Lifting high-dimensional non-linear models with Gaussian regressors

Christos Thrampoulidis
University of California, Santa Barbara

Ankit Singh Rawat*
Massachusetts Institute of Technology

Abstract

We study the problem of recovering a structured signal $\mathbf{x}_0$ from high-dimensional data $\mathbf{y}_i = f(\mathbf{a}_i^T\mathbf{x}_0)$ for some nonlinear (and potentially unknown) link function f, when the regressors $\mathbf{a}_i$ are iid Gaussian. Brillinger (1982) showed that ordinary least-squares estimates $\mathbf{x}_0$ up to a constant of proportionality μ_ℓ, which depends on f. Recently, Plan & Vershynin (2015) extended this result to the high-dimensional setting deriving sharp error bounds for the generalized Lasso. Unfortunately, both least-squares and the Lasso fail to recover $\mathbf{x}_0$ when $\mu_\ell = 0$. For example, this includes all even link functions. We resolve this issue by proposing and analyzing an alternative convex recovery method. In a nutshell, our method treats such link functions as if they were linear in a lifted space of higher-dimension. Interestingly, our error analysis captures the effect of both the nonlinearity and the problem's geometry in a few simple summary parameters.

1 Introduction

We consider the problem of estimating an unknown signal $\mathbf{x}_0 \in \mathbb{R}^n$ from a vector $\mathbf{y} = (y_1, y_2, \ldots, y_m)^T$ of m *generalized linear measurements* of the following form:

$$\mathbf{y}_i = f_i(\mathbf{a}_i^T\mathbf{x}_0), \quad i = 1, 2, \ldots, m. \tag{1}$$

Here, each $\mathbf{a}_i \in \mathbb{R}^m$ represents a (known) measurement vector and the f_i's are independent copies of a (possibly random) link function f. A few examples include: $f_i(x) = x + z_i$, with say z_i being random noise, for standard linear regression setup; $f_i(x) = |x|^2 + z_i$, for quadratic (noisy) measurements; $f_i(x) = Q_b(|x|)$, where $Q_b(\cdot)$ denotes a quantizer on b number of bits. In the statistics and econometrics literature, (1) is popular under the name *single-index model* (also regarded as a special case of the *sufficient dimension reduction* problem). As a slight generalization of (1), we allow measurements that are drawn independently according to a conditional distribution of some probability density function as follows:

$$\mathbf{y}_i \sim p(y \mid \mathbf{a}_i^T\mathbf{x}_0), \quad i = 1, 2, \ldots, m. \tag{2}$$

This expands the model by including instances such as logistic-regression and $y_i \sim \text{Poisson}(|\mathbf{a}_i^T\mathbf{x}_0|)$, $i \in [m]$.

We seek recovery methods that ensure the following favorable features: (i) computational efficiency; (ii) provable performance guarantees; (iii) flexibility to exploit various forms of possible prior structural knowledge about $\mathbf{x}_0$ (e.g., sparsity); and (iv) flexibility to the underlying link-function. Here, "flexible" refers to a method that can easily adapt to different structural information on $\mathbf{x}_0$ and different link-functions with minimal changes, such as appropriate tuning of involved regularization parameters. Lastly, we remark that information about the magnitude of $\mathbf{x}_0$ might in general be lost in the nonlinearity. Thus, for partial identifiability, we assume throughout that $\|\mathbf{x}_0\|_2 = 1$ and aim to recover an estimate that is highly correlated with the true signal (often referred to as *weak recovery*, e.g., [17]).

1.1 Prior art

In the simplest case with *linear* link function, i.e., $f_i(x) = x + z_i$, perhaps the most popular approach of estimating $\mathbf{x}_0$ is via solving the generalized Lasso:

$$\hat{\mathbf{x}} := \arg\min_{\mathbf{x}} \sum_{i=1}^{m} (\mathbf{y}_i - \mathbf{a}_i^T\mathbf{x}_0)^2 \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{K}_{\mathcal{R}}. \tag{3}$$

Here, for a properly chosen *regularizer* function $\mathcal{R} : \mathbb{R}^n \to \mathbb{R}$, $\mathcal{K}_{\mathcal{R}} \subset \mathbb{R}^n$ is a set that encodes the available information about $\mathbf{x}_0$. For instance, $\mathcal{K}_{\mathcal{R}} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathcal{R}(\mathbf{x}) \leq K\}$, where $K > 0$ is a (tuning) parameter. The generalized Lasso comes with provable performance guarantees under general assumptions on the choice of $\mathcal{R}$ and on the measurement vectors. The Lasso objective is by nature tailored to linear measurement models, but one can always employ it as a candidate recovery algorithm for the more general model (1). This immediately raises the following question: *For such non-linear measurements, when (if ever) is the solution $\hat{\mathbf{x}}$ of the Lasso still a good estimate of $\mathbf{x}_0$?*

*Ankit Singh Rawat is now with Google Research, NY.

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

This question has been recently addressed in a quantitative way in [22], under the assumption that the measurement vectors are independent Gaussians. Naturally, the answer depends on both: (i) the specific non-linearity f in (1); and (ii) the structure of $\mathbf{x}_0$ and the associated choice of the regularizer. The dependence on these factors can be conveniently summarized in terms of a few easily computable model parameters. In particular, the effect of the non-linearity is entirely captured by the following:

$$\mu_\ell := \mathbb{E}[\gamma f(\gamma)] \quad \text{and} \quad \tau_\ell^2 := \mathbb{E}[(f(\gamma)-\mu\gamma)^2], \tag{4}$$

where, the expectations are over $\gamma \sim \mathcal{N}(0,1)$ and the (possibly) random link function f. For simplicity, we focus on the case of sparse recovery, but the results of [22] are more generally applicable. When, $\mathbf{x}_0$ is k-sparse and $\hat{\mathbf{x}}$ is the solution to (3) with ℓ_1-regularization, then [22, Thm. 1.4] shows the following. Provided that $m \gtrsim k\log(n/k)$ and n is large enough, it holds with high probability that[1]:

$$\|\hat{\mathbf{x}} - \mu_\ell \cdot \mathbf{x}_0\|_2 \lesssim \tau_\ell \cdot \sqrt{k\log(n/k)}/\sqrt{m}. \tag{5}$$

What if the Lasso fails? It becomes clear from this discussion that (at least for Gaussian regressors) the generalized Lasso satisfies all the four aforementioned favorable properties that we seek in our recovery methods. However, a closer inspection of (5) reveals that the Lasso fails to produce a good estimate for all functions for which $\mu_\ell = 0$; e.g., this includes all *even* link functions!

1.2 Our contribution

In this paper, we provide an affirmative answer to the following natural questions that arise from the inadequacy of the Lasso for a large class of link functions: *Is there a generic convex program that can recover structured signals from even non-linear measurements? And if so, can we quantify its error performance?*

As we discuss next, our algorithm is motivated by a simple observation. It is well known that one can expand the link function f as a series in terms of the Hermite polynomials, i.e., $f(x) = \sum_{i=0}^{+\infty} \mu_i H_i(x)$, where $H_i(x)$ is the i^{th}-order Hermite polynomial with leading coefficient 1, and, $\mu_i = \frac{1}{n!}\mathbb{E}_\gamma[f(\gamma)H_i(\gamma)]$ for $\gamma \sim \mathcal{N}(0,1)$. In particular, we may expand $y_i = f(\mathbf{a}_i^T\mathbf{x}_0), i \in [m]$ as follows:

$$y_i = \mu_0 + \mu_1(\mathbf{a}_i^T\mathbf{x}_0) + \mu_2((\mathbf{a}_i^T\mathbf{x}_0)^2 - 1) + \dots \tag{6}$$

with $\mu_0 = \mathbb{E}[f(\gamma)], \mu_1 = \mathbb{E}[\gamma f(\gamma)]$, and $\mu_2 = \frac{1}{2}\mathbb{E}[(\gamma^2 - 1)f(\gamma)]$. First, observe that $\mu_1 = \mu_\ell$ in (4); thus, we may interpret the objective function of the Lasso in (3) as one that attempts to approximate (in ℓ_2-sense) each y_i in (6) by only keeping the first-order (linear) term. Clearly, this approximation fails if $\mu_1 = 0$, and so does the Lasso. For such cases, we naturally propose keeping only the second-order (quadratic) term of the expansion in (6). Moreover, in order to obtain a favorable convex program, we apply the *lifting* technique. In particular, approximating (6), we write

$$\begin{aligned} y_i \approx \mu_2((\mathbf{a}_i^T\mathbf{x}_0)^2 - 1) &= \mathrm{tr}((\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}}\text{-}\mathbf{I}) \cdot \mu_2\mathbf{x}_0\mathbf{x}_0^{\mathrm{T}}) \\ &= \mathrm{tr}((\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}}\text{-}\mathbf{I}) \cdot \widetilde{\mathrm{X}}_0), \end{aligned} \tag{7}$$

where $\widetilde{X}_0$ denotes the rank-1 matrix $\mu_2 X_0 = \mu_2\mathbf{x}_0\mathbf{x}_0^T$, and the first equality follows as $\|\mathbf{x}_0\|_2 = 1$. Note that (after lifting) the quantity on the right-hand side (RHS) in (7) is now a linear function of the unknown rank-1 matrix $\widetilde{X}_0$.

We are now ready to describe our algorithm. From (7), one can attempt to reconstruct $\widetilde{X}_0$ (i.e., a scaled version of $X_0 = \mathbf{x}_0\mathbf{x}_0^T$) by searching for a positive-semidefinite and low-rank matrix X that minimizes the residual between y_i and $\mathrm{tr}((\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}}\text{-}\mathbf{I}) \cdot \mathrm{X})$. Further relaxing the rank constraint leads to the following convex method:

$$\begin{aligned} \widehat{X} = \arg\min_{X \succeq 0} \; & \sum_{i=1}^{m} \big(\mathbf{y}_i - \mathrm{tr}((\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}}\text{-}\mathbf{I}) \cdot \mathrm{X})\big)^2 + \lambda \cdot \mathrm{tr}(\mathrm{X}), \\ & \text{subject to} \quad X \in \mathcal{K}_{\mathcal{R}}, \end{aligned} \tag{8}$$

where $\lambda > 0$ is a regularization parameter. We have further added a constraint $X \in \mathcal{K}_{\mathcal{R}}$ to promote the structure of the signal-defined matrix X_0 (inherited by the structure of $\mathbf{x}_0$). For instance, if $\mathbf{x}_0$ is k-sparse, then X_0 is k^2-sparse and a natural choice becomes $X \in \mathcal{K}_{\ell_1} := \{W \mid \|W\|_1 \le K\}$, for $\|W\|_1 = \sum_{i,j=1}^n |W_{ij}|$ and regularizer $K > 0$. As a last step, our algorithm obtains a final estimate $\hat{\mathbf{x}}$ of (a scaled) $\mathbf{x}_0$ by the leading eigenvector of $\widehat{X}$.

