and $|\overline{T}| \leq \binom{\frac{d(d-1)}{2}}{k}(\frac{9}{\varepsilon})^k$ because it is included in the unit ball of k-sparse vector in dimension $\frac{d(d-1)}{2}$ (see *e.g.* [26]). Consider $\overline{\mathfrak{W}_k} = \{\mathbf{D}_* + \mathbf{T}_* + \mathbf{T}_*^\top, (\mathbf{D}_*, \mathbf{T}_*) \in \overline{D} \times \overline{T}\}$. Then $|\overline{\mathfrak{W}_k}| \leq \binom{\frac{d(d-1)}{2}}{k}(\frac{9}{\varepsilon})^k(\frac{9}{\varepsilon})^d = \binom{\frac{d(d-1)}{2}}{k}(\frac{9}{\varepsilon})^{d+k}$. Also, for any $\mathbf{\Theta} = \mathbf{D} + \mathbf{T} + \mathbf{T}^\top$ there exists $(\mathbf{D}_*, \mathbf{T}_*) \in \overline{D} \times \overline{T}$ such that $\|\mathbf{D} - \mathbf{D}_*\|_{\text{Fro}} \leq \varepsilon/3, \|\mathbf{T} - \mathbf{T}_*\|_{\text{Fro}} \leq \varepsilon/3$. Hence:

$$\|\mathbf{D} + \mathbf{T} + \mathbf{T}^\top - (\mathbf{D}_* + \mathbf{T}_* + \mathbf{T}_*^\top)\|_{\text{Fro}} \leq \|\mathbf{D} - \mathbf{D}_*\|_{\text{Fro}} + 2\|\mathbf{T} - \mathbf{T}_*\|_{\text{Fro}} \leq \varepsilon\,. \tag{66}$$

Hence,

$$\mathcal{N}(\mathfrak{W}_k, \|\cdot\|_{\text{Fro}}, \varepsilon) \leq \binom{\frac{d(d-1)}{2}}{k}(\frac{9}{\varepsilon})^{d+k} \leq (\frac{ed(d-1)}{2k})^k(\frac{9}{\varepsilon})^{d+k} \leq (\frac{ed^2}{2k})^k(\frac{9}{\varepsilon})^{d+k}\,, \tag{67}$$

where in the last inequality we used the bound [26, Lemma C.5]. Note that we only considered the fact that $\mathfrak{W}$ is the space of symmetric an $d+2k$ sparse matrices. Restricting to positive definite matrices could be an avenue for further improvements. □

As a consequence we can prove Lemma 2 as follows.

Proof of Lemma 2. Recall that

$$\mathfrak{S}_{k,a,b}^{-1} = \{\mathbf{\Theta} \in S_d^{++}(\mathbb{R}) \ : \|\mathbf{\Theta}\|_0 \leq d + 2k, \mathrm{spec}(\mathbf{\Theta}) \subseteq [a, b]\}\,.$$

Consider $\mathbf{\Theta} \in \mathfrak{S}_{k,a,b}^{-1}$. We have $\|\mathbf{\Theta}\|_{2\to 2} \leq b$ which implies $\|\mathbf{\Theta}\|_{\text{Fro}} \leq \sqrt{d}\|\mathbf{\Theta}\|_{2\to 2} \leq \sqrt{d}b$. Thus $\mathfrak{S}_{k,a,b}^{-1} \subset \{\mathbf{\Theta} \in S_d^{++}(\mathbb{R}) : \|\mathbf{\Theta}\|_0 \leq d + 2k, \|\mathbf{\Theta}\|_{\text{Fro}} \leq \sqrt{d}b\} = \sqrt{d}b\mathfrak{W}$. Consequently,

$$\mathcal{N}(\mathfrak{S}_{k,a,b}^{-1}, \|\cdot\|_{\text{Fro}}, \varepsilon) \leq \mathcal{N}(\sqrt{d}b\mathfrak{W}, \|\cdot\|_{\text{Fro}}, \varepsilon/2) \leq \mathcal{N}(\mathfrak{W}, \|\cdot\|_{\text{Fro}}, \varepsilon/2\sqrt{d}b) \overset{\text{Lemma 7}}{\leq} (\frac{ed^2}{2k})^k(\frac{18\sqrt{d}b}{\varepsilon})^{d+k}\,, \tag{68}$$

which concludes the first part. For the second part we recall the definition of the normalized secant set

$$S[\mathfrak{S}_{k,a,b}^{-1}] = \{\frac{\mathbf{\Theta}_1 - \mathbf{\Theta}_2}{\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_{\text{Fro}}} \ : \ (\mathbf{\Theta}_1, \mathbf{\Theta}_2) \in (\mathfrak{S}_{k,a,b}^{-1})^2, \|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_{\text{Fro}} > 0\}\,.$$

Moreover, if $\mathbf{U} = \mathbf{\Theta}_1 - \mathbf{\Theta}_2 \in \mathfrak{S}_{k,a,b}^{-1} - \mathfrak{S}_{k,a,b}^{-1}$ then $\|\mathbf{U}\|_0 \leq d + 4k$. Indeed, $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ have d nonzeros elements on the diagonal (since both are postitive definite) and since these matrices are symmetric then they have at most k nonzeros elements in the upper-triangular (*resp.* lower-triangular) part. Thus $\mathbf{\Theta}_1 - \mathbf{\Theta}_2$ has at most $2k$ nonzeros elements in the upper-triangular (*resp.* lower-triangular) part. Consequently,

$$S[\mathfrak{S}_{k,a,b}^{-1}] \subset \{\mathbf{M} \in S_d(\mathbb{R}) : \|\mathbf{M}\|_{\text{Fro}} \leq 1, \|\mathbf{M}\|_0 \leq d + 4k\} = \mathfrak{W}_{2k}\,.$$

This gives

$$\mathcal{N}(S[\mathfrak{S}_{k,a,b}^{-1}],\|\cdot\|_{\mathrm{Fro}},\varepsilon)\leq\mathcal{N}(\mathfrak{W}_{2k},\|\cdot\|_{\mathrm{Fro}},\varepsilon/2)\leq(\frac{ed^2}{4k})^{2k}(\frac{18}{\varepsilon})^{d+2k}\,. \tag{69}$$

□

7) Proof of Corollary 1 : We recall the statement.

Corollary 1. *Assuming the existence of the constant $C_{\mathrm{Fro}},c_{\mathrm{Fro}}$ as in Lemma 1, there exist absolute constants $c_0,c_1\geq 1$, such that, for any $\varepsilon>0$, the covering number of $S[\mathfrak{S}_{k,a,b}]$ verifies*

$$\mathcal{N}\left(S[\mathfrak{S}_{k,a,b}],\|\cdot\|_{\Lambda},\varepsilon\right)\leq\left(\frac{ed^2}{2k}\right)^{4k}\left[\left(c_0\frac{\sqrt{d}C_{\mathrm{Fro}}^2}{\varepsilon^2c_{\mathrm{Fro}}^2}\frac{b^5}{a^5}\right)^{2(d+k)}+\left(c_1\frac{\sqrt{d}C_{\mathrm{Fro}}^2}{\varepsilon^2c_{\mathrm{Fro}}}(\frac{2}{c_{\mathrm{Fro}}}+1)\frac{b^5}{a^5}\right)^{d+2k+1}\right].$$

Proof. From Lemma 1, the assumptions of Theorem 4 hold for $f=\mathrm{inv}$, $\mathfrak{S}=\mathfrak{S}_{k,a,b}^{-1}$, $\|\cdot\|_E=\|\cdot\|_{\mathrm{Fro}}$ and $\|\cdot\|_F=\|\cdot\|_{\Lambda}$. Thus, for any $\eta,\varepsilon'>0$ we have the inequality

$$\begin{aligned}\mathcal{N}(S[\mathfrak{S}_{k,a,b}],\|\cdot\|_{\Lambda},2\left[\varepsilon'+\frac{L}{\alpha^2}\eta\right])\leq&\mathcal{N}(\mathfrak{S}_{k,a,b}^{-1},\|\cdot\|_{\mathrm{Fro}},\frac{\eta}{8\beta}\left[\varepsilon'+\frac{L}{\alpha^2}\eta\right])^2\\&+\frac{\zeta+\beta\alpha}{\alpha^2\varepsilon'}\mathcal{N}(\mathfrak{S}_{k,a,b}^{-1},\|\cdot\|_{\mathrm{Fro}},\frac{\alpha\varepsilon'}{\zeta+\beta\alpha})\times\mathcal{N}(S[\mathfrak{S}_{k,a,b}^{-1}],\|\cdot\|_{\mathrm{Fro}},\frac{\alpha^2\varepsilon'}{2(\zeta+\beta\alpha)})\,.\end{aligned} \tag{70}$$

Let $\varepsilon>0$ be fixed. We define $\eta\triangleq\frac{\alpha^2}{4L}\varepsilon$ and $\varepsilon'\triangleq\varepsilon/4$. Then we have $\varepsilon=2\left[\varepsilon'+\frac{L}{\alpha^2}\eta\right]$. With these ε',η and Lemma 2 we obtain

$$\begin{aligned}&\mathcal{N}(\mathfrak{S}_{k,a,b}^{-1},\|\cdot\|_{\mathrm{Fro}},\frac{\eta}{8\beta}\left[\varepsilon'+\frac{L}{\alpha^2}\eta\right])^2=\mathcal{N}(\mathfrak{S}_{k,a,b}^{-1},\|\cdot\|_{\mathrm{Fro}},\frac{\eta}{16\beta}\varepsilon)^2\leq\left(\frac{ed^2}{2k}\right)^{2k}\left(\frac{288\beta\sqrt{d}b}{\eta\varepsilon}\right)^{2(d+k)}\\&\mathcal{N}(\mathfrak{S}_{k,a,b}^{-1},\|\cdot\|_{\mathrm{Fro}},\frac{\alpha\varepsilon}{4(\zeta+\beta\alpha)})\leq\left(\frac{ed^2}{2k}\right)^{k}(\frac{72\sqrt{d}b(\zeta+\beta\alpha)}{\alpha\varepsilon})^{d+k}\\&\mathcal{N}(S[\mathfrak{S}_{k,a,b}^{-1}],\|\cdot\|_{\mathrm{Fro}},\frac{\alpha^2\varepsilon}{8(\zeta+\beta\alpha)})\leq(\frac{ed^2}{4k})^{2k}(\frac{144(\zeta+\beta\alpha)}{\alpha^2\varepsilon})^{d+2k}\,.\end{aligned} \tag{71}$$

Thus, inequality (70) yields

$$\begin{aligned}\mathcal{N}\left(S[\mathfrak{S}_{k,a,b}],\|\cdot\|_{\Lambda},\varepsilon\right)&\overset{(71)}{\leq}\left(\frac{ed^2}{2k}\right)^{2k}\left(\frac{288\beta\sqrt{d}b}{\eta\varepsilon}\right)^{2(d+k)}\\&\quad+\frac{4(\zeta+\beta\alpha)}{\alpha^2\varepsilon}\left(\frac{ed^2}{2k}\right)^{k}\left(\frac{72b\sqrt{d}(\zeta+\beta\alpha)}{\alpha\varepsilon}\right)^{d+k}\times\left(\frac{ed^2}{4k}\right)^{2k}\left(\frac{144(\zeta+\beta\alpha)}{\alpha^2\varepsilon}\right)^{d+2k}\\&\leq\left(\frac{ed^2}{2k}\right)^{4k}\left[\left(\frac{288\beta\sqrt{d}b}{\eta\varepsilon}\right)^{2(d+k)}+\frac{4(\zeta+\beta\alpha)}{\alpha^2\varepsilon}\left(\frac{72b\sqrt{d}(\zeta+\beta\alpha)}{\alpha\varepsilon}\right)^{d+k}\left(\frac{144(\zeta+\beta\alpha)}{\alpha^2\varepsilon}\right)^{d+2k}\right]\end{aligned}$$

Using the definition $\eta=\frac{\alpha^2}{4L}\varepsilon$ we obtain

$$\mathcal{N}\left(S[\mathfrak{S}_{k,a,b}],\|\cdot\|_{\Lambda},\varepsilon\right)\leq\left(\frac{ed^2}{2k}\right)^{4k}\left[\left(\frac{1152\sqrt{d}b\beta L}{\alpha^2\varepsilon^2}\right)^{2(d+k)}+\frac{4(\zeta+\beta\alpha)}{\alpha^2\varepsilon}\left(\frac{72b\sqrt{d}(\zeta+\beta\alpha)}{\alpha\varepsilon}\right)^{d+k}\left(\frac{144(\zeta+\beta\alpha)}{\alpha^2\varepsilon}\right)^{d+2k}\right].$$