Error guarantees. We characterize the estimation performance of the following constrained version of (8):

$$\begin{aligned} \widehat{X} = \arg\min_{X \succeq 0} \; & \sum_{i=1}^{m} \big(\mathbf{y}_i - \mathrm{tr}((\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}}\text{-}\mathbf{I}) \cdot \mathrm{X})\big)^2 \\ & \text{subject to} \quad \mathrm{tr}(\mathrm{X}) \le \widetilde{\mu} \quad \text{and} \quad X \in \mathcal{K}_{\mathcal{R}}, \end{aligned} \tag{9}$$

where $\widetilde{\mu} > 0$ is a tuning parameter. Here, we state a version of our result in the ideal case where $\widetilde{\mu} = \mu_2$ (cf. (6) and recall that the goal of (9) is to recover $\widetilde{X} = \mu_2 X_0$ for which $\mathrm{tr}(\widetilde{\mathrm{X}}) = \mu_2$). In Sec. 3.4), we also analyze the performance of (9) when $\widetilde{\mu}$ is not ideally tuned. Our result captures the effect of f via two simple parameters. For $\gamma \sim \mathcal{N}(0,1)$:

$$\begin{aligned} \mu_q := \mu_2 = \frac{1}{2}\mathbb{E}[(\gamma^2-1)f(\gamma)] \quad & \text{and} \\ \tau_q^2 := \mathbb{E}\big[\big(f(\gamma) - \mu\cdot(\gamma^2-1)\big)^2\big]. & \end{aligned} \tag{10}$$

For example, for recovery of a k-sparse signal we show the following about (9) with ℓ_1-regularization (Thm. 2.1 for

[1] Here and in the rest of the paper, a statement is said to hold with high probability if it holds with probability at least 0.99 (say). Also, the symbol "$\lesssim$" is used to hide universal constants (in particular ones that do not depend on f).

details): For $m \gtrsim k^2 \log(n/k)$ and n large enough, it holds with high probability that

$$\|\hat{X} - \mu_q X_0\|_F \lesssim \tau_q \cdot k \log(n/k)/\sqrt{m}. \qquad (11)$$

From eigenvalue perturbation theory, this further bounds the deviation of the leading eigenvector $\hat{\mathbf{x}}$ of $\widehat{X}$ from $\mathbf{x}_0$ (cf. Sec. 2.1). Observe the resemblance between (5) and our result in (11). The Lasso-related parameters μ_ℓ and τ_ℓ are replaced by μ_q and τ_q, respectively. It is evident from (11) that (9) is efficient for those link functions f for which μ_q is nonzero. In contrast with (5), in terms of sample complexity, (11) pays a penalty of $\mathcal{O}\left(k^2 \log(n/k)\right)$ rather than $\mathcal{O}\left(k \log(n/k)\right)$. This is not surprising since the same gap holds for the known algorithms for quadratic measurements (e.g., [20]), a special case of even link function.

We note that (11) is only an instantiation to sparse signals of our general result that characterizes the recovery performance of (9) for general convex regularizers $\mathcal{R}$ (cf. Sec. 2.2). Also, inherent in the formulation of (9) is the condition $\mu_q > 0$ (since, $X \succeq 0$ implies that $\mathrm{tr(X)} \geq 0$). Note that if $\mu_q < 0$, then the same method works only by replacing the constraint $X \succeq 0$ with $X \preceq 0$. Overall, the proposed algorithm requires that $\mu_q \neq 0$.

Remark 1 (Extensions)**.** Our algorithm is motivated by the expansion of the link function in (6), which we attempt to approximate in (8) by solely keeping the second-order term. Naturally, one can imagine the possibility of developing extensions to (8) that achieve better performance by more accurate approximations in (6). For example, one may keep both the zeroth- and the second-order terms:

$$\widehat{X} = \arg\min_{X \succeq 0} \sum_{i=1}^{m} (\mathbf{y}_i - \hat{\mu}_0 - \mathrm{tr}((\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}}\text{-}\mathbf{I}) \cdot \mathrm{X}))^2$$
$$\text{subject to} \quad \mathrm{tr(X)} \leq \widetilde{\mu} \quad \text{and} \quad X \in \mathcal{K}_{\mathcal{R}}, \qquad (12)$$

where, $\hat{\mu}_0 = \frac{1}{m}\sum_{i=1}^m y_i$. Note that for large enough number of measurements m, it holds $\hat{\mu}_0 \approx \mathbb{E}[f(\gamma)] = \mu_0$. We remark that our analysis of (9) directly translates to guarantees about (12) when $\hat{\mu}_0$ is estimated by a fresh batch of measurements (for example, this is achieved in practice by sample splitting). Other interesting extensions of this flavor are certainly possible, albeit may require additional effort.

Remark 2 (Relation to PhaseLift)**.** The lifting technique is by now well established in the literature, the most prominent example being its use in phase retrieval (i.e., quadratic measurements). In fact, the popular PhaseLift method [5] is very similar to (8), the only difference being in the loss function: the PhaseLift penalizes $\sum_{i=1}^m (\mathbf{y}_i - \mathrm{tr}(\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}} \cdot \mathrm{X}))^2$, instead. In reference to (6), note that in the special case of quadratic measurements, it holds $\mu_0 = \mu_2 = 1$, which explains the choice of that particular loss function in the PhaseLift. However, our arguments show the need for modifications to be able to treat a wide class of functions beyond quadratics. This leads to (12), which can be viewed as a canonical extension or robust version of the PhaseLift to more general link-functions.

1.3 Relevant literature and outlook

The vast majority of the literature on structured signal recovery in the high-dimensional regime assumes a linear measurement model. In this vanilla setting, convex methods offer many desired features: computational efficiency, tractable analysis, flexibility to adjust to different problem instances (such as, sparse recovery with ℓ_1-regularization and low-rank matrix recovery with nuclear-norm minimization). By now, performance guarantees for these methods are well-established ([9] and references therein).

Extensions to nonlinear link functions, have been only more recently considered in high dimensions. While, the key observation that for Gaussian measurement vectors it holds $\mu_\ell \mathbf{x}_0 = \arg\min_{\mathbf{x}} \mathbb{E}(f(\langle \mathbf{a}, \mathbf{x}_0\rangle) - \langle \mathbf{a}, \mathbf{x}\rangle)^2$ goes back to [4] (also, [14]), Plan and Vershynin [22, 23] were the first to show how that idea extends to the high-dimensional setting by applying it to the generalized Lasso. Subsequently [29] extended [22] to the regularized Lasso, obtained the exact constants in the analysis, and used the results to design optimal quantization schemes. More recent works involve extensions to other loss-functions beyond least squares [10], to elliptically symmetric distributions [11], to projected gradient-descent [21], to signal-demixing applications [24], and to demonstrating computational speedups compared to maximum-likelihood estimation [8]. Finally, Yang et al. [32] have appropriately modified Brillinger's original observation to sub-Gaussian vectors, based on which they propose and analyze generic convex solvers that work in this more general setting.

Unfortunately, all these works, starting with the original result by Brillinger, assume that the link function satisfies $\mu_\ell \neq 0$. Instead, our method and analysis works for these; and more generally for link functions satisfying $\mu_q \neq 0$. In that sense, our paper is a direct counterpart of [22] for "even-like" nonlinearities. Given the aforementioned extensions that followed [22], it is natural to expect analogous extensions of our results as part of future research. Also, perhaps an interesting research question that is raised is related to unifying the efficacy of the Lasso and of (8) in a single generic algorithm that would combine the best of the two worlds and would apply to nonlinearities satisfying either $\mu_\ell \neq 0$ or $\mu_q \neq 0$ (cf. Remark 1). On the one hand, since $\mu_\ell \neq 0$, one can obtain an estimate of $\mathbf{x}_0$ by the Lasso; however, this entirely ignores the hidden "quadratic part" in the measurements . On the other hand, seemingly natural lifting procedure leads to a semidefinite-optimization estimator that successfully accounts for both "linear and quadratic parts", but suffers from worse sample complexity. We point out the interesting and relevant work by Yi et al. [35] in this direction. Their algorithm is appli-

cable to such a wide class of link-functions, but, it imposes other restrictions compared to our work, such us requiring binary measurements and a sparse signal $\mathbf{x}_0$.

Out of all the nonlinear link functions, the quadratic case (cf. phase-retrieval) deserves special attention; there has been a surge in the provable recovery methods for this case over the past few years. Among the most well-established and the first chronologically to enjoy rigorous recovery guarantees are the methods based on semidefinite relaxation [3, 5]. Such methods operate by lifting the original n-dimensional natural parameter space to a higher dimensional matrix space. Here, we extend the lifting idea and the recovery guarantees to general link functions beyond quadratics. Naturally, the increase in dimensionality often introduces challenges in computational complexity and memory requirements. To overcome these issues, subsequent works on phase retrieval develop non-convex formulations that start with a careful spectral initialization (see [25, Sec. 6] for references), which is then iteratively refined by a gradient-descent-like scheme of low computational complexity. See also [2, 12] for an alternative *convex* formulation of the problem in the natural parameter space. Naturally, many of these solution methods can be combined with the ideas introduced in this paper to extend their reach to non-quadratic link functions. In fact, while preparing our paper, we became aware of the recent work by Yang et al. [34] exactly along these lines, which proposes an iterative non-convex method for sparse-signal recovery from link functions beyond quadratics. Interestingly, in order to arrive at their algorithm, Yang et. al. identify necessary modifications to the vanilla non-convex methods for quadratics, which are in exact agreement with our modification to the PhaseLift needed to arrive at (8). Thus, (8) and (12) can be interpreted as the convex counterparts to the algorithm of [34]. Despite the similarities, (a) the two papers arrive at the proposed algorithms through different ideas (cf. Remark 2); (b) our algorithm and analysis support structures beyond sparsity; (c) our performance bounds precisely capture the effect of the link-function.

Another closely related work [33] takes a somewhat different view than ours on the measurement model in (1) and arrives at a different semidefinite optimization algorithm closer to the algorithms for sparse PCA. Notably, [33] also addresses the design matrices with sub-gaussian entries. For Gaussian regressors, a parallel work to ours [28] extends the analysis of [33] beyond sparse recovery. Finally, [18, 15, 17] study the performance of spectral initialization for measurements as in (1). In contrast to our work, the results in [15, 17] are asymptotic and do not exploit structural information on $\mathbf{x}_0$. Also, compared to [18, 33, 28] our analysis is sharp with respect to the link function f. Such sharp results are directly useful in various applications where one has control over some parameters of f, and also, to the design of pre-processing functions h (see [29, 15, 17] for precursors of this idea). For example, by inspecting (11), the estimation error is minimized by choosing h so that μ_q and τ_q result in the effective noise parameter τ_q/μ_q being as small as possible (see Sec. 4). We postpone further investigations of such implications to future work.

Notation. We use $\mathbb{S}_n$ ($\mathbb{S}_n^+$) to denote the sets of real $n\times n$ symmetric (resp., positive semidefinite) matrices. For $U, V \in \mathbb{S}_n$, $\langle U, V\rangle = \mathrm{tr}(\mathrm{UV})$ denotes the standard inner product in $\mathbb{S}_n$, $\|V\|_2$ ($\|V\|_F$) denotes the spectral (resp., Frobenius) norm, and $\|V\|_1 = \sum_{j\in[m]}\sum_{i\in[n]}|V_{ji}|$. For $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|_p$ denotes its ℓ_p norm, $p \geq 1$. We denote by $\mathcal{S}_{F,n}$ the unit sphere in $\mathbb{S}_n$, i.e., $\mathcal{S}_{F,n} := \{V \in \mathbb{S}_n \mid \|V\|_F = 1\}$.

Organization. We formally state our results along with some technical background in Sec. 2. The case of sparse recovery is formally treated in Sec. 2.1, while error guarantees for the general setting are provided in Sec. 2.2. We highlight the key steps of the proofs in Sec. 3 (full details are relegated to the supplementary material). Also, in Sec. 3.4, we study the performance of (9) under imperfect tuning. Numerical simulations are presented in Sec. 4.

2 Error bounds for nonlinear measurements

2.1 An example: Sparse recovery

Assume that the true signal $\mathbf{x}_0$ is k-sparse. Then, $X_0 = \mathbf{x}_0\mathbf{x}_0^T$ is at most k^2-sparse. Thus, we solve (9) with an ℓ_1-norm constraint. Thm. 2.1 below characterizes the performance of the algorithm. Before that, recall the definitions of μ_q and τ_q in (10). Here onwards, for ease of exposition, we refer to μ_q and τ_q simply as μ and τ, respectively. If f is twice differentiable with f'' being its second derivative, then by integration by parts we have $\mu = \mathbb{E}[f''(\gamma)]/2$. Moreover, it is easily shown that $\tau^2 = \mathbb{E}[f^2(\gamma)] - 2\mu^2$.

For all the theorems that follow: *a statement is said to hold with high probability if it holds with probability at least 0.99 (say)*. Also, the appearing constants $c, C > 0$ may only depend on the probability of success.

Theorem 2.1 (Sparse recovery)**.** *Suppose that $\mathbf{x}_0$ is k-sparse, $\mathbf{a}_i \sim \mathcal{N}(0, \mathbf{I})$, and that $\mathbf{y}$ follows the generalized linear model of* (1)*. Assume that $\mu > 0$, and let $\widehat{X}$ be the solution to* (9) *with $\widetilde{\mu} = \mu$ and $\mathcal{K}_{\mathcal{R}} = \{X : \|X\|_1 \leq \mu\|X_0\|_1\}$. There exist universal constants $c, C > 0$ such that, if the number of observations obeys*

$$m \geq c \cdot k^2 \log(n/k) \tag{13}$$

for sufficiently large n, then, with high probability, we have

$$\|\widehat{X} - \mu X_0\|_F \leq C \cdot \tau \cdot k \log(n/k)/\sqrt{m}. \tag{14}$$

We defer the proof of Theorem 2.1 to Sec. A.1 in the supplementary. The theorem does not claim that $\hat{X}$ is rank one.

As usual, we obtain an estimate $\hat{\mathbf{x}}$ of $\mathbf{x}_0$ by extracting the rank-one component (e.g. [5]). In particular, if λ_1 and $\mathbf{v}_1$ denote the maximum eigenvalue and the principal eigenvector of $\hat{X}$, respectively, then we obtain $\hat{\mathbf{x}} = \sqrt{\lambda_1}\mathbf{v}_1$. It follows from (19) that $\|\hat{\mathbf{x}} - \sqrt{\mu}\mathbf{x}_0\|_2 \leq C \cdot \min\{\sqrt{\mu}, \frac{E}{\sqrt{\mu}}\}$, where E denotes the expression in the RHS of (14). The proof is based on the Davis-Kahan $\sin(\theta)$-theorem and is the same as in [5, Sec. 6]; thus it is omitted for brevity.

Thm. 2.1 implies the sample complexity of $\mathcal{O}\left(k^2\log(n/k)\right)$. Notably, for quadratic measurements, this is the same as the guarantees of [5]. In fact, the same k^2-barrier appears in most of the algorithms that have been proposed for sparse recovery from quadratic measurements (e.g., [25, Sec. 6] and references therein). However, given more information about the link-function (e.g., f having sub-exponential moments), the numerator of the RHS in (14) can be improved to $k^2\sqrt{\log(n/k)}$.

2.2 General result

Thm. 2.2 below characterizes the performance of (9) for general signal structure and $\mathcal{K}_{\mathcal{R}}$. The bounds are given in terms of specific summary parameters. We distinguish between: (i) *Geometric parameters* that capture the efficacy of the imposed geometric constraints in (9) in promoting solutions of desired structure (positive semidefinite, low-rank, sparse, etc.); (ii) *Model parameters* that capture f.

2.2.1 Geometric parameters

First, we introduce the notions of the tangent cone and the (local) Gaussian width.

Definition 1 (Tangent cone)**.** *The tangent cone of a subset $\mathcal{K} \subset \mathbb{S}_n$ at $X \in \mathbb{S}_n$ is defined as $\mathcal{D}(\mathcal{K}, X) := \{\tau V : \tau \geq 0, V \in \mathcal{K} - X\}$.*

Definition 2 ((Local) Gaussian width)**.** *The local Gaussian width of a set $\mathcal{C} \subset \mathbb{S}_n$ is a function of a scale parameter $t > 0$, which is defined as*

$$\omega_{g,t}(\mathcal{C}) := \mathbb{E}_G\big[\sup_{V \in \mathcal{C} \cap t\mathcal{S}_{F,n}} \langle G, V\rangle\big], \tag{15}$$

where G is a matrix from the Gaussian orthogonal ensemble (GOE), i.e. $G = G^T$, $G_{ii} \overset{iid}{\sim} \mathcal{N}(0,1)$ for $i \in [n]$, and, $G_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1/2)$ for $i > j \in [n]$.

For $t = 1$, we refer to $\omega_{g,1}(\mathcal{C})$ simply as the Gaussian width, and, we use the shorthand $\omega_g(\mathcal{C})$.

The results of this section only involve the Gaussian width $\omega_{g,1}(\mathcal{C}) = \omega_g(\mathcal{C})$ ($t = 1$, above). The general definition of local Gaussian width becomes useful when we study the performance of (9) under imperfect parameter tuning in Sec. 3.4. The Gaussian width plays a central role in asymptotic convex geometry. It also appears as a key quantity in the study of random linear inverse problems [7, 1]: for a cone $\mathcal{C} \subset \mathbb{S}_n$, $\omega_g(\mathcal{C})^2$ can be formally described as a measure of the effective dimension of the cone $\mathcal{C}$ [31, 1]. Importantly, while it is an abstract geometric quantity, it is possible in many instances to derive sharp numerical bounds that are explicit in terms of the parameters of interest (such as sparsity level, rank) [26, 7, 1]. We make use of these ideas in the proof of Thm. 2.2 (see Sec. 3.3.2).

Finally, we need two more geometric parameters: Talagrand's γ_1 and γ_2-functionals. To streamline the presentation, we defer the formal definitions of these parameters to Sec. C in the supplementary. For a set $\mathcal{K} \subset \mathbb{S}_n$ we write $\gamma_1(\mathcal{K}, \|\cdot\|_2)$ and $\gamma_2(\mathcal{K}, \|\cdot\|_F)$ for the γ_1 and γ_2-functionals with respect to the spectral and the Frobenius norms, respectively. The γ-functionals are fundamental in the study of suprema of random processes and specifically in the theory of generic chaining [27]. In general, explicit calculation of the γ-functionals can be challenging depending on the specific set $\mathcal{K}$; however, it is often possible to control them in a sufficient way. Specifically, for the term $\gamma_2(\mathcal{C}; \|\cdot\|_F)$, one can appeal to Talagrand's majorizing measure theorem that establishes a tight (up to constants) relations to the Gaussian width [27, Thm. 2.1.1], which can in turn be often well approximated: $\gamma_2(\mathcal{K}; \|\cdot\|_F) \leq C \cdot \omega_g(\mathcal{K})$. More generally, Dudley's integral produces bounds on γ_1, γ_2 in terms of the metric entropy of the set (see (66) in the supplementary) [2].

2.2.2 Model parameters

First, recall the definition of μ and τ in (10). Moreover, for $\gamma \sim \mathcal{N}(0,1)$ and all expectations take over γ and f:

$$\upsilon^2 := \mathbb{E}\big[\gamma^2 \cdot \big(f(\gamma) - \mu \cdot (\gamma^2 - 1)\big)^2\big] \quad \text{and}$$
$$\chi^2 := \mathbb{E}\big[(\gamma^2-1)^2 \cdot \big(f(\gamma) - \mu \cdot (\gamma^2 - 1)\big)^2\big]. \tag{16}$$

Remark 3. The results extend to the more general model of (2) with a natural modification in (10) and (16). In particular, $f(\gamma)$ is substituted by a (random variable) $y \sim p(y|\gamma)$, and the expectation is over the conditional distribution $p(y|\gamma)$ and γ, e.g., $\mu = \frac{1}{2}\int y\mathbb{E}_\gamma[(\gamma^2-1)p(y|\gamma)]dy$.

2.2.3 Main result

We are now ready to state the main result of this section.

Theorem 2.2 (General result)**.** *Suppose that $\mathbf{a}_i \sim \mathcal{N}(0, \mathbf{I})$, and that $\mathbf{y}$ follows the model in* (1)*. Recall the definitions of $\mu, \tau, \upsilon, \chi$ in* (10) *and* (16)*. Assume that $\mu > 0$ and that $\widetilde{\mu} = \mu$ and $\mu X_0 \in \mathcal{K}_{\mathcal{R}}$ in* (9)*, where $X_0 = \mathbf{x}_0\mathbf{x}_0^T$. Denote*

$$\Gamma := \min\big\{\sqrt{n}\,,\ \gamma_2\big(\mathcal{D}(\mathcal{K}_{\mathcal{R}}, \mu X_0) \cap \mathcal{S}_{F,n}, \|\cdot\|_F\big) + \gamma_1\big(\mathcal{D}(\mathcal{K}_{\mathcal{R}}, \mu X_0) \cap \mathcal{S}_{F,n}, \|\cdot\|_2\big)\big\}. \tag{17}$$

[2] See also [19] for recent progress on bounding the γ_1 functional of constraint sets with "good" covering number in terms of their Gaussian width. In particular, [19, Lemma D.19] can be combined with Thm. 2.2 to obtain specific results for other structures beyond sparsity.

There exist universal constants $c, C > 0$ such that, if

$$m \geq c \cdot \Gamma^2, \tag{18}$$

then, with high-probability, the solution $\widehat{X}$ of (9) *satisfies*

$$\|\widehat{X} - \mu X_0\|_F \leq \frac{C}{\sqrt{m}} \cdot \Big(\tau \cdot \Gamma + \chi + \upsilon \cdot \min\big\{\sqrt{n}, \omega_g\big(\mathcal{D}(\mathcal{K}_{\mathcal{R}}, \mu X_0)\big)\big\}\Big), \tag{19}$$

Note that Thm. 2.2 holds for all values of n. Typically, for large enough n, the first term in the right hand side (RHS) of (19), i.e., $\tau \cdot \Gamma/\sqrt{m}$, becomes the dominant term. For a simple illustration of the theorem, consider the case of a generic true signal $\mathbf{x}_0$ without any prior structural information. In this case, we solve (9) with no additional constraints other than $X \succeq 0$ and $\mathrm{tr(X)} \leq \mu$. Hence, $\Gamma \leq \sqrt{n}$ and from (18) the sample requirement is $m \geq c' \cdot n$.