Now, from Lemma 1, we have $\beta = \frac{C_{\text{Fro}}}{a^2}, \quad L = \frac{C_{\text{Fro}}}{a^3}, \quad \alpha = \frac{c_{\text{Fro}}}{b^2}, \quad \zeta = 2L = 2\frac{C_{\text{Fro}}}{a^3}$. So,

$$\begin{aligned}\mathcal{N}\left(S[\mathfrak{S}_{k,a,b}], \|\cdot\|_\Lambda, \varepsilon\right) &\leq \left(\frac{ed^2}{2k}\right)^{4k}\left[\left(\frac{1152b\sqrt{d}\frac{C_{\text{Fro}}}{a^2}\frac{C_{\text{Fro}}}{a^3}}{\frac{c_{\text{Fro}}^2}{b^4}\varepsilon^2}\right)^{2(d+k)}\right. \\ &\left. + \frac{4(2\frac{C_{\text{Fro}}}{a^3} + \frac{C_{\text{Fro}}}{a^2}\frac{c_{\text{Fro}}}{b^2})}{\frac{c_{\text{Fro}}^2}{b^4}\varepsilon}\left(\frac{72b\sqrt{d}(2\frac{C_{\text{Fro}}}{a^3} + \frac{C_{\text{Fro}}}{a^2}\frac{c_{\text{Fro}}}{b^2})}{\frac{c_{\text{Fro}}}{b^2}\varepsilon}\right)^{d+k}\left(\frac{144(2\frac{C_{\text{Fro}}}{a^3} + \frac{C_{\text{Fro}}}{a^2}\frac{c_{\text{Fro}}}{b^2})}{\frac{c_{\text{Fro}}^2}{b^4}\varepsilon}\right)^{d+2k}\right] \\ &= \left(\frac{ed^2}{2k}\right)^{4k}\left[\left(\frac{1152\sqrt{d}}{\varepsilon^2}\frac{C_{\text{Fro}}^2}{c_{\text{Fro}}^2}\frac{b^5}{a^5}\right)^{2(d+k)}\right. \\ &\left. + \frac{4}{\varepsilon}\frac{C_{\text{Fro}}}{c_{\text{Fro}}}(\frac{2}{c_{\text{Fro}}}\frac{b^4}{a^3} + \frac{b^2}{a^2})\left(\frac{72}{\varepsilon}\frac{C_{\text{Fro}}}{c_{\text{Fro}}}(2\frac{b^3}{a^3} + \frac{b}{a^2}c_{\text{Fro}})\sqrt{d}\right)^{d+k}\left(\frac{144}{\varepsilon}\frac{C_{\text{Fro}}}{c_{\text{Fro}}}(\frac{2}{c_{\text{Fro}}}\frac{b^4}{a^3} + \frac{b^2}{a^2})\right)^{d+2k}\right].\end{aligned} \tag{72}$$

We will simplify this expression using the homogeneity of the normalized secant set. More precisley if $\frac{\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2}{\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_\Lambda} \in S[\mathfrak{S}_{k,a,b}]$ then for any $t > 0$, $\frac{t\mathbf{\Sigma}_1 - t\mathbf{\Sigma}_2}{\|t\mathbf{\Sigma}_1 - t\mathbf{\Sigma}_2\|_\Lambda} \in S[\mathfrak{S}_{k,a,b}]$. This implies that $\forall t > 0,\ S[\mathfrak{S}_{k,a,b}] = S[\mathfrak{S}_{k,t\cdot a,t\cdot b}]$. In particular for $t \triangleq \frac{a}{b^2}$ the previous expression (72) gives

$$\begin{aligned}\mathcal{N}\left(S[\mathfrak{S}_{k,a,b}], \|\cdot\|_\Lambda, \varepsilon\right) &= \mathcal{N}\left(S[\mathfrak{S}_{k,t\cdot a,t\cdot b}], \|\cdot\|_\Lambda, \varepsilon\right) \\ &\leq \left(\frac{ed^2}{2k}\right)^{4k}\left[\left(\frac{1152\sqrt{d}}{\varepsilon^2}\frac{C_{\text{Fro}}^2}{c_{\text{Fro}}^2}\frac{b^5}{a^5}\right)^{2(d+k)}\right. \\ &\left. + \frac{4}{\varepsilon}\frac{C_{\text{Fro}}}{c_{\text{Fro}}}(\frac{2}{c_{\text{Fro}}}t\frac{b^4}{a^3} + \frac{b^2}{a^2})\left(\frac{72}{\varepsilon}\frac{C_{\text{Fro}}}{c_{\text{Fro}}}(2\frac{b^3}{a^3} + \frac{b}{ta^2}c_{\text{Fro}})\sqrt{d}\right)^{d+k}\left(\frac{144}{\varepsilon}\frac{C_{\text{Fro}}}{c_{\text{Fro}}}(\frac{2}{c_{\text{Fro}}}t\frac{b^4}{a^3} + \frac{b^2}{a^2})\right)^{d+2k}\right] \\ &= \left(\frac{ed^2}{2k}\right)^{4k}\left[\left(\frac{1152\sqrt{d}}{\varepsilon^2}\frac{C_{\text{Fro}}^2}{c_{\text{Fro}}^2}\frac{b^5}{a^5}\right)^{2(d+k)}\right. \\ &\left. + \frac{4}{\varepsilon}\frac{C_{\text{Fro}}}{c_{\text{Fro}}}(\frac{2}{c_{\text{Fro}}} + 1)\frac{b^2}{a^2}\left(\frac{72}{\varepsilon}\frac{C_{\text{Fro}}}{c_{\text{Fro}}}(2 + c_{\text{Fro}})\frac{b^3}{a^3}\sqrt{d}\right)^{d+k}\left(\frac{144}{\varepsilon}\frac{C_{\text{Fro}}}{c_{\text{Fro}}}(\frac{2}{c_{\text{Fro}}} + 1)\frac{b^2}{a^2}\right)^{d+2k}\right] \\ &\leq \left(\frac{ed^2}{2k}\right)^{4k}\left[\left(\frac{1152\sqrt{d}}{\varepsilon^2}\frac{C_{\text{Fro}}^2}{c_{\text{Fro}}^2}\frac{b^5}{a^5}\right)^{2(d+k)}\right. \\ &\left. + \left(\frac{72}{\varepsilon}\frac{C_{\text{Fro}}}{c_{\text{Fro}}}(2 + c_{\text{Fro}})\frac{b^3}{a^3}\sqrt{d}\right)^{d+k}\left(\frac{144}{\varepsilon}\frac{C_{\text{Fro}}}{c_{\text{Fro}}}(\frac{2}{c_{\text{Fro}}} + 1)\frac{b^2}{a^2}\right)^{d+2k+1}\right].\end{aligned}$$

Now, to simplify the expression, remark that $\frac{72}{\varepsilon}\frac{C_{\text{Fro}}}{c_{\text{Fro}}}(2 + c_{\text{Fro}})\frac{b^3}{a^3}\sqrt{d} \geq 1$ as $C_{\text{Fro}}/c_{\text{Fro}} \geq 1$. Therefore we can increase its power from $d + k$ to $d + 2k + 1$ to match with the other multiplicative term. This yields

$$\begin{aligned}\mathcal{N}\left(S[\mathfrak{S}_{k,a,b}], \|\cdot\|_\Lambda, \varepsilon\right) &\leq \left(\frac{ed^2}{2k}\right)^{4k}\left[\left(\frac{1152C_{\text{Fro}}^2\sqrt{d}}{\varepsilon^2 c_{\text{Fro}}^2}\frac{b^5}{a^5}\right)^{2(d+k)} + \left(\frac{72\times 144C_{\text{Fro}}^2\sqrt{d}}{\varepsilon^2 c_{\text{Fro}}}(\frac{2}{c_{\text{Fro}}} + 1)\frac{b^5}{a^5}\right)^{d+2k+1}\right] \\ &= \left(\frac{ed^2}{2k}\right)^{4k}\left[\left(c_0\frac{\sqrt{d}C_{\text{Fro}}^2}{\varepsilon^2 c_{\text{Fro}}^2}\frac{b^5}{a^5}\right)^{2(d+k)} + \left(c_1\frac{\sqrt{d}C_{\text{Fro}}^2}{\varepsilon^2 c_{\text{Fro}}}(\frac{2}{c_{\text{Fro}}} + 1)\frac{b^5}{a^5}\right)^{d+2k+1}\right],\end{aligned}$$

where c_0, c_1 are absolute constants greater than 1. This concludes the proof. □

C. Proofs of Section VI-B

1) Proofs of the rank-one projection operator properties : The goal of this section is to prove Proposition 3 and 4. Before that, we prove several results that will become handy afterwards. Firstly, we state a result that will be usefull to leverage results from the Gaussian case to the uniform case.

Lemma 8. *Let $\mathbf{u}$ and ρ be independent variables with the following distributions : $\mathbf{u} \sim \mathcal{U}(\sqrt{d}\mathbb{S}^{d-1})$ is a uniform vector on the hyper-sphere of radius $\sqrt{d}$ and $\rho^2 \sim \chi^2(d)$ is a chi-square variable with d degrees of freedom. Then, $\rho\frac{\mathbf{u}}{\sqrt{d}} \sim \mathcal{N}(0, \mathbf{I}_d)$ is a standard normal vector.*

Now, a lower bound is derived for the Λ-norm.

Proposition 5. *For any $\mathbf{M} \in S_d(\mathbb{R})$ and $\Lambda \in \{\Lambda_{\mathrm{G}}, \Lambda_{\mathrm{U}}\}$, we have*

$$\|\mathbf{M}\|_{\Lambda} \geq \frac{2}{9\sqrt{15}} \left(\|\mathbf{M}\|_{\mathrm{Fro}} + |\mathrm{tr}(\mathbf{M})|\right) . \tag{73}$$

Proof. We use the fact that for any real random variable X, whenever its fourth moment exists, we have[12]

$$\mathbb{E}\left[|X|\right] \geq \sqrt{\frac{\mathbb{E}\left[X^2\right]^3}{\mathbb{E}\left[X^4\right]}} .$$

First, we focus on the Gaussian case. Using $X = \mathbf{a}^\top \mathbf{M} \mathbf{a} \overset{(d)}{=} \sum \lambda_k b_k^2$, where $\mathbf{a} \sim \mathcal{N}(0, I_d)$, the (λ_k) are the eigenvalues of $\mathbf{M}$ and the (b_k) are *i.i.d* standard normal random variables[13], we can obtain the following bounds :

$$\mathbb{E}\left[X^2\right] = 2\|\mathbf{M}\|_{\mathrm{Fro}}^2 + \mathrm{tr}(\mathbf{M})^2 \geq \frac{2}{3}\left(\|\mathbf{M}\|_{\mathrm{Fro}} + |\mathrm{tr}(\mathbf{M})|\right)^2 ,$$

$$\begin{aligned}
\mathbb{E}\left[X^4\right] = & \sum_i \lambda_i^4 \mathbb{E}\left[b_i^8\right] + \sum_{i\neq j} \lambda_i \lambda_j^3 \mathbb{E}\left[b_i^2\right]\mathbb{E}\left[b_j^6\right] + \sum_{i\neq j} \lambda_i^2 \lambda_j^2 \mathbb{E}\left[b_i^4\right]\mathbb{E}\left[b_j^4\right] \\
& + \sum_{i\neq j\neq k} \lambda_i \lambda_j \lambda_k^2 \mathbb{E}\left[b_i^2\right]\mathbb{E}\left[b_j^2\right]\mathbb{E}\left[b_k^4\right] + \sum_{i\neq j\neq k\neq l} \lambda_i\lambda_j\lambda_k\lambda_l \mathbb{E}\left[b_i^2\right]\mathbb{E}\left[b_j^2\right]\mathbb{E}\left[b_k^2\right]\mathbb{E}\left[b_l^2\right] \\
= & \, 105 \sum_i \lambda_i^4 + 15 \sum_{i\neq j} \lambda_i \lambda_j^3 + 9 \sum_{i\neq j} \lambda_i^2\lambda_j^2 + 3 \sum_{i\neq j\neq k} \lambda_i\lambda_j\lambda_k^2 + \sum_{i\neq j\neq k\neq l} \lambda_i\lambda_j\lambda_k\lambda_l \\
= & \, 90\|\lambda\|_4^4 + 12\mathrm{tr}(\mathbf{M}) \sum_i \lambda_i^3 + 6\|\mathbf{M}\|_{\mathrm{Fro}}^4 + 2\mathrm{tr}(\mathbf{M})^2\|\mathbf{M}\|_{\mathrm{Fro}}^2 + \mathrm{tr}(\mathbf{M})^4 \\
\leq & \, 90\left(\|\mathbf{M}\|_{\mathrm{Fro}} + |\mathrm{tr}(\mathbf{M})|\right)^4 .
\end{aligned}$$

This yields equation (73) for the Gaussian case.

For the uniform case, considering the independent random variables $\mathbf{u} \sim \mathcal{U}(\sqrt{d}\mathbb{S}^{d-1})$, $\rho^2 \sim \chi^2(d)$, from Lemma 8 we have that

$$\mathbb{E}\left[\left|\mathbf{a}^\top\mathbf{M}\mathbf{a}\right|\right] = \mathbb{E}\left[\left|(\rho\frac{1}{\sqrt{d}}\mathbf{u})^\top \mathbf{M} (\rho\frac{1}{\sqrt{d}}\mathbf{u})\right|\right] = \frac{1}{d}\mathbb{E}\left[\rho^2\right]\mathbb{E}\left[\left|\mathbf{u}^\top\mathbf{M}\mathbf{u}\right|\right] = \mathbb{E}\left[\left|\mathbf{u}^\top\mathbf{M}\mathbf{u}\right|\right] .$$

So (73) also holds in the uniform case. □

The proof of Proposition 3 is based on a concentration inequality for subexponential variables. Here, we prove that the variables at play are indeed subexponential by providing an upper-bound on their subexponenital norm. Recall that for a random variable X, its subexponential norm is defined by $\|X\|_{\psi_1} = \inf\{s > 0,\ \mathbb{E}\left[e^{|X|/s}\right] \leq 2\}$.