3 Technical results and proofs

We now outline the proof of Thm. 2.2, which has two main steps. First, in Thm. 3.1, we upper bound the error $\|\widehat{X} - \mu X_0\|_F$ of (9) in terms of an appropriate geometric quantity, namely the *weighted empirical width*. We state the theorem and outline its proof in Sec. 3.1 and 3.2, respectively. Next, in Sec. 3.3, we show how to control the weighted empirical width to finally arrive at Thm. 2.2.

3.1 General upper bound

We begin the section by defining the local weighted empirical width. Similar to the definition of the local Gaussian width, this geometric is also a function of a scale parameter t. In this section, we only make use of the case $t = 1$, but the generality of the definition will prove useful in Sec. 3.4.

Definition 3 ((Local) empirical width)**.** *Let $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$ be independent copies of a standard normal vector $\mathcal{N}(0, \mathbf{I}_n)$ and $\varepsilon_1, \ldots, \varepsilon_m$ be independent Rademacher random variables. For a set $\mathcal{C} \subset \mathbb{S}_n$ and a vector $\mathbf{p} := (p_1, \ldots, p_m)$ the local weighted empirical width $\omega_{e,t}(\mathcal{C}; \mathbf{p})$ is a function of a scale parameter $t > 0$ defined as follows:*

$$\omega_{e,t}(\mathcal{C}; \mathbf{p}) := \mathbb{E}\big[\sup_{V \in \mathcal{C} \cap t\mathcal{S}_{F,n}} \langle V, H_{\mathbf{p}} \rangle\big], \tag{20}$$

where $H_{\mathbf{p}} := \frac{1}{\sqrt{m}} \sum_{i=1}^m p_i \cdot \varepsilon_i \cdot \mathbf{a}_i \mathbf{a}_i^T$, and, the expectation is over the randomness of $\{\mathbf{a}_i\}$ and of $\{\varepsilon_i\}$. In particular, when $\mathbf{p} = \mathbf{1}$ we write $\omega_{e,t}(\mathcal{C}) := \omega_{e,t}(\mathcal{C}; \mathbf{1})$, and call this the local empirical width.

If $t = 1$, we simply call $\omega_{e,1}(\mathcal{C}; \mathbf{p})$ the weighted empirical width, and, we use the shorthand $\omega_e(\mathcal{C}; \mathbf{p})$.

We study further the weighted empirical width in Sec. 3.3. Now, we are ready to state a general upper bound on the error of the estimate obtained by (9). For convenience, let:

$$\mathcal{K}_0 := \{X \ : \ X \succeq 0 \text{ and } \mathrm{tr(X)} \leq \widetilde{\mu}\}. \tag{21}$$

Theorem 3.1. *Let the same assumptions as in Thm. 2.2 hold (including $\tilde{\mu} = \mu$). Further let*

$$\mathcal{C}_0 := \mathcal{D}(\mathcal{K}_0, \mu X_0) \cap \mathcal{D}(\mathcal{K}_{\mathcal{R}}, \mu X_0). \tag{22}$$

Finally, define $\boldsymbol{\eta} := (\eta_1, \ldots, \eta_m)$ with $\eta_i := \big(f(\gamma_i) - \mu \cdot (\gamma_i^2 - 1)\big)$ and $\gamma_i \overset{iid}{\sim} \mathcal{N}(0,1)$, $i \in [m]$. There exist constants $c, C > 0$ such that, if the number of observations obeys

$$m \geq c \cdot \left(\omega_e(\mathcal{C}_0)\right)^2, \tag{23}$$

then, the solution $\widehat{X}$ of (9) *satisfies with high probability:*

$$\|\widehat{X} - \mu X_0\|_F \leq C \cdot \big(\mathbb{E}_{\boldsymbol{\eta}}\big[\omega_e(\mathcal{C}_0; \boldsymbol{\eta})\big] + \upsilon\sqrt{2} \cdot \omega_g(\mathcal{C}_0) + \chi\big)/\sqrt{m}. \tag{24}$$

3.2 Proof outline of Theorem 3.1

Let $\mathbf{y} = (y_1, \ldots, y_m)$ consist of m observations: $y_i = f_i(\mathbf{a}_i^T \mathbf{x}_0)$, $i \in [m]$. We use X_0 and $\widehat{X}$ to denote $\mathbf{x}_0 \mathbf{x}_0^T$ and the solution to (9), respectively. For convenience, we write the loss function in (9) as $\mathcal{L}(X) := \|\mathbf{y} - \mathcal{A}(X)\|_2^2$, where the operator $\mathcal{A} : \mathbb{R}^{n \times n} \to \mathbb{R}^m$ returns $\mathcal{A}(X) := \big(\mathbf{a}_1^T X \mathbf{a}_1 - \mathrm{tr(X)} \ , \ \ldots \ , \ \mathbf{a}_m^T X \mathbf{a}_m - \mathrm{tr(X)}\big)^T$. We define the *error matrix* $\widehat{V} = \widehat{X} - \mu X_0$. We need to upper bound $\|\widehat{V}\|_F$ to establish Thm. 2.2. Towards this direction, let us consider the excess loss function

$$\begin{aligned} 0 &\leq \mathcal{L}(\mu X_0) - \mathcal{L}(\widehat{X}) \\ &= \|\mathbf{y} - \mathcal{A}(\mu X_0)\|_2^2 - \|\mathbf{y} - \mathcal{A}(\mu X_0 + \widehat{V})\|_2^2. \end{aligned} \tag{25}$$

The nonnegativity follows by optimality of $\widehat{X}$ and feasibility of μX_0 (recall that $\mu > 0$ and $\widetilde{\mu} = \mu$). Therefore,

$$\|\mathcal{A}(\widehat{V})\|_2^2 \leq 2\langle \mathbf{y} - \mu \cdot \mathcal{A}(X_0), \mathcal{A}(\widehat{V}) \rangle. \tag{26}$$

On the one hand, recall from (9) that $\widehat{V}$ satisfies $\mathrm{tr}(\widehat{\mathrm{V}}) \leq 0$, $\mu X_0 + \widehat{V} \succeq 0$, and $\mu X_0 + \widehat{V} \in \mathcal{K}_{\mathcal{R}}$, i.e., $\widehat{V} \in \mathcal{C}_0$. Thus,

$$\widehat{V}/\|\widehat{V}\|_F \in \mathcal{E} := \mathcal{C}_0 \cap \mathcal{S}_{F,n}. \tag{27}$$

On the other hand, observe in (26) that the LHS (resp., RHS) is homogeneous of degree 2 (resp., 1). With these, it follows from (26) that

$$\|\widehat{V}\|_F \cdot \inf_{V \in \mathcal{E}} \|\mathcal{A}(V)\|_2^2 \leq 2 \sup_{V \in \mathcal{E}} \langle \mathcal{A}(V), \mathbf{y} - \mu \cdot \mathcal{A}(X_0) \rangle. \tag{28}$$

The strategy is now clear: we will obtain high-probability lower and upper bounds on the LHS and on the RHS, respectively. This will immediately upper bound $\|\widehat{V}\|_F$.

Lower bound: We lower bound the LHS of (28) by employing Mendelson's *Small Ball method* [16]. We defer details to Sec. B.1 of the supplementary material.

Lemma 3.1 (Lower bound)**.** *There exists positive absolute constant $c > 0$ such that for any $s > 0$ it holds with probability at least $1 - e^{-s^2/4}$:*

$$\inf_{V\in\mathcal{E}} \|\mathcal{A}(V)\|_2 \geq c\sqrt{m} - 4\,\omega_e(\mathcal{C}_0) - s. \tag{29}$$

Upper bound: The key observation for upper bounding the RHS of (28) is that we have set up the objective function of (9) and we have chosen the value of μ in such a way that the following holds. Let $z_i = f(\mathbf{a}_i^T\mathbf{x}_0) - \mu\cdot((\mathbf{a}_i^T\mathbf{x}_0)^2 - 1)$, $i \in [m]$. Then, setting $\gamma_i := \mathbf{a}_i^T\mathbf{x}_0$, $i \in [m]$, observing that $\gamma_i \sim \mathcal{N}(0,1)$, and recalling the definition of μ, we obtain

$$\mathbb{E}\big[z_i\cdot(\gamma_i^2 - 1)\big] = 0. \tag{30}$$

Therefore, z_i is uncorrelated with $((\mathbf{a}_i^T\mathbf{x}_0)^2 - 1)$, and (by definition) it only depends on the measurement vector $\mathbf{a}_i$ through $\mathbf{a}_i^T\mathbf{x}_0$. Using these, it can be shown that $\mathbb{E}\langle \mathbf{y} - \mu\,\mathcal{A}(X_0), \mathcal{A}(V)\rangle = 0$ for each fixed V (see (59) in Sec. B.3 of the supplementary material). On the other hand, in (28), we need to uniformly bound the deviation over *all* V. To do that while exploiting the key observation above, we decompose the measurement vectors as follows:

$$\mathbf{a}_i = \mathbf{x}_0\mathbf{x}_0^T\mathbf{a}_i + \big(\mathbf{I} - \mathbf{x}_0\mathbf{x}_0^T\big)\,\mathbf{a}_i =: P\mathbf{a}_i + P^{\perp}\mathbf{a}_i,$$

where $P, P^{\perp}$ denote the projection operators to the direction of $\mathbf{x}_0$ and to its orthogonal subspace, respectively. With this representation, it can be shown that

$$\langle \mathbf{y} - \mu\cdot\mathcal{A}(X_0), \mathcal{A}(V)\rangle = \text{Term I} + \text{Term II} + \text{Term III}$$

where,

$$\text{Term I} := \sum_{i\in[m]} \mathrm{tr}(\mathrm{V}\mathrm{X}_0)\,(\gamma_i^2 - 1)\,z_i, \tag{31a}$$

$$\text{Term II} := \sum_{i\in[m]} \big(\mathbf{a}_{i,\perp}^T V\mathbf{a}_{i,\perp} + \mathrm{tr}(\mathrm{V}(\mathrm{X}_0\text{-}\mathbf{I}))\big)\,z_i, \tag{31b}$$

$$\text{Term III} := \sum_{i\in[m]} \big(\mathbf{x}_0^T V\mathbf{a}_{i,\perp} + \mathbf{a}_{i,\perp}^T V\mathbf{x}_0\big)\,\gamma_i\,z_i. \tag{31c}$$

and, we denote $\mathbf{a}_{i,\perp} = P^{\perp}\mathbf{a}_i$ and $z_i = f(\gamma_i) - \mu(\gamma_i^2 - 1)$. In Sec. B.2 we upper bound $\mathbb{E}[\sup_{V\in\mathcal{E}} \text{Term } \star]$ for all three terms, i.e., $\star = I, II, III$; each one giving rise to one of the terms in the final upper bound of the following lemma.