Proposition 6. *For any $\mathbf{M} \in S_d(\mathbb{R})$ and for $a \sim \mathcal{N}(0, I_d)$ and $u \sim \mathcal{U}(\sqrt{d}\mathbb{S}^{d-1})$, the following controls hold :*

$$\left\||\mathbf{a}^\top\mathbf{M}\mathbf{a}| - \mathbb{E}\left[|\mathbf{a}^\top\mathbf{M}\mathbf{a}|\right]\right\|_{\psi_1} \leq \frac{2}{\log 2}\left(\frac{76}{9}e^2 \cdot \|\mathbf{M}\|_{\mathrm{Fro}} + |\mathrm{tr}(\mathbf{M})|\right), \tag{74}$$

$$\left\||\mathbf{u}^\top\mathbf{M}\mathbf{u}| - \mathbb{E}\left[|\mathbf{u}^\top\mathbf{M}\mathbf{u}|\right]\right\|_{\psi_1} \leq \frac{8e}{\log 2}\left(\frac{76}{9}e^2 \cdot \|\mathbf{M}\|_{\mathrm{Fro}} + |\mathrm{tr}(\mathbf{M})|\right). \tag{75}$$

Proof of Equation (74). This proof revolves around the different characterizations of subexponentiality (indexed from (*a*) to (*e*)) presented in Proposition 2.7.1 of [89]. First from the centering Lemma (see Exercise 2.7.10 in [89]), we have the existence of a constant $C_1 > 0$ such that $\left\||\mathbf{a}^\top\mathbf{M}\mathbf{a}| - \mathbb{E}\left[|\mathbf{a}^\top\mathbf{M}\mathbf{a}|\right]\right\|_{\psi_1} \leq C_1 \left\|\,|\mathbf{a}^\top\mathbf{M}\mathbf{a}|\,\right\|_{\psi_1} = C_1\|\mathbf{a}^\top\mathbf{M}\mathbf{a}\|_{\psi_1}$. Let us denote by X our random variable of interest $X = \mathbf{a}^\top\mathbf{M}\mathbf{a}$, and Y its centered version $Y = X - \mathbb{E}\left[X\right]$. Working with Y, we now characterize the constant $K_5(Y)$ appearing in statement (*e*) of Proposition 2.7.1 in [89]. Remark that the centered χ^2 variables $b_k^2 - 1$ verify (*e*) with a certain constant K_{χ^2}. For $|t| \leq 1/(K_{\chi^2} \cdot \|\mathbf{M}\|_{\mathrm{Fro}})$ we have that

$$\mathbb{E}\left[e^{tY}\right] = \prod_{k=1}^d \mathbb{E}\left[e^{t\lambda_k(b_k^2-1)}\right] \leq \prod_{k=1}^d \mathbb{E}\left[e^{K_{\chi^2}^2\lambda_k^2 t^2}\right] = e^{K_{\chi^2}^2\ \|\mathbf{M}\|_{\mathrm{Fro}}^2 t^2} .$$

This yield $K_5(Y) \leq K_{\chi^2}\ \|\mathbf{M}\|_{\mathrm{Fro}}$. Now, by considering statement (*c*), we have that $K_3(X) \leq K_3(Y) + |\mathbb{E}\left[X\right]| \leq C_{3,5}K_{\chi^2} \cdot \|\mathbf{M}\|_{\mathrm{Fro}} + |\mathrm{tr}(\mathbf{M})|$ (where $C_{3,5}$ is the universal constant allowing to pass from (*c*) to (*e*)). Finally, gathering up the pieces we have

$$\left\||\mathbf{a}^\top\mathbf{M}\mathbf{a}| - \mathbb{E}\left[|\mathbf{a}^\top\mathbf{M}\mathbf{a}|\right]\right\|_{\psi_1} \leq C_1C_{4,3}(C_{3,5}K_{\chi^2} \cdot \|\mathbf{M}\|_{\mathrm{Fro}} + |\mathrm{tr}(\mathbf{M})|) ,$$

[12] It comes from applying the Hölder inequality to $\mathbb{E}\left[|X|^{2/3}|X|^{4/3}\right]$ with $1/p = 2/3$ and $1/q = 1/3$.

[13] The last equality is obtained from the rotation invariance of the standard multivariate normal distribution.

where $C_{4,3}$ is the constant allowing to pass from (*d*) (statement defining the ψ_1-norm) to (*c*). To conclude the proof, it suffices to find the values of the various constants. This is completely general and does not depend on the rank-one projection considered here. The various constants can be set as follows :

$$C_1 = 2\,, \qquad C_{3,5} = 4e^2\,,$$
$$C_{4,3} = \frac{1}{\log 2}\,, \qquad K_{\chi^2} = \frac{19}{9}\,.$$

Let us start by computing C_1. Let X be a subexponential random variable. We have, for any $s > 0$,

$$\mathbb{E}\left[e^{\frac{||X|-\mathbb{E}[|X|]|}{s}}\right] \overset{(\triangle \text{ ineq.})}{\leq} \mathbb{E}\left[e^{\frac{|X|}{s}}\right] e^{\frac{\mathbb{E}[|X|]}{s}} \overset{\text{(Jensen)}}{\leq} \mathbb{E}\left[e^{\frac{|X|}{s}}\right]^2 \overset{\text{(Jensen)}}{\leq} \mathbb{E}\left[e^{\frac{2|X|}{s}}\right]\,.$$

The last term is smaller than 2 for $s/2 \geq \|X\|_{\psi_1}$, so we have that $\||X| - \mathbb{E}\left[|X|\right]\|_{\psi_1} \leq 2\|X\|_{\psi_1}$, yielding $C_1 = 2$.

For the constant $C_{3,5}$ we use that $C_{3.5} \leq C_{3,2}C_{2,5}$. In [89], the value of $C_{2,5}$ is given and equals $2e$. Let us focus on $C_{3,2}$ and assume that $K_2(X) = 1$. For any λ such that $0 \leq \lambda \leq 1/(2e)$, we have the following inequalities

$$\mathbb{E}\left[e^{\lambda|X|}\right] = 1 + \sum_{p\geq 1} \frac{\lambda^p \mathbb{E}\left[|X|^p\right]}{p!} \overset{K_2(X)=1}{\leq} 1 + \sum_{p\geq 1} \frac{\lambda^p p^p}{p!} \leq 1 + \sum_{p\geq 1} \frac{\lambda^p p^p}{(p/e)^p} = \frac{1}{1-\lambda e} \leq e^{2e\lambda}\,.$$

Thus, $K_3(X) \leq 2e$. So we can take $C_{3,5} = 4e^2$.

For $C_{4,3}$ assume that $K_3(X) = 1$. For $\lambda \leq \log 2$ we have

$$\mathbb{E}\left[e^{\lambda|X|}\right] \overset{K_3(X)=1}{\leq} e^{\lambda} \leq 2\,.$$

So we can take $C_{4,3} = 1/\log 2$.

The value of K_{χ^2} can be obtain from the following computation. Let $b \sim \mathcal{N}(0,1)$ be a standard normal distribution, then for $\lambda < 1/2$

$$\mathbb{E}\left[e^{\lambda(b^2-1)}\right] = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}\,.$$

We can show that for $K = \frac{2}{1-x_0}$ where x_0 is the smallest solution of $e^x = e^3 x$, we have for all λ such that $|\lambda| \leq 1/K$, $\mathbb{E}\left[e^{\lambda(b^2-1)}\right] \leq e^{K^2\lambda^2}$. A numerical approximation gives $K \simeq 2.1107\ldots$. So we can take $K_{\chi^2} = 19/9$. This finishes the proof. □

Proof of Equation (75). As in the proof of (74), we start by decentering : $\||\mathbf{u}^\top\mathbf{M}\mathbf{u}| - \mathbb{E}\left[|\mathbf{u}^\top\mathbf{M}\mathbf{u}|\right]\|_{\psi_1} \leq C_1\|\mathbf{u}^\top\mathbf{M}\mathbf{u}\|_{\psi_1}$. Then, recall (see [89]) that there exists an absolute constant $C_{4,2}$ such that $\|\mathbf{u}^\top\mathbf{M}\mathbf{u}\|_{\psi_1} \leq C_{4,2}K_2(\mathbf{u}^\top\mathbf{M}\mathbf{u})$ where $K_2(\mathbf{u}^\top\mathbf{M}\mathbf{u})$ is the smallest constant K such that for all $p \geq 1$, $\mathbb{E}\left[|\mathbf{u}^\top\mathbf{M}\mathbf{u}|^p\right] \leq K^p p^p$. From Lemma 8, we have that

$$\mathbb{E}\left[|\mathbf{a}^\top\mathbf{M}\mathbf{a}|^p\right] = \frac{1}{d^p}\mathbb{E}\left[\rho^{2p}\right]\mathbb{E}\left[|\mathbf{u}^\top\mathbf{M}\mathbf{u}|^p\right] = \frac{d(d+2)\ldots(d+2(p-1))}{d^p}\mathbb{E}\left[|\mathbf{u}^\top\mathbf{M}\mathbf{u}|^p\right] \geq \mathbb{E}\left[|\mathbf{u}^\top\mathbf{M}\mathbf{u}|^p\right].$$

Hence, we have that $K_2(\mathbf{u}^\top\mathbf{M}\mathbf{u}) \leq K_2(\mathbf{a}^\top\mathbf{M}\mathbf{a})$. Then, using (74) and the existence of a constant $C_{2,4}$ such that $K_2(\mathbf{a}^\top\mathbf{M}\mathbf{a}) \leq C_{2,4}\|\mathbf{a}^\top\mathbf{M}\mathbf{a}\|_{\psi_1}$. We have

$$\|\mathbf{u}^\top\mathbf{M}\mathbf{u}\|_{\psi_1} \leq C_{4,2}C_{2,4}\|\mathbf{a}^\top\mathbf{M}\mathbf{a}\|_{\psi_1} \leq C_{4,2}C_{2,4}C_{4,3}(C_{3,5}K_{\chi^2}\cdot\|\mathbf{M}\|_{\mathrm{Fro}} + |\mathrm{tr}(\mathbf{M})|)\,.$$

To finish the proof, we show that $C_{4,2}$ and $C_{2,4}$ can be chosen as

$$C_{4,2} = 2e\,, \qquad C_{2,4} = 2.$$

For $C_{4,2}$, we need to prove that $\|X\|_{\psi_1} \leq C_{4,2}K_2(X)$, for any subexponential variable X. Without loos of generality, we can always assume that $K_2(X) = 1$. Thus, for $s > e$ we have

$$\mathbb{E}\left[e^{|X|/s}\right] = 1 + \sum_{k=1}^{\infty} \frac{\mathbb{E}\left[|X|^k\right]}{k!\ s^k}$$
$$\leq 1 + \sum_{k=1}^{\infty} \frac{k^k}{k! s^k} \overset{(\star)}{\leq} \sum_{k=0}^{\infty} \left(\frac{e}{s}\right)^k = \frac{1}{1-e/s}\,,$$

where the $(\star)$ inequality comes from the Stirling approximation $k! \geq (k/e)^k$. Thus, for $s \geq 2e$ we have $\mathbb{E}\left[e^{|X|/s}\right] \leq 2$. So we can take $C_{4,2} = 2e$.

For $C_{2,4}$, assume that $\|X\|_{\psi_1} = 1$. Then, for any $p \geq 1$,

$$\mathbb{E}\left[|X|^p\right] = \int_0^\infty \mathbb{P}\left(|X|^p > u\right) du = \int_0^\infty \mathbb{P}\left(|X| > t\right) pt^{p-1} dt$$
$$\leq \int_0^\infty \mathbb{E}\left[e^{|X|}\right] e^{-t} pt^{p-1} dt \leq 2\int_0^\infty e^{-t} pt^{p-1} dt = 2p! \leq (2p)^p .$$

Thus, we can take $C_{2,4} = 2$. □

All this previous results allow now to prove Proposition 3, which is recalled below.

Proposition 3. *For any $\mathbf{U} \in S_d(\mathbb{R})$ satisfying $\|\mathbf{U}\|_\Lambda = 1$ and for any $t > 0$, we have*

$$\mathbb{P}\left(\left|\|\mathcal{A}(\mathbf{U})\|_1 - 1\right| > t\right) \leq 2\exp\left(-\frac{m}{8e^2}\min\left(\frac{t^2}{K_\Lambda^2}, \frac{t}{K_\Lambda}\right)\right), \tag{32}$$

where K_Λ is an absolute constant given by $K_\Lambda = \frac{76e^2\sqrt{15}}{\log 2}$ for the Gaussian case and $K_\Lambda = \frac{304e^3\sqrt{15}}{\log 2}$ for the uniform case.

Proof of Proposition 3. From our previous results, we have :

$$\||\mathbf{a}^\top\mathbf{U}\mathbf{a}| - \mathbb{E}\left[|\mathbf{a}^\top\mathbf{U}\mathbf{a}|\right]\|_{\psi_1} \overset{(74)}{\leq} \frac{2\times 76\ e^2}{9\log 2}(\|\mathbf{U}\|_{\mathrm{Fro}} + |\mathrm{tr}(\mathbf{U})|) \overset{(73)}{\leq} \frac{76e^2\sqrt{15}}{\log 2}\|\mathbf{U}\|_\Lambda \overset{\|\mathbf{U}\|_\Lambda=1}{\leq} \frac{76e^2\sqrt{15}}{\log 2}. \tag{76}$$

Similarly, in the uniform case, we obtain $\||\mathbf{u}^\top\mathbf{U}\mathbf{u}| - \mathbb{E}\left[|\mathbf{u}^\top\mathbf{U}\mathbf{u}|\right]\|_{\psi_1} \leq \frac{304e^3\sqrt{15}}{\log 2}$. Therfore, in both cases, the subexponential norm is bounded by an absolute constant that will be denoted by K_Λ in the following. Now, let us recall a Bernstein-type concentration inequality for sum of subexponential variables.

Lemma 9 (Proposition 5.16 [90])**.** *Let $X_1, \ldots, X_m$ be independent centered sub-exponential random variables, and $K = \max_i \|X_i\|_{\psi_1}$. Then for every $\gamma = (\gamma_1, \ldots, \gamma_m) \in \mathbb{R}^m$ and every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^m \gamma_i X_i\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{K^2\|\gamma\|_2^2}, \frac{t}{K\|\gamma\|_\infty}\right)\right),$$

where $c = \frac{1}{8e^2}$. (The value of c can be tracked through the proofs of Lemma 5.15 and Proposition 5.16 in [90].)

Taking $X_i = |\mathbf{a}_i^\top\mathbf{U}\mathbf{a}_i| - \mathbb{E}\left[|\mathbf{a}_i^\top\mathbf{U}\mathbf{a}_i|\right] = |\mathbf{a}_i^\top\mathbf{U}\mathbf{a}_i| - 1$ (or $X_i = |\mathbf{u}_i^\top\mathbf{U}\mathbf{u}_i| - 1$ in the uniform case), and $\gamma_i = 1/m$ for all i, yields

$$\mathbb{P}\left(|\|\mathcal{A}\mathbf{U}\|_1 - 1| > t\right) \leq 2\exp\left(-\frac{m}{8e^2}\min\left(\frac{t^2}{K_\Lambda^2}, \frac{t}{K_\Lambda}\right)\right), \qquad \forall t \geq 0 .$$

□

We proceed by proving Proposition 4.

Proposition 4. *Let $\Lambda \in \{\Lambda_{\mathrm{G}}, \Lambda_{\mathrm{U}}\}$ be either the Gaussian or uniform distribution on $\mathbb{R}^d$ and consider the associated Λ-norm defined in (15). Then,*

$$\forall \mathbf{M} \in S_d(\mathbb{R}),\ \frac{2}{9\sqrt{15}}\|\mathbf{M}\|_{\mathrm{Fro}} \leq \|\mathbf{M}\|_\Lambda \leq \sqrt{d}\|\mathbf{M}\|_{\mathrm{Fro}} .$$

This gives $c_{\mathrm{Fro}} = 2/(9\sqrt{15})$ and $C_{\mathrm{Fro}} = \sqrt{d}$ in Lemma 1.

Proof of Proposition 4. The lower bound is a direct consequence of Proposition 5. Let us prove the upper bound. In the Gaussian case, the following inequalities hold

$$\mathbb{E}\left[|\mathbf{a}^\top\mathbf{M}\mathbf{a}|\right] = \mathbb{E}\left[\left|\sum_{k=1}^d \lambda_k b_k^2\right|\right] \leq \sum_{k=1}^d |\lambda_k|\mathbb{E}\left[b_k^2\right] = \sum_{k=1}^d |\lambda_k| \leq \sqrt{d}\sqrt{\sum_{k=1}^d \lambda_k^2} = \sqrt{d}\|\mathbf{M}\|_{\mathrm{Fro}},$$

where $\mathbf{a}, \mathbf{b} \sim \mathcal{N}(0, \mathbf{I}_d)$. Similarly, in the URO case,

$$\mathbb{E}\left[|\mathbf{u}^\top\mathbf{M}\mathbf{u}|\right] \leq \sum_{k=1}^d |\lambda_k|\mathbb{E}\left[v_k^2\right] = \sum_{k=1}^d |\lambda_k| \leq \sqrt{d}\|\mathbf{M}\|_{\mathrm{Fro}},$$

where $\mathbf{u}, \mathbf{v} \sim \mathcal{U}(\sqrt{d}\mathbb{S}^{d-1})$. □

2) Proof of Theorem 3:

Theorem 3. *Let $\mathcal{A} : (S_d(\mathbb{R}), \|\cdot\|_\Lambda) \to (\mathbb{R}^m, \|\cdot\|_1)$ be a rank-one projection operator as defined in (5), with $(\mathbf{a}_j)_j$ either Gaussian or uniform. For all $\delta, \rho \in]0,1[$, there exists $C = C(\delta, \rho, b/a)$, independent of m, k and d, such that, whenever*

$$m \geq C(d + 2k) \log d \,, \tag{16}$$

the operator $\mathcal{A}$ satisfies RIP_δ on $\mathfrak{S}_{k,a,b}$ with probability at least $1-\rho$. In particular the following holds uniformly on $\mathbf{\Sigma} \in \mathfrak{S}_{k,a,b}$ with probability at least $1-\rho$: Let $\mathbf{x}_1, \cdots, \mathbf{x}_n \sim \mu$ with μ a centered probability distribution with covariance $\mathbf{\Sigma}$, $\widehat{\mathbf{\Sigma}}$ the empirical covariance matrix and $\mathbf{s} = \mathcal{A}(\widehat{\mathbf{\Sigma}})$ a sketch of the data. The estimator $\mathbf{\Sigma}^\star = \Delta[\mathbf{s}]$ defined in (7) satisfies

$$\|\mathbf{\Sigma}^\star - \mathbf{\Sigma}\|_\Lambda \leq \frac{2}{1-\delta} \|\mathcal{A}(\widehat{\mathbf{\Sigma}}) - \mathcal{A}(\mathbf{\Sigma})\|_1 \,. \tag{17}$$

Proof. From Theorem 2 and Remark 2, we know that the probability that the operator $\mathcal{A}$ *does not* satisfy the $\mathrm{RIP}(\delta)$ is upper-bounded by

$$\mathcal{N}(S[\mathfrak{S}_{k,a,b}], \|\cdot\|_\Lambda, \varepsilon) C_2(\frac{\delta}{2}) + \mathcal{N}(B_\Lambda, \|\cdot\|_\Lambda, \varepsilon') C_2((1-\varepsilon')\frac{\delta}{2\varepsilon} - 1), \qquad \forall \varepsilon, \varepsilon' > 0 \,, \tag{77}$$

where $B_\Lambda = \{\mathbf{U} \in S_d(\mathbb{R}) : \|\mathbf{U}\|_\Lambda = 1\}$ and $C_2(t) = 2\exp(-\frac{m}{8e^2}\min(t/K_\Lambda, (t/K_\Lambda)^2))$ (see Proposition 3). In the following, we choose $\varepsilon' = 1/2$. Given ρ, the strategy is to find an ε small enough so that the right handside term is smaller that $\rho/2$. Then, a condition on m will be derived to ensure that the left handside term is also smaller than $\rho/2$.

First of all, notice that, from standard covering argument, $\mathcal{N}(B_\Lambda, \|\cdot\|_\Lambda, 1/2) \leq (3/(1/2))^{d(d+1)/2}$. Therefore, looking at the logarithm of the second term in (77), ε should verify

$$\frac{d(d+1)}{2}\log(6) + \log 2 - \frac{m}{8e^2}\min\left[\frac{1}{K_\Lambda}\left(\frac{\delta}{4\varepsilon} - 1\right), \frac{1}{K_\Lambda^2}\left(\frac{\delta}{4\varepsilon} - 1\right)^2\right] \leq \log(\rho/2) \,. \tag{78}$$

Assuming that $\varepsilon \leq \frac{\delta}{4(K_\Lambda+1)}$ to ensure that the minimum in the above expression is $\frac{1}{K_\Lambda}\left(\frac{\delta}{4\varepsilon} - 1\right)$, (78) is equivalent to

$$\varepsilon \leq \frac{\delta}{4}\left[\frac{8e^2 K_\Lambda}{m}\left(\frac{d(d+1)}{2}\log(6) + \log 2 + \log(2/\rho)\right) + 1\right]^{-1} .$$

In order to remove the dependency in m and to simplify the expression, notice that it is sufficient to take

$$\varepsilon \leq \frac{\delta}{32e^2 K_\Lambda}\left[d^2\log(6) + 4\log(2/\rho)\right]^{-1} \triangleq \varepsilon_0 \,.$$

In the following, we take $\varepsilon = \varepsilon_0$. Remark that it satisfies the assumption below (78). In particular, note that $\varepsilon \leq 1$. We now focus on the first term in (77). Notice that from Corollary 1 and Proposition 4 giving $c_{\mathrm{Fro}} = 2/(9\sqrt{15})$ and $C_{\mathrm{Fro}} = \sqrt{d}$, the covering number is controlled as follows:

$$\begin{aligned}
&\log \mathcal{N}(S[\mathfrak{S}_{k,a,b}], \|\cdot\|_\Lambda, \varepsilon) \\
\leq& 4k\log(\frac{ed^2}{2k}) + \log\left[\left(c_0\frac{\sqrt{d}C_{\mathrm{Fro}}^2}{\varepsilon^2 c_{\mathrm{Fro}}^2}\frac{b^5}{a^5}\right)^{2(d+k)} + \left(c_1\frac{\sqrt{d}C_{\mathrm{Fro}}^2}{\varepsilon^2 c_{\mathrm{Fro}}}(\frac{2}{c_{\mathrm{Fro}}}+1)\frac{b^5}{a^5}\right)^{d+2k+1}\right] \\
\leq& 4k\log(\frac{ed^2}{2k}) + \log\left[\left(\frac{c_0' d^{3/2}}{\varepsilon^2}\frac{b^5}{a^5}\right)^{2(d+k)} + \left(\frac{c_1' d^{3/2}}{\varepsilon^2}\frac{b^5}{a^5}\right)^{d+2k+1}\right] .
\end{aligned}$$

where c_0, c_1, c_0', c_1' are absolute constants greater than 1. As $\frac{c_1' d^{3/2}}{\varepsilon^2}\frac{b^5}{a^5} \geq 1$, $\left(\frac{c_0' d^{3/2}}{\varepsilon^2}\frac{b^5}{a^5}\right)^{2(d+k)} + \left(\frac{c_1' d^{3/2}}{\varepsilon^2}\frac{b^5}{a^5}\right)^{d+2k+1} \leq \left(\frac{c_0' d^{3/2}}{\varepsilon^2}\frac{b^5}{a^5}\right)^{2(d+k)} + \left(\frac{c_1' d^{3/2}}{\varepsilon^2}\frac{b^5}{a^5}\right)^{2(d+k)} \leq \left(\frac{c_0' d^{3/2}}{\varepsilon^2}\frac{b^5}{a^5} + \frac{c_1' d^{3/2}}{\varepsilon^2}\frac{b^5}{a^5}\right)^{2(d+k)}$. This gives

$$\begin{aligned}
\log \mathcal{N}(S[\mathfrak{S}_{k,a,b}], \|\cdot\|_\Lambda, \varepsilon) &\leq 4k\log(\frac{ed^2}{2k}) + 2(d+k)\log\left[(c_0' + c_1')\frac{d^{3/2}}{\varepsilon^2}\frac{b^5}{a^5}\right] \\
&= 4k\log(\frac{ed^2}{2k}) + 4(d+k)\log\left[c_{\frac{b}{a}}\frac{d}{\varepsilon}\right] ,
\end{aligned}$$

where

$$c_{\frac{b}{a}} \triangleq \sqrt{(c_0' + c_1')}\frac{b^{5/2}}{a^{5/2}} \text{ only depends on } b/a \,.$$

Therefore, m needs to verify

$$4k\log(\frac{ed^2}{2k}) + 4(d+k)\log\left(c_{\frac{b}{a}}\frac{d}{\varepsilon}\right) + \log 2 - \frac{m\delta^2}{32e^2K_\Lambda^2} \leq \log(\rho/2)\,,$$

which is equivalent to

$$\frac{m\delta^2}{32e^2K_\Lambda^2} \geq 4k\log(\frac{ed^2}{2k}) + 4(d+k)\log\left(c_{\frac{b}{a}}\frac{d}{\varepsilon}\right) + \log 2 + \log(2/\rho)\,.$$

To simplify the expression, we derive a sufficient condition on m given by

$$m \geq \frac{32e^2K_\Lambda^2}{\delta^2}\left[4k\log(\frac{ed^2}{2k}) + 4(d+k)\log\left(\frac{32e^2c_{\frac{b}{a}}K_\Lambda}{\delta}d\left[d^2\log(6) + 4\log(2/\rho)\right]\right) + 2\log(2/\rho)\right]\,. \tag{79}$$

This finishes the proof as we can find a constant $C = C(\delta, \rho, b/a)$ such that $m \geq C(d+2k)\log d$ implies (79). □

D. Connection with Bregman proximal gradient

The iterations of our algorithm (24) can be related to the iterations of Bregman Proximal Gradient (BPG). Originally introduced in [72], BPG is a generalization of the classical proximal gradient method in which the proximal operator is replaced with a Bregman proximal operator. It aims at solving problems of the form $\min f+g$ where f, g are proper, convex and lower semi-continuous. For $\lambda > 0$, we consider the optimization problem

$$\min_{\boldsymbol{\Theta}\succ 0} F(\boldsymbol{\Theta}) + \lambda\|\boldsymbol{\Theta}\|_{1,\text{off}} \text{ where } F(\boldsymbol{\Theta}) \triangleq \frac{1}{2}\|\mathcal{A}(\boldsymbol{\Theta}^{-1}) - \mathbf{s}\|_2^2 = f(\boldsymbol{\Theta}^{-1})\,. \tag{80}$$

Interestingly the function F is convex on S_d^{++} so that (80) is a convex optimization problem. Indeed as shown in [66, Example 3.4] the function $(\mathbf{a}, \boldsymbol{\Theta}) \to \mathbf{a}^\top\boldsymbol{\Theta}^{-1}\mathbf{a}$ is jointly convex on $\mathbb{R}^d \times S_d^{++}(\mathbb{R})$. Consequently, $F(\boldsymbol{\Theta}) = \frac{1}{2}\sum_{j=1}^m(\mathbf{a}_j^\top\boldsymbol{\Theta}^{-1}\mathbf{a}_j - s_j)^2$ is also convex as the sum of convex functions. For a fixed step-size $\gamma > 0$ the BPG iterations with the Bregman divergence (22) are given by

$$\boldsymbol{\Theta}_{t+1} = \operatorname*{argmin}_{\boldsymbol{\Theta}\succ 0} \langle\gamma\nabla F(\boldsymbol{\Theta}_t), \boldsymbol{\Theta}\rangle + D_h(\boldsymbol{\Theta}|\boldsymbol{\Theta}_t) + \lambda\gamma\|\boldsymbol{\Theta}\|_{1,\text{off}}\,. \tag{81}$$

By expressing the divergence D_h and using that $\nabla F(\boldsymbol{\Theta}_t) = -\boldsymbol{\Theta}_t^{-1}\nabla f(\boldsymbol{\Theta}_t^{-1})\boldsymbol{\Theta}_t^{-1}$ these iterations are equivalent to

$$\boldsymbol{\Theta}_{t+1} = \operatorname*{argmin}_{\boldsymbol{\Theta}\succ 0} \langle\boldsymbol{\Theta}_t^{-1} - \gamma\boldsymbol{\Theta}_t^{-1}\nabla f(\boldsymbol{\Theta}_t^{-1})\boldsymbol{\Theta}_t^{-1}, \boldsymbol{\Theta}\rangle - \log\det\boldsymbol{\Theta} + \lambda\gamma\|\boldsymbol{\Theta}\|_{1,\text{off}}\,. \tag{82}$$

These iterations also correspond to a graphical lasso since (82) rewrites as $\boldsymbol{\Theta}_{t+1}^{-1} = \text{GLASSO}_{\lambda\gamma}[\boldsymbol{\Theta}_t^{-1} - \gamma\boldsymbol{\Theta}_t^{-1}\nabla f(\boldsymbol{\Theta}_t^{-1})\boldsymbol{\Theta}_t^{-1}]$. With the change of variable $\boldsymbol{\Sigma}_t = \boldsymbol{\Theta}_t^{-1}$, the BPG iterations for solving (80) equivalently write

$$\boldsymbol{\Sigma}_{t+1} = \text{GLASSO}_{\lambda\gamma}[\boldsymbol{\Sigma}_t - \gamma\boldsymbol{\Sigma}_t\nabla f(\boldsymbol{\Sigma}_t)\boldsymbol{\Sigma}_t]\,. \tag{83}$$

We can notice that these iterations are very similar to the one in (24) but with $\boldsymbol{\Sigma}_t\nabla f(\boldsymbol{\Sigma}_t)\boldsymbol{\Sigma}_t$ instead of $\nabla f(\boldsymbol{\Sigma}_t)$. In fact, the iterations (24) are equivalent to the BPG iterations (83) when using a Riemannian gradient instead of a Euclidean one. More precisely, when considering for $\mathbf{X} \succ 0$, the inner product $\langle\mathbf{U}, \mathbf{V}\rangle_{\mathbf{X}} \triangleq \text{tr}(\mathbf{U}\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1})$ and computing the gradient w.r.t. $\langle\cdot,\cdot\rangle_{\mathbf{X}}$ at $\mathbf{X}$ we get the formula $\text{grad}\, f(\mathbf{X}) = \mathbf{X}\nabla f(\mathbf{X})\mathbf{X}$ [91]. This corresponds to endowing the space $S_d^{++}(\mathbb{R})$ with the affine-invariant geometry [92, Section 11.7]. In conclusion, if we consider our iterations (24) with the Riemannian gradient $\text{grad}\, f$ instead of the Euclidean gradient we get the BPG iterations (83). However, we observe in practice that the algorithm with the BPG has degraded performance compared to the one proposed in (24).