Lemma 3.2 (Upper bound)**.** *Let υ, χ and $\boldsymbol{\eta}$ be defined as in the statement of Thm. 3.1. Then,*

$$\mathbb{E}\sup_{V\in\mathcal{E}} \langle \mathbf{y} - \mu\,\mathcal{A}(X_0), \mathcal{A}(V)\rangle \leq \sqrt{m}\,\big(\chi + 2\,\mathbb{E}\big[\omega_e(\mathcal{C}_0;\boldsymbol{\eta})\big] + \sqrt{2}\,\upsilon\,\omega_g(\mathcal{C}_0)\big). \tag{32}$$

While we defer the proof to the supplementary material, it is not hard to see that (30) is key in bounding the supremum of Term I by $\chi\sqrt{m}$, rather than a trivial bound of order m.

Putting things together: First, Lemma 3.1 and (13) imply that there exists constant $c > 0$ such that, with probability at least 0.995, we have: $\inf_{V\in\mathcal{E}} \frac{1}{m}\|\mathcal{A}(V)\|_2^2 \geq c$. Second, Lemma 3.2 combined with Markov's inequality imply that, with probability at least 0.995, we have: $\sup_{V\in\mathcal{E}}\langle \mathbf{y} - \mu\cdot\mathcal{A}(X_0), \mathcal{A}(V)\rangle \leq C\cdot\sqrt{m}\cdot\big(\chi + \sqrt{2}\cdot\upsilon\cdot\omega_g(\mathcal{C}_0) + 2\mathbb{E}\big[\omega_e(\mathcal{E};\boldsymbol{\eta})\big]\big)$. Then, by union bound, (28) holds with probability 0.99.

3.3 Controlling the weighted empirical width

Note that $\omega_e(\mathcal{C};\mathbf{p})$ depends both on the geometry of the cone $\mathcal{C}$ and on the weights p_i. We present two ways of controlling $\omega_e(\mathcal{C};\mathbf{p})$ in terms of simpler geometric quantities.

3.3.1 First bound: generic chaining

This general bound captures the geometry by the γ_1 and γ_2-functionals with respect to appropriate metrics, and quantifies the role of the weights by ℓ_∞ and ℓ_2-norms of $\mathbf{p}$.

Lemma 3.3 (Generic chaining bound)**.** *For a cone $\mathcal{C} \subset \mathbb{S}_n$ and $\mathbf{p} \in \mathbb{R}^m$ there exists universal constant $C > 0$ such that $\omega_e(\mathcal{C};\mathbf{p}) \leq (C/\sqrt{m})\cdot\big(\|\mathbf{p}\|_2\cdot\gamma_2(\mathcal{C};\|\cdot\|_F) + \|\mathbf{p}\|_\infty\cdot\gamma_1(\mathcal{C};\|\cdot\|_2)\big)$. In particular, $\omega_e(\mathcal{C}) = \omega_e(\mathcal{C};\mathbf{1}) \leq C\cdot\big(\gamma_2(\mathcal{C};\|\cdot\|_F) + \frac{\gamma_1(\mathcal{C};\|\cdot\|_2)}{\sqrt{m}}\big)$.*

See Sec. C of the supplementary for a proof. Note that further using the crude bound $\|\mathbf{p}\|_\infty \leq \|\mathbf{p}\|_2$ implies:

$$\omega_e(\mathcal{C};\mathbf{p}) \leq \frac{C\|\mathbf{p}\|_2}{\sqrt{m}}\cdot\big(\gamma_2(\mathcal{C};\|\cdot\|_F) + \gamma_1(\mathcal{C};\|\cdot\|_2\big). \tag{33}$$

3.3.2 Second bound: polarity

Alternatively, one can apply polarity arguments, by extending recent results regarding the Gaussian width to more general notions such as the weighted empirical width. The idea is as follows. By using polarity it can be shown that $\omega_e(\mathcal{C};\mathbf{p})$ is upper bounded by the expected distance of $H_{\mathbf{p}}$ (cf. Definition 3) to the polar cone $\mathcal{C}^\circ$. When $\mathcal{C}$ is the tangent cone of some convex proper function (say) $\mathcal{R}$, then, $\mathcal{C}^\circ$ is the cone of subdifferential of $\mathcal{R}$. Thus, we arrive at the result below that is based on techniques from [26, 7].

Proposition 3.1 (Polarity bound)**.** *Let $\mathcal{R} : \mathbb{S}_n \to \mathbb{R}$ be a proper convex function, fix $X_0 \in \mathbb{S}_n$, and let $\widetilde{\mathcal{K}}_{\mathcal{R}} := \{V : \mathcal{R}(X_0 + V) \leq \mathcal{R}(X_0)\}$. Assume that the subdifferential $\partial\mathcal{R}(X_0)$ is non-empty and does not contain the origin. Then, for $H_{\mathbf{p}} = \frac{1}{\sqrt{m}}\sum_{i\in[m]} p_i\cdot\varepsilon_i\mathbf{a}_i\mathbf{a}_i^T$ it holds:*

$$\omega_e^2\big(\mathcal{D}(\widetilde{\mathcal{K}}_{\mathcal{R}}, X_0);\mathbf{p}\big) \leq \mathbb{E}\big[\inf_{\lambda\geq 0}\ \inf_{V\in\partial\mathcal{R}(X_0)} \big\|H_{\mathbf{p}} - \lambda\cdot V\big\|_F^2\big].$$

For illustration, in Lemmas 3.4 and 3.5 below we apply Proposition 3.1 to obtain bounds for the following two cones: (a) $\mathcal{D}(\mathcal{K}_0, \mu X_0)$; and, (b) $\mathcal{D}(\mathcal{K}_{\mathcal{R}}, \mu X_0)$. We defer the proofs to Sec. D in the supplementary material.

Lemma 3.4 (Weighted empirical width of $\mathcal{C}_+$)**.** *For* $\mathbf{p} := (p_1, \ldots, p_m)$ *and* $\mathcal{C}_+ = \{V : \mu X_0 + V \succeq 0 \text{ and } \mathrm{tr}(\mathrm{V}) \leq 0\}$, *there exists constant* $C > 0$ *such that* $\omega_e(\mathcal{C}_+; \mathbf{p}) \leq \frac{C}{\sqrt{m}} (\|\mathbf{p}\|_2 \sqrt{n} + \|\mathbf{p}\|_\infty n)$. *In particular,* $\omega_e(\mathcal{C}_+; \mathbf{1}) \leq C\sqrt{n}$, *provided that* $m \geq cn$ *for some constant* $c > 0$.

Lemma 3.5 (Weighted empirical width of $\mathcal{C}_{\text{sparse}}$)**.** *Let* $X_0 = \mathbf{x}_0 \mathbf{x}_0^T$ *where* $\mathbf{x}_0$ *is* k*-sparse. For* $\mathbf{p} := (p_1, \ldots, p_m)$ *and the cone* $\mathcal{C}_{\text{sparse}} := \{V : \|\mu X_0 + V\|_1 \leq \|\mu X_0\|_1\}$, *there exists constant* $C > 0$ *such that* $\omega_e(\mathcal{C}_{\text{sparse}}; \mathbf{p}) \leq \frac{C}{\sqrt{m}} k \sqrt{\log(n/k)} \left(\|\mathbf{p}\|_2 + \|\mathbf{p}\|_\infty \sqrt{2 \log(n/k)} \right)$. *In particular,* $\omega_e(\mathcal{C}_{\text{sparse}}, \mathbf{1}) \leq Ck\sqrt{\log(n/k)}$, *provided that* $m \geq ck^2 \log(n/k)$, *for* $c > 0$.

3.4 Signal recovery with imperfect tuning

Thm. 3.1 assumes ideal tuning $\widetilde{\mu} = \mu$ in (9), which guarantees that X_0 belongs to the *boundary* of the constraint set $\mathcal{K}_0$ (cf. (21)). Moreover, its bounds in terms of the Gaussian/empirical widths of $\mathcal{D}(\mathcal{K}_\mathcal{R}, \mathbf{x}_0)$ are most informative when X_0 is at the boundary of $\mathcal{K}_\mathcal{R}$. Thm. 3.2 below relaxes these assumptions, via a local analysis that gives rise to the *local* versions of the Gaussian/empirical widths. The proof is deferred to the supplementary material. For example, the theorem shows that the error performance of (9) is controlled as long as $\widetilde{\mu} \geq \mu$ (but, not necessarily equal).

Theorem 3.2 (Imperfect tuning)**.** *Suppose that* $\mathbf{a}_i \sim \mathcal{N}(0, \mathbf{I}_n)$, *and that* $\mathbf{y}$ *follows the model in* (1). *Recall the definitions of* μ *and of* τ, υ, χ. *Further let*

$$\mathcal{K}_{\mathbf{x}_0} := \{\mathcal{K}_0 - \mu X_0\} \cap \{\mathcal{K}_\mathcal{R} - \mu X_0\}. \tag{34}$$

Finally, define $\boldsymbol{\eta} := (\eta_1, \ldots, \eta_m)$, *where* $\eta_i := \big(f(\gamma_i) - \mu \cdot (\gamma_i^2 - 1)\big)$, $i \in [m]$ *for* $\gamma_i \overset{iid}{\sim} \mathcal{N}(0, 1)$. *For any* $t > 0$, *there exist constants* $c, C > 0$ *such that, if the number of observations obeys*

$$m \geq c \cdot \big(\omega_{e,t}(\mathcal{K}_{\mathbf{x}_0})/t\big)^2, \tag{35}$$

then, with high probability, the solution $\widehat{X}$ *of* (9) *satisfies:*

$$\|\widehat{X} - \mu X_0\|_F \leq \frac{C}{\sqrt{m}} \cdot \left(\mathbb{E}_{\boldsymbol{\eta}} \big[\frac{\omega_{e,t}(\mathcal{K}_{\mathbf{x}_0}; \boldsymbol{\eta})}{t} \big] + \upsilon \sqrt{2} \, \frac{\omega_{g,t}(\mathcal{K}_{\mathbf{x}_0})}{t} + \chi \right) + t.$$

4 Simulations

In this section, we show numerical results regarding the performance of the proposed method. In order to solve the optimization in (9), we use PhasePack library [6] (designed for quadratic measurements) with appropriate modifications to account for the different objective function compared to the PhaseLift [5] (cf. Remark 2), as well as for the regularization in case of structured signal recovery.