E. Safe step-size strategy

The goal of this section is to provide a step-size $\gamma > 0$ ensuring that the matrix $\boldsymbol{\Sigma}_{t+\frac{1}{2}} := \boldsymbol{\Sigma}_k - \gamma\nabla f(\boldsymbol{\Sigma}_k)$ remains positive definite during the iterations. Recall that $\nabla f(\boldsymbol{\Sigma}_t) = \mathcal{A}^\star(\mathcal{A}(\boldsymbol{\Sigma}_t) - \mathbf{s})$ where $\mathbf{s} = \frac{1}{n}\sum_{i=1}^n\Phi(\mathbf{x}_i) = \frac{1}{m}\left(\mathbf{a}_1^\top\widehat{\boldsymbol{\Sigma}}\mathbf{a}_1, \cdots, \mathbf{a}_m^\top\widehat{\boldsymbol{\Sigma}}\mathbf{a}_m\right)$ is the sketch of the data and $\widehat{\boldsymbol{\Sigma}}$ is the empirical covariance matrix. The adjoint operator $\mathcal{A}^\star$ is given by $\mathbf{y} \to \mathcal{A}^\star(\mathbf{y}) = \frac{1}{m}\sum_{j=1}^m y_j\mathbf{a}_j\mathbf{a}_j^\top$.

The matrix $\boldsymbol{\Sigma}_{t+\frac{1}{2}}$ is positive definite when $\lambda_{\min}(\boldsymbol{\Sigma}_t - \gamma\nabla f(\boldsymbol{\Sigma}_t)) > 0$ that is when

$$\lambda_{\min}\left(\boldsymbol{\Sigma}_t - \gamma\frac{1}{m}\sum_{j=1}^m(\frac{1}{m}\mathbf{a}_j^\top\boldsymbol{\Sigma}_t\mathbf{a}_j - s_j)\mathbf{a}_j\mathbf{a}_j^\top\right) > 0\,. \tag{84}$$

Moreover we have,

$$\begin{aligned}\lambda_{\min}\left(\boldsymbol{\Sigma}_t-\gamma\frac{1}{m}\sum_{j=1}^{m}(\frac{1}{m}\mathbf{a}_j^\top\boldsymbol{\Sigma}_t\mathbf{a}_j-s_j)\mathbf{a}_j\mathbf{a}_j^\top\right)&\geq\lambda_{\min}(\boldsymbol{\Sigma}_t)-\gamma\frac{1}{m}\lambda_{\max}\left(\sum_{j=1}^{m}(\frac{1}{m}\mathbf{a}_j^\top\boldsymbol{\Sigma}_t\mathbf{a}_j-s_j)\mathbf{a}_j\mathbf{a}_j^\top\right)\\&=\lambda_{\min}(\boldsymbol{\Sigma}_t)-\gamma\frac{1}{m^2}\lambda_{\max}\left(\sum_{j=1}^{m}[\mathbf{a}_j^\top(\boldsymbol{\Sigma}_t-\widehat{\boldsymbol{\Sigma}})\mathbf{a}_j]\mathbf{a}_j\mathbf{a}_j^\top\right).\end{aligned}\tag{85}$$

Using $\forall j\in[\![m]\!],\mathbf{a}_j(\boldsymbol{\Sigma}_t-\widehat{\boldsymbol{\Sigma}})\mathbf{a}_j\leq\lambda_{\max}(\boldsymbol{\Sigma}_t-\widehat{\boldsymbol{\Sigma}})\|\mathbf{a}_j\|_2^2$ we have, for any $\mathbf{z}\in\mathbb{R}^d,\|\mathbf{z}\|_2=1$,

$$\mathbf{z}^\top(\sum_{j=1}^{m}[\mathbf{a}_j^\top(\boldsymbol{\Sigma}_t-\widehat{\boldsymbol{\Sigma}})\mathbf{a}_j]\mathbf{a}_j\mathbf{a}_j^\top)\mathbf{z}=\sum_{j=1}^{m}[\mathbf{a}_j^\top(\boldsymbol{\Sigma}_t-\widehat{\boldsymbol{\Sigma}})\mathbf{a}_j]|\mathbf{z}^\top\mathbf{a}_j|^2\leq\lambda_{\max}(\boldsymbol{\Sigma}_t-\widehat{\boldsymbol{\Sigma}})\sum_{j=1}^{m}\|\mathbf{a}_j\|_2^2|\mathbf{z}^\top\mathbf{a}_j|^2\,.\tag{86}$$

By introducing the matrix $\mathbf{A}=(\mathbf{a}_1,\cdots,\mathbf{a}_m)\in\mathbb{R}^{d\times m}$ the previous inequality leads to

$$\begin{aligned}\mathbf{z}^\top(\sum_{j=1}^{m}[\mathbf{a}_j^\top(\boldsymbol{\Sigma}_t-\widehat{\boldsymbol{\Sigma}})\mathbf{a}_j]\mathbf{a}_j\mathbf{a}_j^\top)\mathbf{z}&\leq\lambda_{\max}(\boldsymbol{\Sigma}_k-\widehat{\boldsymbol{\Sigma}})\max_{j\in[\![m]\!]}\|\mathbf{a}_j\|_2^2\sum_{j=1}^{m}|\mathbf{z}^\top\mathbf{a}_j|^2=\lambda_{\max}(\boldsymbol{\Sigma}_t-\widehat{\boldsymbol{\Sigma}})(\max_{j\in[\![m]\!]}\|\mathbf{a}_j\|_2^2)\;\mathbf{z}^\top\mathbf{A}\mathbf{A}^\top\mathbf{z}\\&\leq\lambda_{\max}(\boldsymbol{\Sigma}_t-\widehat{\boldsymbol{\Sigma}})(\max_{j\in[\![m]\!]}\|\mathbf{a}_j\|_2^2)\lambda_{\max}(\mathbf{A}\mathbf{A}^\top)\\&\leq[\lambda_{\max}(\boldsymbol{\Sigma}_t)-\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})](\max_{j\in[\![m]\!]}\|\mathbf{a}_j\|_2^2)\lambda_{\max}(\mathbf{A}\mathbf{A}^\top)\,.\end{aligned}\tag{87}$$

Consequently,

$$\lambda_{\max}\left(\sum_{j=1}^{m}[\mathbf{a}_j^\top(\boldsymbol{\Sigma}_t-\widehat{\boldsymbol{\Sigma}})\mathbf{a}_j]\mathbf{a}_j\mathbf{a}_j^\top\right)\leq[\lambda_{\max}(\boldsymbol{\Sigma}_t)-\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})](\max_{j\in[\![m]\!]}\|\mathbf{a}_j\|_2^2)\lambda_{\max}(\mathbf{A}\mathbf{A}^\top)\,.\tag{88}$$

This shows that if $\lambda_{\max}(\boldsymbol{\Sigma}_t)<\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})$ then the condition (84) is valid for any $\gamma>0$ since in this case $\lambda_{\max}\left(\sum_{j=1}^{m}[\mathbf{a}_j^\top(\boldsymbol{\Sigma}_t-\widehat{\boldsymbol{\Sigma}})\mathbf{a}_j]\mathbf{a}_j\mathbf{a}_j^\top\right)<0$. On the other hand if $\lambda_{\max}(\boldsymbol{\Sigma}_k)>\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})$ then, by (85) and (88), a step-size

$$\gamma<\frac{m^2}{\max_{j\in[\![m]\!]}\|\mathbf{a}_j\|_2^2\;\sigma_{\max}^2(\mathbf{A})}\times\frac{\lambda_{\min}(\boldsymbol{\Sigma}_t)}{\lambda_{\max}(\boldsymbol{\Sigma}_t)-\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}\,,\tag{89}$$

where $\sigma_{\max}(\mathbf{A})$ is the largest singular value of $\mathbf{A}$, ensures that $\boldsymbol{\Sigma}_{k+\frac{1}{2}}$ is positive definite. Overall a safe step-size strategy is given by

$$\gamma\in\begin{cases}(0,+\infty[& \text{if } \lambda_{\max}(\boldsymbol{\Sigma}_t)<\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})\,,\\(0,\frac{m^2}{\max_{j\in[\![m]\!]}\|\mathbf{a}_j\|_2^2\;\sigma_{\max}^2(\mathbf{A})}\frac{\lambda_{\min}(\boldsymbol{\Sigma}_t)}{\lambda_{\max}(\boldsymbol{\Sigma}_t)-\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}) & \text{if } \lambda_{\max}(\boldsymbol{\Sigma}_t)>\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})\,.\end{cases}\tag{90}$$

We emphasize that this strategy is quite conservative, and requires the computation of both the maximum and minimum eigenvalues of $\boldsymbol{\Sigma}_t$ at every iteration. In practical scenarios, we find that searching for γ in $\{1e{-}3,1e{-}2,1e{-}1\}$ is adequate for achieving convergence in our experimental setups.

F. Covering number of a model set with a condition number constraint.

We would like to consider a model set of covariance matrices that does not restrict their spectra but rather their condition numbers. For a square matrix $\mathbf{M}$, the condition number is defined as $\kappa(\mathbf{M})\triangleq\|\mathbf{M}\|_{2\to2}\|\mathbf{M}^{-1}\|_{2\to2}$. In the special case of positive-definite matrices, it is the ratio between the largest and smallest eigenvalues. For some $\kappa_0\geq1$, we consider the following model set

$$\mathfrak{S}_{k,\kappa_0}\triangleq\left\{\boldsymbol{\Sigma}\in S_d^{++}(\mathbb{R}):\boldsymbol{\Theta}=\boldsymbol{\Sigma}^{-1}\succ0,\|\boldsymbol{\Theta}\|_0\leq d+2k,\kappa(\boldsymbol{\Theta})\leq\kappa_0\right\}.$$

Remark that the above definition implies that $\mathfrak{S}_{k,a,b}\subset\mathfrak{S}_{k,\kappa_0}$, for all $a,b>0$ such that $b/a\leq\kappa_0$. In particular, $\mathfrak{S}_{k,1/\kappa_0,1}\subset\mathfrak{S}_{k,\kappa_0}$. However, $\mathfrak{S}_{k,\kappa_0}$ is a much "bigger" set than sets like $\mathfrak{S}_{k,a,b}$, as the latter are bounded sets while the former is not. Interestingly enough, we are able to upper-bound the covering number of the normalized secant of $\mathfrak{S}_{k,\kappa_0}$ that is

$$S[\mathfrak{S}_{k,\kappa_0}]=\left\{\frac{\boldsymbol{\Theta}_1^{-1}-\boldsymbol{\Theta}_2^{-1}}{\|\boldsymbol{\Theta}_1^{-1}-\boldsymbol{\Theta}_2^{-1}\|_\Lambda}:\;(\boldsymbol{\Theta}_1^{-1},\boldsymbol{\Theta}_2^{-1})\in\mathfrak{S}_{k,\kappa_0}^2,\;\|\boldsymbol{\Theta}_1^{-1}-\boldsymbol{\Theta}_2^{-1}\|_\Lambda>0\right\}.$$

The key element to obtain this bound is that we are able to rewrite $S[\mathfrak{S}_{k,\kappa_0}]$ in term of matrices that are in

$$\mathfrak{S}_0\triangleq\left\{\boldsymbol{\Sigma}\in S_d^{++}(\mathbb{R}):\boldsymbol{\Theta}=\boldsymbol{\Sigma}^{-1}\succ0,\|\boldsymbol{\Theta}\|_0\leq d+2k,\;\mathrm{spec}(\boldsymbol{\Theta})\subset[1/\kappa_0,1],\;\|\boldsymbol{\Theta}\|_{2\to2}=1\right\}.$$

Indeed, for an element $\frac{\tilde{\mathbf{\Theta}}_1^{-1}-\tilde{\mathbf{\Theta}}^{-1}}{\|\tilde{\mathbf{\Theta}}^{-1}-\tilde{\mathbf{\Theta}}^{-1}\|_\Lambda} \in S[\mathfrak{S}_{k,\kappa_0}]$, notice that for any $\lambda > 0$ the matrix $\frac{\lambda\tilde{\mathbf{\Theta}}_1^{-1}-\lambda\tilde{\mathbf{\Theta}}^{-1}}{\|\lambda\tilde{\mathbf{\Theta}}^{-1}-\lambda\tilde{\mathbf{\Theta}}^{-1}\|_\Lambda}$ still is on $S[\mathfrak{S}_{k,\kappa_0}]$. This implies that we can normalize the matrices involved in the secant. More precisely, by choosing $\lambda = \|\tilde{\mathbf{\Theta}}_1\|_{2\to 2}$, setting $\mathbf{\Theta}_1 = \lambda^{-1}\tilde{\mathbf{\Theta}}_1$, $\mathbf{\Theta}_2 = \tilde{\mathbf{\Theta}}_2/\|\tilde{\mathbf{\Theta}}_2\|_{2\to 2}$ and $\tau = \|\tilde{\mathbf{\Theta}}_1\|_{2\to 2}/\|\tilde{\mathbf{\Theta}}_2\|_{2\to 2}$, we have $\mathbf{\Theta}_1, \mathbf{\Theta}_2 \in \mathfrak{S}_0$, $\tau > 0$ and $\frac{\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}}{\|\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}\|_\Lambda} \in S[\mathfrak{S}_{k,\kappa_0}]$. This shows that

$$S[\mathfrak{S}_{k,\kappa_0}] = \left\{ \frac{\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}}{\|\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}\|_\Lambda} : \ \tau > 0, \ (\mathbf{\Theta}_1^{-1}, \mathbf{\Theta}_2^{-1}) \in \mathfrak{S}_0^2, \ \|\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}\|_\Lambda > 0 \right\}.$$

Now, up to exchanging the role of $\tilde{\mathbf{\Theta}}_1$ and $\tilde{\mathbf{\Theta}}_2$, which result in changing the sign of $\frac{\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}}{\|\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}\|_\Lambda}$, we can always assume that $\|\tilde{\mathbf{\Theta}}_1\|_{2\to 2} \leq \|\tilde{\mathbf{\Theta}}_2\|_{2\to 2}$, meaning that $0 < \tau \leq 1$, which yields

$$\begin{aligned} S[\mathfrak{S}_{k,\kappa_0}] &= \left\{ \pm\frac{\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}}{\|\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}\|_\Lambda} : \ \tau \in (0,1], \ (\mathbf{\Theta}_1^{-1}, \mathbf{\Theta}_2^{-1}) \in \mathfrak{S}_0^2 : \ \|\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}\|_\Lambda > 0 \right\} \\ &= \overline{S}[\mathfrak{S}_{k,\kappa_0}] \ \cup \ (-\overline{S}[\mathfrak{S}_{k,\kappa_0}]), \end{aligned}$$

where

$$\overline{S}[\mathfrak{S}_{k,\kappa_0}] \triangleq \left\{ \frac{\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}}{\|\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}\|_\Lambda} : \ \tau \in (0,1], \ (\mathbf{\Theta}_1^{-1}, \mathbf{\Theta}_2^{-1}) \in \mathfrak{S}_0^2, \ \|\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}\|_\Lambda > 0 \right\}.$$

Remark that for any $\varepsilon > 0$, $\mathcal{N}(S[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, \varepsilon) \leq 2\mathcal{N}(\overline{S}[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, \varepsilon)$, therefore we only have to control the covering number of $\overline{S}[\mathfrak{S}_{k,\kappa_0}]$. To do so, we follow the same line of proof as in the control of $S[\mathfrak{S}_{k,a,b}]$ by splitting our set of interests into long and short chords. The analysis of these chords is similar, although more technical and more computation-heavy. A slight complication in this new setting is the need to ensure that τ is bounded away from zero in the case of short chords. In the following, we briefly detail the analysis of the long and short chords given for some $\eta > 0$ by

$$\begin{aligned} \overline{S}_\eta^+[\mathfrak{S}_{k,\kappa_0}] &\triangleq \left\{ \frac{\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}}{\|\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}\|_\Lambda} : \ \tau \in (0,1], \ (\mathbf{\Theta}_1^{-1}, \mathbf{\Theta}_2^{-1}) \in \mathfrak{S}_0^2, \ \|\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}\|_\Lambda > \eta \right\}, \\ \overline{S}_\eta^-[\mathfrak{S}_{k,\kappa_0}] &\triangleq \left\{ \frac{\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}}{\|\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}\|_\Lambda} : \ \tau \in (0,1], \ (\mathbf{\Theta}_1^{-1}, \mathbf{\Theta}_2^{-1}) \in \mathfrak{S}_0^2, \ 0 < \|\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}\|_\Lambda \leq \eta \right\}. \end{aligned}$$

1) Control of the long chords: The strategy to control the covering number of long chords is to express $\overline{S}_\eta^+[\mathfrak{S}_{k,\kappa_0}]$ as the image of a Lipschitz-continuous function and control the covering number of the original set. Let us consider the set $\overline{\mathfrak{X}}_\eta \triangleq \{(\tau, \mathbf{\Theta}_1, \mathbf{\Theta}_2) \in (0,1] \times \mathfrak{S}_0^{-1} \times \mathfrak{S}_0^{-1}, \ \|\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}\|_\Lambda > \eta\}$ equipped with the norm $\|(\tau, \mathbf{M}_1, \mathbf{M}_2)\|_\otimes = |\tau| + \|\mathbf{M}_1\|_{\mathrm{Fro}} + \|\mathbf{M}_2\|_{\mathrm{Fro}}$. Then we have the following lemma. For the sake of conciseness, its proof is not provided, but it is based on the one of Proposition 1 presented in Appendix B3.

Lemma 10. *Let $g : (\overline{\mathfrak{X}}_\eta, \|\cdot\|_\otimes) \to (\overline{S}_\eta^+[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda)$ be the function defined by*

$$g(\tau, \mathbf{\Theta}_1, \mathbf{\Theta}_2) \triangleq \frac{\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}}{\|\mathbf{\Theta}_1^{-1}-\tau\mathbf{\Theta}_2^{-1}\|_\Lambda}.$$

Then, g is surjective and L_0/η-lipschitz continuous with $L_0 = 2C_{\mathrm{Fro}}\kappa_0^2\sqrt{d}$.

As a consequence, we can control the covering number of $\overline{S}_\eta^+[\mathfrak{S}_{k,\kappa_0}]$ using the one of $\overline{\mathfrak{X}}_\eta$ which is easier to handle. This yields the following proposition which proof is also based on the one of Proposition 1.

Proposition 7. *For all $\varepsilon > 0$ and $\eta > 0$, we have*

$$\mathcal{N}(\overline{S}_\eta^+[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, \varepsilon) \leq \mathcal{N}((0,1], |\cdot|, \frac{\eta\varepsilon}{6L_0}) \times \mathcal{N}(\mathfrak{S}_0^{-1}, \|\cdot\|_{\mathrm{Fro}}, \frac{\eta\varepsilon}{6L_0})^2.$$

Note that $\mathcal{N}((0,1], |\cdot|, \varepsilon)$ is bounded by ε^{-1} and the control of the covering of $\mathfrak{S}_0^{-1}$ will be provided later by Lemma 13.

2) Control of the short chords: In order to control the covering number of the short chords we follow these two steps: 1) we show that any element of $\overline{S}_\eta^-[\mathfrak{S}_{k,\kappa_0}]$ is close to an element of a certain "tangent space" 2) we will control the covering number of this space.

Assumption: In all of this section we will assume that $0 < \eta \leq c_{\mathrm{Fro}}/2$. This requirement will be useful for various simplification and will be met when we calibrate η for a good balance between the covering numbers of both long and short chords.

Point 1) is done through the following lemma which proof can be found in Appendix G1.

Lemma 11. *For $\eta > 0$, consider the set of short chords $\overline{S}_\eta^-[\mathfrak{S}_{k,\kappa_0}]$ and the normalized secant set $S[\mathfrak{S}_{k,\kappa_0}^{-1}]$. Define $C \triangleq \{\lambda \mathbf{V} : \ \lambda \in]0, \lambda_0], \mathbf{V} \in S[\mathfrak{S}_{k,\kappa_0}^{-1}]\}$ and*

$$T_C \triangleq \left\{ \mathrm{D}\,\mathrm{inv}_{\tilde{\mathbf{\Theta}}}(\mathbf{C}) : (\tilde{\mathbf{\Theta}}, \mathbf{C}) \in \mathfrak{S}_{k,\frac{1}{\kappa_0},2}^{-1} \times C \right\},$$

with $\lambda_0 \triangleq \frac{2}{c_{\mathrm{Fro}}}$. Defining $Z_0 \triangleq \frac{C_{\mathrm{Fro}}\kappa_0^3}{c_{\mathrm{Fro}}^2}$, we have

$$\forall \mathbf{U} \in \overline{S}_\eta^-[\mathfrak{S}_{k,\kappa_0}], \exists \mathbf{T} \in \mathcal{T}, \ \|\mathbf{U} - \mathbf{T}\|_\Lambda \leq Z_0 \eta \,.$$

As T_C is a good approximation of $\overline{S}_\eta^-[\mathfrak{S}_{k,\kappa_0}]$, we can bound the covering number of the latter by the covering of the former (with a differnt scale), see Lemma 3. Hence, we need to control the covering number of T_C.

Lemma 12. *For all $\varepsilon > 0$, we have*

$$\mathcal{N}(T_C, \|\cdot\|_\Lambda, \varepsilon) \leq \mathcal{N}(\mathfrak{S}_{k,\frac{1}{\kappa_0},2}^{-1}, \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon}{C_0}) \times \mathcal{N}(C, \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon}{C_0}), \tag{91}$$

with $C_0 = C_{\mathrm{Fro}}(2\kappa_0^3\lambda_0 + \kappa_0^2)$.

Now, combining the above results, we are able to provide a control of the covering number of the short chords. See Appendix G1 for the proof.

Proposition 8 (Similar to Proposition 2)**.** *For any $\varepsilon > 0$ and $\eta > 0$, we have*

$$\mathcal{N}(\overline{S}_\eta^-[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, 2(\varepsilon + Z_0\eta)) \leq 2\lambda_0 C_0 \varepsilon^{-1} \mathcal{N}(\mathfrak{S}_{k,\frac{1}{\kappa_0},2}^{-1}, \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon}{C_0}) \times \mathcal{N}(S[\mathfrak{S}_{k,\kappa_0}^{-1}], \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon}{2\lambda_0 C_0}),$$

where $Z_0 = \frac{C_{\mathrm{Fro}}\kappa_0^3}{c_{\mathrm{Fro}}^2}$, $\lambda_0 = \frac{2}{c_{\mathrm{Fro}}}$, $C_0 = C_{\mathrm{Fro}}(2\kappa_0^3\lambda_0 + \kappa_0^2)$.

3) Combining the results: To finish this section and obtain the covering of $\overline{S}_\eta[\mathfrak{S}_{k,\kappa_0}]$, we need the control of the covering of $\mathfrak{S}_0^{-1}$, $\mathfrak{S}_{k,\frac{1}{\kappa_0},2}^{-1}$ and $S[\mathfrak{S}_{k,\kappa_0}^{-1}]$ as they appear in the control of the covering number of the long and short chords. This is done in the following lemma which proof can be found in Appendix G2.

Lemma 13. *For any $\varepsilon > 0$, and $0 < a \leq b$, we have*

$$\mathcal{N}(\mathfrak{S}_0^{-1}, \|\cdot\|_{\mathrm{Fro}}, \varepsilon) \leq \left(\frac{ed^2}{2k}\right)^k \left(\frac{18\sqrt{d}}{\varepsilon}\right)^{d+k},$$

$$\mathcal{N}(\mathfrak{S}_{k,\frac{1}{\kappa_0},2}^{-1}, \|\cdot\|_{\mathrm{Fro}}, \varepsilon) \leq \left(\frac{ed^2)}{2k}\right)^k \left(\frac{2 \times 18\sqrt{d}}{\varepsilon}\right)^{d+k},$$

$$\mathcal{N}(S[\mathfrak{S}_{k,\kappa_0}^{-1}], \|\cdot\|_{\mathrm{Fro}}, \varepsilon) \leq \left(\frac{ed^2}{4k}\right)^{2k} \left(\frac{18}{\varepsilon}\right)^{d+2k}.$$

Gathering up all the pieces, we obtain the following theorem.

Theorem 5. *There exist absolute constants $\tilde{c_1}$ and $\tilde{c_2}$ such that for any ε such that $0 < \varepsilon \leq \frac{2\kappa_0^3}{c_{\mathrm{Fro}}}\sqrt{d}$, we have*

$$\mathcal{N}(S_\eta[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, \varepsilon) \leq 2\left(\frac{ed^2}{2k}\right)^{3k} \left[\left(\frac{\tilde{c_1}\kappa_0^5 d^2}{\varepsilon^2}\right)^{2(d+k)+1} + \left(\frac{\tilde{c_2}\kappa_0^3\sqrt{d}}{\varepsilon}\right)^{2d+3k+1}\right].$$

See Appendix G3 for the proof.

G. Proof for the coverings with a condition number hypothesis

Before diving into the control of the covering numbers of the long and short chords, let us claim various inequalities related to the inverse function on matrices that will be useful in the following.