In our setup, the unknown unit-norm signal $\mathbf{x}_0$ has dimension $n = 50$. m measurements are generated according to (1) with $\mathbf{a}_i \sim \mathcal{N}(0, \mathbf{I})$, for the following link functions: (1) $f^{(1)}(u) = \big(\lfloor |u| \rfloor + \frac{1}{2}\big) \cdot \mathbb{1}_{\{|u| < 3\}} + \frac{7}{2} \cdot \mathbb{1}_{\{|u| \geq 3\}}$; and (2) $f^{(2)}(u) = \frac{1}{2} \cdot \big(\lfloor 2|u|^2 \rfloor + \frac{1}{2}\big) \cdot \mathbb{1}_{\{|u|^2 < 4\}} + \frac{17}{4} \cdot \mathbb{1}_{\{|u|^2 \geq 4\}}$. Note that $f^{(1)}$ and $f^{(2)}$ correspond to quantized measurements with alphabets sizes 4 and 9, respectively. We also experiment with random link functions as per (2); in particular, $y_i \sim \mathrm{Poisson}(|\mathbf{a}_i^T \mathbf{x}_0|^2)$, for all $i \in [m]$. In all cases, we solve (9) with $\tilde{\mu} = \mu$, with appropriate μ for each link function. The results shown are averages over 40 trials.

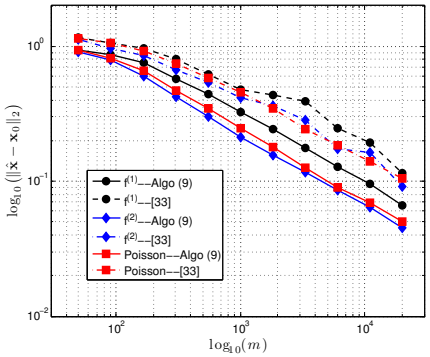

Figure 1: Performance for unstructured signal recovery.

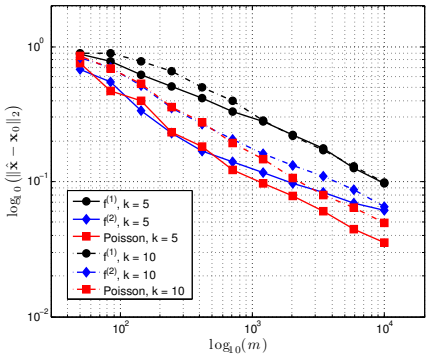

Figure 2: Performance of (9) for sparse signal recovery.

Figure 1 shows the estimation error of (9) as a function of m for a generic $\mathbf{x}_0$ (aka no constraint $X \in \mathcal{K}_\mathcal{R}$). We also compare the proposed method with the PCA-based method in [33, 28]; we empirically observe that the former method outperforms the latter for all tested link functions. Similarly, Figure 2 illustrates the performance of (9) in a sparse-signal recovery setting with sparsity levels $k \in \{5, 10\}$.

Our simulations also confirm the point briefly made in Sec. 1.3. The precise nature of our results can be useful in optimal design of system parameters. In particular, our analysis suggests that link functions with smaller effective noise parameter τ/μ result in smaller error. Indeed, the ratio τ/μ of $f^{(1)}$ and $f^{(2)}$ is equal to 2.25 and 1.51, respectively, which is in agreement to the better recovery performance of $f^{(2)}$ in our simulations.

Acknowledgements

C.T. would like to thank Pablo Parrilo and Joel Tropp for helpful discussions.

References

[1] Dennis Amelunxen, Martin Lotz, Michael B. Mccoy, and Joel A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference*, 3(3):224–294, Sep. 2014.

[2] Sohail Bahmani and Justin Romberg. Phase Retrieval Meets Statistical Learning Theory: A Flexible Convex Relaxation. *arXiv preprint arXiv:1610.04210*, 2016.

[3] Radu Balan, Bernhard G Bodmann, Peter G Casazza, and Dan Edidin. Painless reconstruction from magnitudes of frame coefficients. *Journal of Fourier Analysis and Applications*, 15(4):488–501, 2009.

[4] David R Brillinger. A generalized linear model with "gaussian" regressor variables. *A Festschrift For Erich L. Lehmann*, page 97, 1982.

[5] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[6] Rohan Chandra, Ziyuan Zhong, Justin Hontz, Val McCulloch, Christoph Studer, and Tom Goldstein. PhasePack: A phase retrieval library. *Asilomar Conference on Signals, Systems, and Computers*, 2017.

[7] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[8] Murat A Erdogdu, Lee H Dicker, and Mohsen Bayati. Scaled least squares estimator for glms in large-scale problems. In *Advances in Neural Information Processing Systems*, pages 3324–3332, 2016.

[9] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer Science & Business Media, 2013.

[10] Martin Genzel. High-dimensional estimation of structured signals from non-linear observations with general convex loss functions. *IEEE Transactions on Information Theory*, 63(3):1601–1619, 2017.

[11] Larry Goldstein, Stanislav Minsker, and Xiaohan Wei. Structured signal recovery from non-linear and heavy-tailed measurements. *arXiv preprint arXiv:1609.01025*, 2016.

[12] Tom Goldstein and Christoph Studer. PhaseMax: Convex Phase Retrieval via Basis Pursuit. *arXiv preprint arXiv:1610.07531*, 2016.

[13] Daniel Hsu, Sham Kakade, Tong Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

[14] Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *The Annals of Statistics*, pages 1009–1052, 1989.

[15] Yue M. Lu and Gen Li. Phase transitions of spectral initialization for high-dimensional nonconvex estimation. *arXiv preprint arXiv:1702.06435*, 2017.

[16] Shahar Mendelson. Learning without concentration. *Journal of the ACM (JACM)*, 62(3):21, 2015.

[17] Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. *arXiv preprint arXiv:1708.05932*, 2017.

[18] Matey Neykov, Zhaoran Wang, and Han Liu. Agnostic estimation for misspecified phase retrieval models. In *Advances in Neural Information Processing Systems*, pages 4089–4097, 2016.

[19] Samet Oymak. Learning compact neural networks with regularization. *arXiv preprint arXiv:1802.01223*, 2018.

[20] Samet Oymak, Amin Jalali, Maryam Fazel, Yonina C Eldar, and Babak Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory*, 61(5):2886–2908, 2015.

[21] Samet Oymak and Mahdi Soltanolkotabi. Fast and reliable parameter estimation from nonlinear observations. *arXiv preprint arXiv:1610.07108*, 2016.

[22] Yaniv Plan and Roman Vershynin. The generalized lasso with non-linear observations. *arXiv preprint arXiv:1502.04071*, 2015.

[23] Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *arXiv preprint arXiv:1404.3749*, 2014.

[24] Mohammadreza Soltani and Chinmay Hegde. Fast algorithms for demixing sparse signals from nonlinear observations. *IEEE Transactions on Signal Processing*, 2017.

[25] Mahdi Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *arXiv preprint arXiv:1702.06175*, 2017.

[26] Mihailo Stojnic. Various thresholds for ℓ_1-optimization in compressed sensing. *arXiv preprint arXiv:0907.3666*, 2009.

[27] Michel Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media, 2006.

[28] Yan Shuo Tan. Sparse phase retrieval via sparse pca despite model misspecification: A simplified and extended analysis. *arXiv preprint arXiv:1712.04106*, 2017.

[29] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. LASSO with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems*, pages 3420–3428, 2015.

[30] Joel A Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, pages 67–101. Springer, 2015.

[31] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2017.

[32] Zhuoran Yang, Krishnakumar Balasubramanian, and Han Liu. High-dimensional non-gaussian single index models via thresholded score function estimation. In *International Conference on Machine Learning*, pages 3851–3860, 2017.

[33] Zhuoran Yang, Krishnakumar Balasubramanian, Zhaoran Wang, and Han Liu. Estimating high-dimensional non-gaussian multiple index models via stein's lemma. In *Advances in Neural Information Processing Systems*, pages 6099–6108, 2017.

[34] Zhuoran Yang, Lin F Yang, Ethan X Fang, Tuo Zhao, Zhaoran Wang, and Matey Neykov. Misspecified nonconvex statistical optimization for phase retrieval. *arXiv preprint arXiv:1712.06245*, 2017.

[35] Xinyang Yi, Zhaoran Wang, Constantine Caramanis, and Han Liu. Optimal linear estimation under unknown nonlinear transform. In *Advances in neural information processing systems*, pages 1549–1557, 2015.

5 Supplementary Material

A Proofs for Section 2

A.1 Proof of Theorem 2.1

We start from Theorem 3.1 and we directly control the weighted empirical width using the polarity strategy of Section 3.3.2. Specifically, we use Lemma 3.5.

Note that $\mathcal{C}_{\text{sparse}}$ is the tangent cone of the constraint set $\mathcal{K}_{\mathcal{R}} = \{X \,:\, \|X\|_1 \leq \mu\|X_0\|_1\}$ at μX_0. Also, since $\mathcal{C}_0 \subset \mathcal{C}_{\text{sparse}}$, it holds $\omega_e(\mathcal{C}_0;\mathbf{p}) \leq \omega_e(\mathcal{C}_{\text{sparse}};\mathbf{p})$; thus the lemma directly applies to Theorem 3.1. In more detail, we have

$$\mathbb{E}_{\boldsymbol{\eta}}[\omega_e(\mathcal{C}_0;\boldsymbol{\eta})] \leq C\cdot k\cdot\sqrt{\log(n/k)}\cdot\left(\frac{\mathbb{E}_{\boldsymbol{\eta}}\|\boldsymbol{\eta}\|_2}{\sqrt{m}}+\frac{\mathbb{E}_{\boldsymbol{\eta}}\|\boldsymbol{\eta}\|_\infty}{\sqrt{m}}\sqrt{2\log(n/k)}\right). \tag{36}$$

Recall the definitions of $\boldsymbol{\eta}$ and note that

$$\mathbb{E}_{\boldsymbol{\eta}}\|\boldsymbol{\eta}\|_2 \leq \sqrt{\mathbb{E}_{\boldsymbol{\eta}}\|\boldsymbol{\eta}\|_2^2} \leq \sqrt{m}\cdot\mathbb{E}[\eta_1^2] = \sqrt{m}\tau.$$

Thus using the crude bound $\|\boldsymbol{\eta}\|_\infty \leq \|\boldsymbol{\eta}\|_2$ results in the following:

$$\mathbb{E}_{\boldsymbol{\eta}}[\omega_e(\mathcal{C}_0;\boldsymbol{\eta})] \leq C_2\cdot\tau\cdot k\cdot\sqrt{\log(n/k)}\left(1+\sqrt{2\log(n/k)}\right). \tag{37}$$

It only remains to compute the Gaussian width of $\mathcal{C}_{\text{sparse}}$. It is know that (e.g., [7, Prop. 3.10])

$$\omega_g(\mathcal{C}_{\text{sparse}}) \leq 3\sqrt{2}k\sqrt{\log(n/k)}. \tag{38}$$

Putting (37) and (38) in (24), we have shown that if $m \geq ck^2\log(n/k)$ it holds with high probability:

$$\|\widehat{X}-\mu X_0\|_F \leq C_1\cdot\left(\frac{C_2\cdot\tau\cdot k\sqrt{\log(n/k)}\left(1+\sqrt{2\log(n/k)}\right)}{\sqrt{m}} \;+\; \frac{6\upsilon\cdot k\sqrt{\log(n/k)}+\chi}{\sqrt{m}}\right). \tag{39}$$

Then, the statement of Theorem 2.1 follow from (39) by taking n (in fact n/k) large enough (which will depend on C_2, τ, ρ and χ).