Lemma 14 (Inverse function properties)**.** *Assume that there exist constants c_{Fro} and C_{Fro} such that $c_{\mathrm{Fro}}\|\mathbf{M}\|_{\mathrm{Fro}} \leq \|\mathbf{M}\|_{\Lambda} \leq C_{\mathrm{Fro}}\|\mathbf{M}\|_{\mathrm{Fro}}$, for all $\mathbf{M} \in S_d$. Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be two matrices in S_d^{++}. Then we have the following inequalities:*

$$\|\mathbf{M}_1^{-1} - \mathbf{M}_2^{-1}\|_{\Lambda} \leq C_{\mathrm{Fro}}\ \|\mathbf{M}_1^{-1}\|_{2\to 2}\ \|\mathbf{M}_2^{-1}\|_{2\to 2}\ \|\mathbf{M}_1 - \mathbf{M}_2\|_{\mathrm{Fro}}\,, \tag{92}$$

$$\|\mathbf{M}_1 - \mathbf{M}_2\|_{\mathrm{Fro}} \leq \frac{1}{c_{\mathrm{Fro}}}\ \|\mathbf{M}_1\|_{2\to 2}\ \|\mathbf{M}_2\|_{2\to 2}\ \|\mathbf{M}_1^{-1} - \mathbf{M}_2^{-1}\|_{\Lambda}\,, \tag{93}$$

$$\|\mathbf{M}_1^{-1} - \mathbf{M}_2^{-1} - \mathrm{D}\,\mathrm{inv}_{\mathbf{M}_2}(\mathbf{M}_1 - \mathbf{M}_2)\|_{\Lambda} \leq C_{\mathrm{Fro}}\ \|\mathbf{M}_1^{-1}\|_{2\to 2}\ \|\mathbf{M}_2^{-1}\|_{2\to 2}^2\ \|\mathbf{M}_1 - \mathbf{M}_2\|_{\mathrm{Fro}}^2\,, \tag{94}$$

$$\|\mathrm{D}\,\mathrm{inv}_{\mathbf{M}_1}(\mathbf{M}) - \mathrm{D}\,\mathrm{inv}_{\mathbf{M}_2}(\mathbf{M})\|_{\mathrm{op}} \leq C_{\mathrm{Fro}}\left(\|\mathbf{M}_1^{-1}\|_{2\to 2} + \|\mathbf{M}_2^{-1}\|_{2\to 2}\right)\|\mathbf{M}_1^{-1}\|_{2\to 2}\|\mathbf{M}_2^{-1}\|_{2\to 2}\|\mathbf{M}_1 - \mathbf{M}_2\|_{\mathrm{Fro}}\,. \tag{95}$$

Idea of the proof. It follows the ideas of the proof of Lemma 1. □

1) Control of the short chords : We begin with the following preliminary result that ensures that τ can not be too close to 0.

Lemma 15. *Assume that $(\tau, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \in (0,1] \times \mathfrak{S}_0^{-1} \times \mathfrak{S}_0^{-1}$ verifies $0 < \|\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1}\|_{\Lambda} \leq \eta$ and $\eta/leqc_{\mathrm{Fro}}/2$. Then τ is bounded away from 0 i.e.*

$$\tau \geq 1 - \frac{\eta}{c_{\mathrm{Fro}}} \geq \frac{1}{2}\,.$$

Proof. From the inverse-lipschitz property of the inverse function in (93), we have :

$$\|\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2\|_{\mathrm{Fro}} \leq \frac{1}{c_{\mathrm{Fro}}}\|\boldsymbol{\Theta}_1\|_{2\to 2}\|\tau^{-1}\boldsymbol{\Theta}_2\|_{2\to 2}\|\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1}\|_{\Lambda} \leq \frac{\eta}{c_{\mathrm{Fro}}\tau}.$$

Moreover,

$$\|\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2\|_{\mathrm{Fro}} \geq \|\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2\|_{2\to 2} \geq \tau^{-1}\|\boldsymbol{\Theta}_2\|_{2\to 2} - \|\boldsymbol{\Theta}_1\|_{2\to 2} = \frac{1}{\tau} - 1\,.$$

Combining the two inequalities yields $\tau \geq 1 - \frac{\eta}{c_{\mathrm{Fro}}}$. □

Let us now prove Lemma 11.

Proof of Lemma 11. Take $\mathbf{U} = \frac{\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1}}{\|\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1}\|_{\Lambda}} \in \overline{S}_{\eta}^{-}[\mathfrak{S}_{k,\kappa_0}]$ so we have $0 < \|\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1}\|_{\Lambda} \leq \eta$. Note that by using (92) and (93), we have

$$0 \overset{(92)}{<} \|\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2\|_{\mathrm{Fro}} \overset{(93)}{\leq} \frac{\eta}{c_{\mathrm{Fro}}\tau}\|\boldsymbol{\Theta}_1\|_{2\to 2}\|\boldsymbol{\Theta}_2\|_{2\to 2} \leq \frac{\eta}{c_{\mathrm{Fro}}\tau}. \tag{96}$$

Now, from (94) with $\mathbf{M}_1 = \boldsymbol{\Theta}_1$ and $\mathbf{M}_2 = \tau^{-1}\boldsymbol{\Theta}_2$, we have

$$\begin{aligned}\|\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1} - \mathrm{D}\,\mathrm{inv}_{\tau^{-1}\boldsymbol{\Theta}_2}(\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2)\|_{\Lambda} &\overset{(94)}{\leq} C_{\mathrm{Fro}}\|\boldsymbol{\Theta}_1^{-1}\|_{\mathrm{Fro}}\|\tau\boldsymbol{\Theta}_2^{-1}\|_{\mathrm{Fro}}^2\|\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2\|_{\mathrm{Fro}}^2 \\ &\leq C_{\mathrm{Fro}}\kappa_0^3\tau^2\|\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2\|_{\mathrm{Fro}}^2\,.\end{aligned}$$

Dividing by $\|\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1}\|_{\Lambda} > 0$ in the above inequality yields:

$$\begin{aligned}\Big\|\frac{\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1}}{\|\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1}\|_{\Lambda}} - \mathrm{D}\,\mathrm{inv}_{\tau^{-1}\boldsymbol{\Theta}_2}\frac{\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2}{\|\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1}\|_{\Lambda}}\Big\|_{\Lambda} &\leq C_{\mathrm{Fro}}\kappa_0^3\tau^2\|\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2\|_{\mathrm{Fro}}\frac{\|\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2\|_{\mathrm{Fro}}}{\|\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1}\|_{\Lambda}} \\ \text{(applying (96) and (93))} \quad &\leq C_{\mathrm{Fro}}\kappa_0^3\tau^2\frac{\eta}{c_{\mathrm{Fro}}\tau}\frac{1}{c_{\mathrm{Fro}}}\|\boldsymbol{\Theta}_1\|_{2\to 2}\|\tau^{-1}\boldsymbol{\Theta}_2\|_{2\to 2} \\ &= \frac{C_{\mathrm{Fro}}\kappa_0^3}{c_{\mathrm{Fro}}^2}\eta = Z_0\eta\,.\end{aligned}$$

Moreover, as $\boldsymbol{\Theta}_2 \in \mathfrak{S}_0^{-1} \subset \mathfrak{S}_{k,\frac{1}{\kappa_0},1}^{-1}$, setting $\tilde{\boldsymbol{\Theta}} = \tau^{-1}\boldsymbol{\theta}_2$, we have $\tilde{\boldsymbol{\Theta}} \in \mathfrak{S}_{k,\frac{1}{\tau\kappa_0},\frac{1}{\tau}}^{-1} \subset \mathfrak{S}_{k,\frac{1}{\kappa_0},2}^{-1}$ (the inclusion comes from Lemma 15). Remark that the element in the differential can be expressed as

$$\frac{\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2}{\|\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1}\|_{\Lambda}} = \frac{\|\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2\|_{\mathrm{Fro}}}{\|\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1}\|_{\Lambda}}\frac{\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2}{\|\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2\|_{\mathrm{Fro}}}.$$

So, if we define $\lambda = \frac{\|\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2\|_{\mathrm{Fro}}}{\|\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1}\|_{\Lambda}}$, by (93) and Lemma 15 it satisfies $0 < \lambda \leq \frac{1}{c_{\mathrm{Fro}}\tau} \leq \frac{2}{c_{\mathrm{Fro}}} \triangleq \lambda_0$. Thus, there exists $\lambda \in]0, \lambda_0]$ and $\tilde{\boldsymbol{\Theta}} \in \mathfrak{S}_{k,\frac{1}{\kappa_0},2}^{-1}$ such that:

$$\Big\|\frac{\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1}}{\|\boldsymbol{\Theta}_1^{-1} - \tau\boldsymbol{\Theta}_2^{-1}\|_{\Lambda}} - \mathrm{D}\,\mathrm{inv}_{\tilde{\boldsymbol{\Theta}}}\,\lambda\frac{\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2}{\|\boldsymbol{\Theta}_1 - \tau^{-1}\boldsymbol{\Theta}_2\|_{\mathrm{Fro}}}\Big\|_{\Lambda} \leq Z_0\eta\,.$$

Now we set $\mathbf{V} = \frac{\mathbf{\Theta}_1 - \tau^{-1}\mathbf{\Theta}_2}{\|\mathbf{\Theta}_1 - \tau^{-1}\mathbf{\Theta}_2\|_{\mathrm{Fro}}}$ and we have $\mathbf{V} \in S[\mathfrak{S}^{-1}_{k,\kappa_0}]$, which finishes the proof. □

Proof of Lemma 12. First observe that for all $\mathbf{C} \in C$, we have $\|\mathbf{C}\|_{\mathrm{Fro}} \leq \lambda_0$ since $\forall \mathbf{V} \in S[\mathfrak{S}^{-1}_{k,\kappa_0}], \|\mathbf{V}\|_{\mathrm{Fro}} = 1$. Then, take $\overline{\mathfrak{S}}^{-1}_{k,\frac{1}{\kappa_0},2}$ an ε-net of $\mathfrak{S}^{-1}_{k,\frac{1}{\kappa_0},2}$ and $\overline{C}$ an ε-net of C. Take $\mathbf{T} = \mathrm{D}\,\mathrm{inv}_{\tilde{\mathbf{\Theta}}}(\mathbf{C}) \in T_C$ and consider $(\overline{\mathbf{\Theta}}, \overline{\mathbf{C}}) \in \overline{\mathfrak{S}}^{-1}_{k,\frac{1}{\kappa_0},2} \times \overline{C}$ such that $\|\overline{\mathbf{C}} - \mathbf{C}\|_{\mathrm{Fro}} \leq \varepsilon$ and $\|\overline{\mathbf{\Theta}} - \tilde{\mathbf{\Theta}}\|_{\mathrm{Fro}} \leq \varepsilon$. Then with $\overline{\mathbf{T}} = \mathrm{D}\,\mathrm{inv}_{\overline{\mathbf{\Theta}}}(\overline{\mathbf{C}}) \in T_C$,

$$\begin{aligned}
\|\mathbf{T} - \overline{\mathbf{T}}\|_\Lambda &= \|\mathrm{D}\,\mathrm{inv}_{\tilde{\mathbf{\Theta}}}(\mathbf{C}) - \mathrm{D}\,\mathrm{inv}_{\overline{\mathbf{\Theta}}}(\overline{\mathbf{C}})\|_\Lambda \\
&\leq \|\mathrm{D}\,\mathrm{inv}_{\tilde{\mathbf{\Theta}}}(\mathbf{C}) - \mathrm{D}\,\mathrm{inv}_{\overline{\mathbf{\Theta}}}(\mathbf{C})\|_\Lambda + \|\mathrm{D}\,\mathrm{inv}_{\overline{\mathbf{\Theta}}}(\mathbf{C}) - \mathrm{D}\,\mathrm{inv}_{\overline{\mathbf{\Theta}}}(\overline{\mathbf{C}})\|_\Lambda \\
&\leq \|\mathrm{D}\,\mathrm{inv}_{\tilde{\mathbf{\Theta}}} - \mathrm{D}\,\mathrm{inv}_{\overline{\mathbf{\Theta}}}\|_{\mathrm{op}} \|\mathbf{C}\|_{\mathrm{Fro}} + \|\mathrm{D}\,\mathrm{inv}_{\overline{\mathbf{\Theta}}}\|_{\mathrm{op}} \|\mathbf{C} - \overline{\mathbf{C}}\|_{\mathrm{Fro}}\,.
\end{aligned}$$

According to (95),

$$\begin{aligned}
&\|\mathrm{D}\,\mathrm{inv}_{\tilde{\mathbf{\Theta}}} - \mathrm{D}\,\mathrm{inv}_{\overline{\mathbf{\Theta}}}\|_{\mathrm{op}} \\
\leq& C_{\mathrm{Fro}} \left(\|\tilde{\mathbf{\Theta}}^{-1}\|_{2\to 2} + \|\overline{\mathbf{\Theta}}^{-1}\|_{2\to 2} \right) \|\tilde{\mathbf{\Theta}}^{-1}\|_{2\to 2} \|\overline{\mathbf{\Theta}}^{-1}\|_{2\to 2} \|\tilde{\mathbf{\Theta}} - \overline{\mathbf{\Theta}}\|_{\mathrm{Fro}} \\
\leq& 2C_{\mathrm{Fro}} \kappa_0^3 \|\tilde{\mathbf{\Theta}} - \overline{\mathbf{\Theta}}\|_{\mathrm{Fro}}.
\end{aligned}$$

We can also show that $\|\mathrm{D}\,\mathrm{inv}_{\overline{\mathbf{\Theta}}}\|_{\mathrm{op}} \leq C_{\mathrm{Fro}} \kappa_0^2$. Therefore,

$$\begin{aligned}
\|\mathbf{T} - \overline{\mathbf{T}}\|_\Lambda &\leq 2C_{\mathrm{Fro}} \kappa_0^3 \|\tilde{\mathbf{\Theta}} - \overline{\mathbf{\Theta}}\|_{\mathrm{Fro}} \delta + \|\mathrm{D}\,\mathrm{inv}_{\overline{\mathbf{\Theta}}}\|_{\mathrm{op}} \varepsilon \\
&\leq 2C_{\mathrm{Fro}} \kappa_0^3 \varepsilon \delta + C_{\mathrm{Fro}} \kappa_0^2 \varepsilon \\
&= C_{\mathrm{Fro}} (2\kappa_0^3 \lambda_0 + \kappa_0^2) \varepsilon\,.
\end{aligned}$$

This gives $\mathcal{N}(T_C, \|\cdot\|_\Lambda, C_{\mathrm{Fro}}(2\kappa_0^3\lambda_\eta + \kappa_0^2)\varepsilon) \leq \mathcal{N}(\mathfrak{S}^{-1}_{k,\frac{1}{\kappa_0},2}, \|\cdot\|_{\mathrm{Fro}}, \varepsilon) \times \mathcal{N}(C, \|\cdot\|_E, \varepsilon)$. Therefor, by setting $C_0 = C_{\mathrm{Fro}}(2\kappa_0^3\lambda_0 + \kappa_0^2)$, we have

$$\mathcal{N}(T_C, \|\cdot\|_\Lambda, \varepsilon) \leq \mathcal{N}(\mathfrak{S}^{-1}_{k,\frac{1}{\kappa_0},2}, \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon}{C_0}) \times \mathcal{N}(C, \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon}{C_0})\,.$$