A.2 Proof of Theorem 2.2

We apply Theorem 3.1, but we need to control the weighted empirical width. Recall the definition of $\mathcal{C}_0$ from (22). Note that $\mathcal{C}_0 = \mathcal{C}_+ \cap \mathcal{C}_{\mathcal{R}}$, where

$$\mathcal{C}_+ = \mathcal{D}(\mathcal{K}_0,\mu X_0) = \{V : \mu X_0 + V \succeq 0 \text{ and } \mathrm{tr(V)} \leq 0\}$$

and

$$\mathcal{C}_{\mathcal{R}} = \mathcal{D}(\mathcal{K}_{\mathcal{R}},\mu X_0) = \{V : \mathcal{R}(\mu X_0+V) \leq \mathcal{R}(\mu X_0)\}.$$

Hence

$$\omega_e(\mathcal{C}_0;\mathbf{p}) \leq \min\{\omega_e(\mathcal{C}_+;\mathbf{p}),\omega_e(\mathcal{C}_{\mathcal{R}};\mathbf{p})\}.$$

First, we prove that

$$(18) \Rightarrow (23). \tag{40}$$

On the one hand, if $\sqrt{m} \geq c\sqrt{n}$, then from polarity arguments in Lemma 3.4 we have that $\omega_e(\mathcal{C}_+;\mathbf{1}) \leq C\sqrt{n} \leq C'\sqrt{m}$. On the other hand, if

$$\sqrt{m} \geq c\left(\gamma_2(\mathcal{C}_{\mathcal{R}}\cap\mathcal{S}_{F,n};\|\cdot\|_F)+\gamma_1(\mathcal{C}_{\mathcal{R}}\cap\mathcal{S}_{F,n};\|\cdot\|_2)\right),$$

then by Lemma 3.3:

$$\begin{aligned}\omega_e(\mathcal{C}_{\mathcal{R}}\cap\mathcal{S}_{F,n}) &\leq C\left(\gamma_2(\mathcal{C}_{\mathcal{R}}\cap\mathcal{S}_{F,n};\|\cdot\|_F)+\gamma_1(\mathcal{C}_{\mathcal{R}}\cap\mathcal{S}_{F,n};\|\cdot\|_2)/\sqrt{m}\,\right)\\ &\leq C'\sqrt{m}.\end{aligned} \tag{41}$$

Thus, we have shown (40).

Next, we show that

$$\mathbb{E}_{\boldsymbol{\eta}}[\omega_e(\mathcal{C}_0;\boldsymbol{\eta})] \le C\tau \, \min\big\{\sqrt{n}, \gamma_2(\mathcal{C}_{\mathcal{R}} \cap \mathcal{S}; \|\cdot\|_F) + \gamma_1(\mathcal{C}_{\mathcal{R}} \cap \mathcal{S}; \|\cdot\|_2)\big\}. \tag{42}$$

We will repeatedly use the fact that

$$\mathbb{E}_{\boldsymbol{\eta}}\|\boldsymbol{\eta}\|_\infty \le \mathbb{E}_{\boldsymbol{\eta}}\|\boldsymbol{\eta}\|_2 \le \tau\sqrt{m}.$$

On the one hand, if $\sqrt{m} \ge c\sqrt{n}$ then from Lemma 3.4:

$$\mathbb{E}_{\boldsymbol{\eta}}[\omega_e(\mathcal{C}_0;\boldsymbol{\eta})] \le C \cdot \tau \frac{\sqrt{n}}{\sqrt{m}}(1+\sqrt{n}) \le C'\tau\sqrt{n}.$$

On the other hand, from Eqn. (33),

$$\mathbb{E}_{\boldsymbol{\eta}}[\omega_e(\mathcal{C}_{\mathcal{R}};\boldsymbol{\eta})] \le C \cdot \tau\big(\gamma_2(\mathcal{C}_{\mathcal{R}} \cap \mathcal{S}; \|\cdot\|_F) + \gamma_1(\mathcal{C}_{\mathcal{R}} \cap \mathcal{S}; \|\cdot\|_2)\big).$$

Thus, we have shown (42).

Finally, the proof of the theorem is complete by establishing that

$$\omega_g(\mathcal{C}_0) \le \min\{\sqrt{n}, \omega_g(\mathcal{C}_{\mathcal{R}})\}.$$

This follows using $\omega_g(\mathcal{C}_0) \le \min\{\omega_e(\mathcal{C}_+), \omega_e(\mathcal{C}_{\mathcal{R}})\}$ and the well-known bound $\omega_g(\mathcal{C}_+) \le 6\sqrt{n}$ (e.g. [7]).

B Proofs for Section 3

B.1 Proof of Lemma 3.1

Here, we will prove the following slightly more general version of Lemma 3.1. The modification is needed for the proof of the lower bound when imperfect tuning is studied in Sec. 3.4. Lemma 3.1 follows immediately by setting $t = 1$.

Lemma B.1 (lower bound)**.** *Fix $\mathcal{C} \in \mathbb{S}_n$. For any $t > 0$, there exists positive absolute constant $c > 0$ such that for any $s > 0$ it holds with probability at least $1 - e^{-s^2/4}$:*

$$\inf_{V \in \mathcal{C} \cap t\mathcal{S}_{F,n}} \frac{\|\mathcal{A}(V)\|_2}{t} \ge c\sqrt{m} - 4\,\frac{\omega_{e,t}(\mathcal{C}_0)}{t} - s. \tag{43}$$

To prove the lemma, we need with the following result, which is an application of [16, Thm. 5.4] to our setting. In particular, we follow here the exposition in [30, Prop. 5.1] with minor modifications (that are omitted for brevity).

Lemma B.2 (Lower bound for a nonnegative empirical process [16])**.** *Fix a set $\mathcal{C} \in \mathbb{S}_n$ and let $\mathcal{C}_t := \mathcal{C} \cap t\mathcal{S}_{F,n}$ for some $t > 0$. Let $\mathbf{a} \sim \mathcal{N}(0, \mathbf{I})$, and $\mathbf{a}_1, \dots, \mathbf{a}_m$ be independent copies of $\mathbf{a}$. Introduce the marginal tail function*

$$Q_\xi(\mathcal{C}_t; \mathbf{a}) := \inf_{V \in \mathcal{C}_t} \mathbb{P}(\, |\langle \mathbf{a}\mathbf{a}^T - \mathbf{I}, V\rangle| \ge \xi\,), \quad \textit{where } \xi \ge 0.$$

Let $\varepsilon_1, \dots, \varepsilon_m$ be independent Rademacher random variables, independent from everything else. Then, for any $\xi > 0$ and $s > 0$, with probability at least $1 - e^{-s^2/2}$ it holds,

$$\inf_{V \in \mathcal{C}_t} \|\mathcal{A}(V)\|_2 \ge \xi\sqrt{m} \cdot Q_{2\xi}(\mathcal{C}_t; \mathbf{a}) - 4\omega_{e,t} - \xi s. \tag{44}$$

Proof of Lemma B.1: We will apply (44) for $\xi = t$. We only need to show that there exists absolute constant $c_0 > 0$ such that $Q_{2t}(\mathcal{C}_t; \mathbf{a}) \ge c_0$. Let us denote $\tilde{A} = \mathbf{a}\mathbf{a}^T - \mathbf{I}$. By application of Paley-Zygmund inequality, followed by Gaussian hypercontractivity of second-order Gaussian chaos, it can be shown as in [30, Sec. 8.5.1] that

$$\Pr\Big(|\langle \tilde{A}, V\rangle|^2 \ge \frac{1}{2}\mathbb{E}[|\langle \tilde{A}, V\rangle|^2]\Big) \ge \frac{1}{4}\frac{\mathbb{E}[|\langle \tilde{A}, V\rangle|^2]^2}{\mathbb{E}[|\langle \tilde{A}, V\rangle|^4]} \ge \frac{1}{4 \cdot 3^4}.$$

Moreover, an easy calculation shows that $\mathbb{E}[|\langle \tilde{A}, V\rangle|^2] = 2\|V\|_F^2$ (cf. (58)). Combining these two and noting that $V \in \mathcal{C}_t \implies \|V\|_F = t$, shows the desired. □

B.2 Proof of Lemma 3.2

Here, we present the proof of Lemma 3.2. Recall from Section 3.2 that

$$\begin{aligned}\langle \mathbf{y} - \mu \cdot \mathcal{A}(X_0), \mathcal{A}(V)\rangle &= \sum_{i\in[m]} (\mathbf{a}_i^T V \mathbf{a}_i - \mathrm{tr(V)}) \cdot \big(f(\gamma_i) - \mu\cdot(\gamma_i^2-1)\big) \\ &= \sum_{i\in[m]} \big((P\mathbf{a}_i)^T V (P\mathbf{a}_i) + (P^\perp \mathbf{a}_i)^T V (P^\perp \mathbf{a}_i) + 2\cdot (P\mathbf{a}_i)^T V (P^\perp \mathbf{a}_i) - \mathrm{tr(V)}\big) \cdot \big(f(\gamma_i) - \mu\cdot(\gamma_i^2-1)\big) \\ &= (\text{Term I}) + (\text{Term II}) + (\text{Term III}),\end{aligned} \tag{45}$$

where $\gamma_i := \mathbf{a}_i^T \mathbf{x}_0$ and the three terms are as defined in (46). Note that $\gamma_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. In what follows, we separately upper bound each one of the three terms, which we repeat here for the reader's convenience:

$$\text{Term I} := \sum_{i\in[m]} \big(\mathbf{x}_0^T V \mathbf{x}_0\big)\cdot(\gamma_i^2-1)\cdot\big(f(\gamma_i) - \mu\,(\gamma_i^2-1)\big), \tag{46a}$$

$$\text{Term II} := \sum_{i\in[m]} \big((P^\perp\mathbf{a}_i)^T V (P^\perp\mathbf{a}_i) + \mathbf{x}_0^T V \mathbf{x}_0 - \mathrm{tr(V)}\big)\cdot\big(f(\gamma_i) - \mu\cdot(\gamma_i^2-1)\big), \tag{46b}$$

$$\text{Term III} := \sum_{i\in[m]} \big(\mathbf{x}_0^T V (P^\perp\mathbf{a}_i) + (P^\perp\mathbf{a}_i)^T V \mathbf{x}_0\big)\cdot\gamma_i\cdot\big(f(\gamma_i) - \mu\cdot(\gamma_i^2-1)\big). \tag{46c}$$

1. **Bounding** Term I **:** For convenience, denote

$$\xi_i := (\gamma_i^2-1)(f(\gamma_i) - \mu\cdot(\gamma_i^2-1)). \tag{47}$$

Note that by definition of μ in (10):

$$\mathbb{E}[\xi_i] = 0,\ i\in[m]. \tag{48}$$

Then,

$$\begin{aligned}\mathbb{E}\sup_{V\in\mathcal{E}} \sum_{i\in[m]} \xi_i\cdot \mathbf{x}_0^T V \mathbf{x}_0 &\le \mathbb{E}\sup_{V\in\mathcal{E}} \|V\|_F \cdot \Big|\sum_{i\in[m]}\xi_i\Big| \\ &\overset{(i)}{\le} \mathbb{E}\Big|\sum_{i\in[m]}\xi_i\Big| = \mathbb{E}\sqrt{\Big(\sum_{i\in[m]}\xi_i\Big)^2} \\ &\overset{(ii)}{\le} \sqrt{\mathbb{E}\Big(\sum_{i\in[m]}\xi_i\Big)^2} \\ &\overset{(iii)}{=} \sqrt{\sum_{i\in[m]}\mathbb{E}\xi_i^2} \\ &= \sqrt{m}\cdot\underbrace{\sqrt{\mathbb{E}\xi_1^2}}_{=\chi},\end{aligned} \tag{49}$$

where: (i) follows from $\|\mathbf{x}_0\|_2 = 1$ and $V\in\mathcal{E} \implies \|V\|_F = 1 \implies \sup_{V\in\mathcal{E}} \mathbf{x}_0^T V \mathbf{x}_0 \le 1$; (ii) follows from Jensen's inequality; and, (iii) by independence of the ξ_i's and (48).