□

Proof of Proposition 8. Recall that from Lemma 12 we have

$$\mathcal{N}(T_C, \|\cdot\|_\Lambda, \varepsilon) \leq \mathcal{N}(\mathfrak{S}^{-1}_{k,\frac{1}{\kappa_0},2}, \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon}{C_0}) \times \mathcal{N}(C, \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon}{C_0})\,,$$

with $C_0 = C_{\mathrm{Fro}}(2\kappa_0^3\lambda_0 + \kappa_0^2)$. Using Lemma 11, we also have the approximation of $\overline{S^-_\eta}[\mathfrak{S}_{k,\kappa_0}]$ by the set T_C:

$$\forall \mathbf{U} \in \overline{S^-_\eta}[\mathfrak{S}_{k,\kappa_0}], \exists \mathbf{T} \in T_C,\ \|\mathbf{U} - \mathbf{T}\|_\Lambda \leq Z_0 \eta\,.$$

Therefore, we can apply Lemma 3 (with $\delta = Z_0\eta$) to prove that for any $\varepsilon > 0$:

$$\mathcal{N}(\overline{S^-_\eta}[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, 2(\varepsilon + Z_0\eta)) \leq \mathcal{N}(T_C, \|\cdot\|_\Lambda, \varepsilon)\,. \tag{97}$$

Thus, combining (91) and (97) yields

$$\mathcal{N}(\overline{S^-_\eta}[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, 2(\varepsilon + Z_0\eta)) \leq \mathcal{N}(\mathfrak{S}^{-1}_{k,\frac{1}{\kappa_0},2}, \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon}{C_0}) \times \mathcal{N}(C, \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon}{C_0})\,.$$

All we need now is to control $\mathcal{N}(C, \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon}{C_0})$. Take $\overline{S[\mathfrak{S}^{-1}_{k,\kappa_0}]}$ a ε-net of $S[\mathfrak{S}^{-1}_{k,\kappa_0}]$ and $\overline{(0,\lambda_0]}$ an $(\lambda_0\varepsilon)$-net of $(0,\lambda_0]$. Take $\mathbf{C} = \lambda\mathbf{V} \in C$ with $\lambda \in (0,\lambda_0]$ and $\mathbf{V} \in S[\mathfrak{S}^{-1}_{k,\kappa_0}]$. Then there exists $\overline{\mathbf{V}} \in \overline{S[\mathfrak{S}^{-1}_{k,\kappa_0}]}$ such that $\|\mathbf{V} - \overline{\mathbf{V}}\|_{\mathrm{Fro}} \leq \varepsilon$ and $\overline{\lambda} \in \overline{(0,\lambda_0]}$ such that $|\lambda - \overline{\lambda}| \leq \lambda_0\varepsilon$. Define $\overline{\mathbf{C}} = \overline{\lambda}\overline{\mathbf{V}}$. We clearly have that $\overline{\mathbf{C}} \in C$ and

$$\begin{aligned}
\|\overline{\mathbf{C}} - \mathbf{C}\|_{\mathrm{Fro}} &= \|\overline{\lambda}\overline{\mathbf{V}} - \lambda\mathbf{V}\|_{\mathrm{Fro}} \\
&\leq \|\overline{\lambda}\overline{\mathbf{V}} - \lambda\overline{\mathbf{V}}\|_{\mathrm{Fro}} + \|\lambda\overline{\mathbf{V}} - \lambda\mathbf{V}\|_{\mathrm{Fro}} \\
&\leq \|\mathbf{V}\|_{\mathrm{Fro}} |\lambda - \overline{\lambda}| + \lambda \|\overline{\mathbf{V}} - \mathbf{V}\|_{\mathrm{Fro}} \leq 2\lambda_0\varepsilon\,.
\end{aligned}$$

Thus, for any $\varepsilon > 0$ we have $\mathcal{N}(C, \|\cdot\|_{\mathrm{Fro}}, 2\lambda_0\varepsilon) \leq \mathcal{N}((0,\lambda_0], |\cdot|, \lambda_0\varepsilon)\mathcal{N}(S[\mathfrak{S}^{-1}_{k,\kappa_0}], \|\cdot\|_{\mathrm{Fro}}, \varepsilon) \leq \varepsilon^{-1}\mathcal{N}(S[\mathfrak{S}^{-1}_{k,\kappa_0}], \|\cdot\|_{\mathrm{Fro}}, \varepsilon)$ or equivalently for any $\varepsilon > 0$ we have $\mathcal{N}(C, \|\cdot\|_{\mathrm{Fro}}, \varepsilon) \leq 2\lambda_0\varepsilon^{-1}\mathcal{N}(S[\mathfrak{S}^{-1}_{k,\kappa_0}], \|\cdot\|_{\mathrm{Fro}}, \varepsilon/(2\lambda_0))$. In conclusion we have

$$\mathcal{N}(\overline{S^-_\eta}[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, 2(\varepsilon + Z_0\eta)) \leq 2\lambda_\eta C_0 \varepsilon^{-1} \mathcal{N}(\mathfrak{S}^{-1}_{k,\frac{1}{\kappa_0},2}, \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon}{C_0}) \times \mathcal{N}(S[\mathfrak{S}^{-1}_{k,\kappa_0}], \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon}{2\lambda_0 C_0})\,,$$

which concludes the proof. □

2) Covering number for bounded sparse symmetric matrices and their secant: We now prove Lemma 13.

Proof of Lemma 13. Recall that Lemma 7 provides the following control of the covering number of the set of symmetric sparse matrices with bounded Frobenius norm $\mathfrak{W}_k = \{\mathbf{\Theta} \in S_d(\mathbb{R}) \, ; \|\mathbf{\Theta}\|_0 \leq d + 2k, \|\mathbf{\Theta}\|_{\mathrm{Fro}} \leq 1\}$:

$$\mathcal{N}(\mathfrak{W}_k, \|\cdot\|_{\mathrm{Fro}}, \varepsilon) \leq (\frac{ed^2}{2k})^k (\frac{9}{\varepsilon})^{d+k} \, ,$$

Noticing that $\mathfrak{S}_0^{-1} \subset \mathfrak{W}_k$, $\mathfrak{S}_{k,\frac{1}{\kappa_0},2}^{-1} \subset 2\,\mathfrak{W}_k$ and that $S[\mathfrak{S}_{k,\kappa_0}^{-1}] \subset \mathfrak{W}_{2k}$ yields the result. □

3) Proof of Theorem 5 :

Proof. First recall that for any $\varepsilon' > 0$ and $\eta > 0$, by Proposition 8 we have

$$\mathcal{N}(\overline{S}_\eta^-[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, 2(\varepsilon' + Z_0\eta)) \leq 2\lambda_\eta C_{0,\eta} \varepsilon'^{-1} \mathcal{N}(\mathfrak{S}_{k,\frac{1}{\kappa_0},\frac{1}{\tau_\eta}}^{-1}, \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon'}{C_{0,\eta}}) \times \mathcal{N}(S[\mathfrak{S}_{k,\kappa_0}^{-1}], \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon'}{2\lambda_\eta C_{0,\eta}}) \, ,$$

Let us fix ε such that $0 < \varepsilon \leq \frac{2\kappa_0^3}{c_{\mathrm{Fro}}}\sqrt{d} = c_{\mathrm{Fro}} Z_0$, as in the hypothesis of Theorem 5. We now set $\varepsilon' = \varepsilon/4$ and $\eta = \varepsilon/(4Z_0)$ such that $2(\varepsilon' + Z_0\eta) = \varepsilon$. Remark that these choices satisfy $\eta \leq c_{\mathrm{Fro}}/2$ as desired in the previous section. Hence,

$$\begin{aligned}
&\mathcal{N}(\overline{S}_\eta[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, \varepsilon) \\
\leq &\mathcal{N}(\overline{S}_\eta^+[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, \varepsilon) + \mathcal{N}(\overline{S}_\eta^-[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, \varepsilon) \\
\leq &\mathcal{N}((0,1], |\cdot|, \frac{\eta\varepsilon}{6L_0}) \times \mathcal{N}(\mathfrak{S}_0^{-1}, \|\cdot\|_{\mathrm{Fro}}, \frac{\eta\varepsilon}{6L_0})^2 \\
&+ 2\lambda_0 C_0 \varepsilon'^{-1} \mathcal{N}(\mathfrak{S}_{k,\frac{1}{\kappa_0},2}^{-1}, \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon'}{C_0}) \times \mathcal{N}(S[\mathfrak{S}_{k,\kappa_0}^{-1}], \|\cdot\|_{\mathrm{Fro}}, \frac{\varepsilon'}{2\lambda_0 C_0}) \, .
\end{aligned}$$

Straightforward computations show that $\frac{\eta\varepsilon}{6L_0} = c_0\varepsilon^2$ with $c_0 = \frac{1}{24Z_0L_0}$, $\frac{\varepsilon'}{C_{0,\eta}} = c_1\varepsilon$ where $c_1 = \frac{c_{\mathrm{Fro}}}{4C_{\mathrm{Fro}}(4\kappa_0^3 + c_{\mathrm{Fro}}\kappa_0^2)}$ and $\frac{\varepsilon'}{2\lambda_0 C_0} = c_2\varepsilon$ with $c_2 = \frac{c_{\mathrm{Fro}}^2}{16C_{\mathrm{Fro}}(4\kappa_0^3 + c_{\mathrm{Fro}}\kappa_0^2)}$ $(= c_{\mathrm{Fro}}c_1/4)$. This allows us to continue the computation as follows:

$$\begin{aligned}
&\mathcal{N}(\overline{S}_\eta[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, \varepsilon) \\
\leq &\mathcal{N}((0,1], |\cdot|, c_0\varepsilon^2) \times N(\mathfrak{S}_0^{-1}, \|\cdot\|_{\mathrm{Fro}}, c_0\varepsilon^2)^2 + \frac{1}{c_2}\varepsilon^{-1}\mathcal{N}(\mathfrak{S}_{k,\frac{1}{\kappa_0},2}^{-1}, \|\cdot\|_{\mathrm{Fro}}, c_1\varepsilon) \times \mathcal{N}(S[\mathfrak{S}_{k,\kappa_0}^{-1}], \|\cdot\|_{\mathrm{Fro}}, c_2\varepsilon). \\
\leq &\frac{1}{c_0}\varepsilon^{-2} \times \left(\frac{ed^2}{2k}\right)^{2k} \left(\frac{18\sqrt{d}}{c_0\varepsilon^2}\right)^{2(d+k)} + \frac{1}{c_2}\varepsilon^{-1} \left(\frac{ed^2}{2k}\right)^k \left(\frac{2\times 18\sqrt{d}}{c_1\varepsilon}\right)^{d+k} \times \left(\frac{ed(d-1)}{4k}\right)^{2k} \left(\frac{18\sqrt{d}}{c_2\varepsilon}\right)^{d+2k} \\
\leq &\frac{1}{c_0}\left(\frac{ed^2}{2k}\right)^{2k} \left(\frac{18\sqrt{d}}{c_0}\right)^{2(d+k)} \varepsilon^{-(4d+4k+2)} + \frac{1}{c_2}\left(\frac{ed^2}{2k}\right)^k \left(\frac{2\times 18\sqrt{d}}{c_1}\right)^{d+k} \times \left(\frac{ed^2}{4k}\right)^{2k} \left(\frac{18\sqrt{d}}{c_2}\right)^{d+2k} \varepsilon^{-(2d+3k+1)} \\
\leq &\left(\frac{ed^2}{2k}\right)^{3k} \left[\frac{(18\sqrt{d})^{2(d+k)}}{c_0^{2(d+k)+1}}\varepsilon^{-(4d+4k+2)} + \frac{2^{d+k}\,(18\sqrt{d})^{2d+3k}}{c_1^{d+k}\,c_2^{d+2k+1}}\varepsilon^{-(2d+3k+1)}\right].
\end{aligned}$$

Recall that c_0, c_1 and c_2 depends on d (directly or from $C_{\mathrm{Fro}} = \sqrt{d}$) and κ_0. Here are the explicit dependence (not considering the constant factors): $c_0 \propto \kappa_0^{-5} d^{-3/2}$, $c_1 \propto \kappa_0^{-3} d^{-1/2}$ $c_2 \propto \kappa_0^{-3} d^{-1/2}$. So we get, that there exists absolute constants $\tilde{c_1}$ and $\tilde{c_2}$ such that

$$\begin{aligned}
&\mathcal{N}(\overline{S}_\eta[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, \varepsilon) \\
\leq &\left(\frac{ed^2}{2k}\right)^{3k} \left[\tilde{c_1}^{d+k} \frac{d^{d+k}\,(\kappa_0^5 d^{3/2})^{2(d+k)+1}}{\varepsilon^{4d+4k+2}} + \tilde{c_2}^{d+k} \frac{\sqrt{d}^{2d+3k}(\kappa_0\sqrt{d})^{2d+3k+1}}{\varepsilon^{2d+3k+1}}\right] \\
\leq &\left(\frac{ed^2}{2k}\right)^{3k} \left[\left(\frac{\tilde{c_1}\kappa_0^5 d^2}{\varepsilon^2}\right)^{2(d+k)+1} + \left(\frac{\tilde{c_2}\kappa_0^3 d}{\varepsilon}\right)^{2d+3k+1}\right],
\end{aligned}$$

where we change the value of the absolute constant $\tilde{c_1}$ and $\tilde{c_2}$ at the last inequality. To finish the proof, recall that $\mathcal{N}(S_\eta[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, \varepsilon) \leq 2\mathcal{N}(\overline{S}_\eta[\mathfrak{S}_{k,\kappa_0}], \|\cdot\|_\Lambda, \varepsilon)$. □