2. **Bounding** Term II **:** Note that that set of random variables $\{\gamma_i = \mathbf{a}_i^T\mathbf{x}_0\}_{i\in[m]}$ are independent of the random vectors $\{P^\perp\mathbf{a}_i\}_{i\in[m]}$ since,

$$\mathbb{E}P^\perp\mathbf{a}_i(\mathbf{a}_i^T\mathbf{x}_0) = P^\perp\mathbb{E}\mathbf{a}_i\mathbf{a}_i^T\mathbf{x}_0 = P^\perp\mathbf{x}_0 = \mathbf{0}. \tag{50}$$

Therefore, given two independent sets of measurement vectors $\{\mathbf{a}_i\}_{i\in[m]}$ and $\{\widetilde{\mathbf{a}}_i\}_{i\in[m]}$, the joint distribution of $\{\gamma_i = \mathbf{a}_i^T\mathbf{x}_0, P^\perp\mathbf{a}_i\}_{i\in[m]}$ is identical to the joint distribution of $\{\gamma_i = \mathbf{a}_i^T\mathbf{x}_0, P^\perp\widetilde{\mathbf{a}}_i\}_{i\in[m]}$. This allows us to introduce the independent copy of random measurement vectors $\{\widetilde{\mathbf{a}}_i\}_{i\in[m]}$ in (46b) as follows.

$$\begin{aligned}\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\big((P^\perp\widetilde{\mathbf{a}}_i)^TV(P^\perp\widetilde{\mathbf{a}}_i)+\mathbf{x}_0^TV\mathbf{x}_0-\mathrm{tr(V)}\big)\cdot\big(f(\gamma_i)-\mu\cdot(\gamma_i^2-1)\big)\\ =\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\big((P^\perp\widetilde{\mathbf{a}}_i)^TV(P^\perp\widetilde{\mathbf{a}}_i)+\mathbf{x}_0^TV\mathbf{x}_0-\mathrm{tr(V)}\big)\cdot\eta_i,\end{aligned}\tag{51}$$

where the expectation is taken over $\{\gamma_i, P^\perp\widetilde{\mathbf{a}}_i\}_{i\in[m]}$, $\{\widetilde{\mathbf{a}}_i\}_{i\in[m]}$ are independent of $\{\gamma_i\}_{i\in[m]}$, and we have defined

$$\eta_i := \big(f(\gamma_i)-\mu\cdot(\gamma_i^2-1)\big).$$

For vectors $\mathbf{g}_i\sim N(0,I)$, we have $\mathbb{E}(\mathbf{g}_i^T\mathbf{x}_0)^2=1$ and $\mathbb{E}(\mathbf{g}_i{}^T\mathbf{x}_0)=0$; hence we can rewrite (51) in the following manner.

$$\begin{aligned}&\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\big((P^\perp\widetilde{\mathbf{a}}_i)^TV(P^\perp\widetilde{\mathbf{a}}_i)+\big(\mathbb{E}(\mathbf{g}_i^T\mathbf{x}_0)^2\big)\cdot\mathbf{x}_0^TV\mathbf{x}_0+\mathbb{E}(\mathbf{g}_i^T\mathbf{x}_0)\cdot\widetilde{\mathbf{a}}_i^TP^\perp V\mathbf{x}_0-\mathrm{tr(V)}\big)\cdot\eta_i\\ &\overset{(i)}{=}\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\mathbb{E}\Big(\big((P^\perp\widetilde{\mathbf{a}}_i)^TV(P^\perp\widetilde{\mathbf{a}}_i)+(\mathbf{g}_i^T\mathbf{x}_0)^2\cdot\mathbf{x}_0^TV\mathbf{x}_0+\mathbf{g}_i^T\mathbf{x}_0\cdot\widetilde{\mathbf{a}}_i^TP^\perp V\mathbf{x}_0-\\ &\qquad\qquad\mathrm{tr(V)}\big)\cdot\eta_i\mid\{\gamma_i,P^\perp\widetilde{\mathbf{a}}_i\}_{i\in[m]}\Big)\\ &\overset{(ii)}{=}\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\mathbb{E}\Big(\big((P^\perp\widetilde{\mathbf{a}}_i)^TV(P^\perp\widetilde{\mathbf{a}}_i)+(\widetilde{\mathbf{a}}_i^T\mathbf{x}_0)^2\cdot\mathbf{x}_0^TV\mathbf{x}_0+\widetilde{\mathbf{a}}_i^T\mathbf{x}_0\cdot\widetilde{\mathbf{a}}_i^TP^\perp V\mathbf{x}_0-\\ &\qquad\qquad\mathrm{tr(V)}\big)\cdot\eta_i\mid\{\gamma_i,P^\perp\widetilde{\mathbf{a}}_i\}_{i\in[m]}\Big)\\ &\overset{(iii)}{\le}\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\big((P^\perp\widetilde{\mathbf{a}}_i)^TV(P^\perp\widetilde{\mathbf{a}}_i)+(\widetilde{\mathbf{a}}_i^T\mathbf{x}_0)^2\cdot\mathbf{x}_0^TV\mathbf{x}_0+\widetilde{\mathbf{a}}_i^T\mathbf{x}_0\cdot\widetilde{\mathbf{a}}_i^TP^\perp V\mathbf{x}_0-\mathrm{tr(V)}\big)\cdot\eta_i\\ &\overset{(iv)}{\le}\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\big(\widetilde{\mathbf{a}}_i^TV\widetilde{\mathbf{a}}_i-\mathrm{tr(V)}\big)\cdot\eta_i=\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\eta_i\big\langle\widetilde{\mathbf{a}}_i\widetilde{\mathbf{a}}_i^T-\mathbf{I},V\big\rangle.\\ &\overset{(v)}{\le}2\cdot\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\eta_i\big\langle\varepsilon_i\widetilde{\mathbf{a}}_i\widetilde{\mathbf{a}}_i^T,V\big\rangle=2\sqrt{m}\cdot\mathbb{E}_{\boldsymbol{\eta}}\big[\omega_e(\mathcal{C}_0;\boldsymbol{\eta})\big]\end{aligned}\tag{52}$$

where (i) follows from the conditioning over $\{\gamma_i, P^\perp\widetilde{\mathbf{a}}_i\}$ which are independent of the random vector $\mathbf{g}$. Since $\{\widetilde{\mathbf{a}}_i^T\mathbf{x}_0\}_i$ are independent of $\{P^\perp\widetilde{\mathbf{a}}_i\}$ (cf. (50)), we replace $\{\mathbf{g}_i^T\mathbf{x}_0\}$ with $\{\widetilde{\mathbf{a}}_i^T\mathbf{x}_0\}$ in order to obtain (ii). The inequalities (iii) and (iv) follow from Jensen's inequality and the fact that

$$\widetilde{\mathbf{a}}_i = P\widetilde{\mathbf{a}}_i + P^\perp\widetilde{\mathbf{a}}_i = (\widetilde{\mathbf{a}}_i^T\mathbf{x}_0)\mathbf{x}_0 + P^\perp\widetilde{\mathbf{a}}_i.$$

Finally, the inequality in (v) follows from standard symmetrization argument: $\varepsilon_1,\ldots,\varepsilon_m$ are independent Rademacher random variables and $\omega_e(\mathcal{C}_0;\boldsymbol{\eta})$ has been defined in Section 3.1.

3. **Bounding** Term III **:** Let us define

$$\zeta_i := \gamma_i\cdot(f(\gamma_i)-\mu\cdot(\gamma_i^2-1)).$$

Repeating the argument that led to (51), we can take the expectation in $\mathbb{E}[\text{Term III}]$ with respect to $\{\gamma_i = \mathbf{a}_i^T\mathbf{x}_0, \widetilde{\mathbf{a}}_i\}$, where $\{\mathbf{a}_i\}$ and $\{\widetilde{\mathbf{a}}_i\}$ two independent copies of the Gaussian measurement vectors. This allows us to write Term III as

$$\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\zeta_i\cdot\big(\mathrm{tr}((\mathrm{P}^\perp\widetilde{\mathbf{a}}_\mathrm{i})\mathbf{x}_0^\mathrm{T}\mathrm{V})+\mathrm{tr}(\mathbf{x}_0(\mathrm{P}^\perp\widetilde{\mathbf{a}}_\mathrm{i})^\mathrm{T}\mathrm{V})\big).\tag{53}$$

Note that the expectation is take with respect to $\{\gamma_i = \mathbf{a}_i^T\mathbf{x}_0, \widetilde{\mathbf{a}}_i\}$, where $\{\mathbf{a}_i\}$ and $\{\widetilde{\mathbf{a}}_i\}$ two independent copies of Gaussian measurement vectors.

For $i \in [m]$, let $\Gamma_{(i,1)}$ be a random matrix with i.i.d. standard normal vectors as its entries. Since $\mathbf{x}_0$ is assumed to be a unit norm vector, for $i \in [m]$, the distribution of $\widetilde{\mathbf{a}}_i$ is identical to the distribution of random vector $\Gamma_{(i,1)}\mathbf{x}_0$. Therefore, we can express (53) as follows.

$$\begin{aligned}
&\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\zeta_i\cdot\big(\mathrm{tr}(\mathrm{P}^{\perp}\Gamma_{(\mathrm{i},1)}\mathbf{x}_0\mathbf{x}_0^{\mathrm{T}}\mathrm{V})+\mathrm{tr}(\mathbf{x}_0\mathbf{x}_0^{\mathrm{T}}\Gamma_{(\mathrm{i},1)}^{\mathrm{T}}\mathrm{P}^{\perp}\mathrm{V})\big)\\
&\quad=\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\zeta_i\cdot\big(\mathrm{tr}((\mathrm{P}^{\perp}\Gamma_{(\mathrm{i},1)}\mathrm{P}+(\mathrm{P}^{\perp}\Gamma_{(\mathrm{i},1)}\mathrm{P})^{\mathrm{T}}\mathrm{V})\big),
\end{aligned}\tag{54}$$

Let's consider three sets of independent random matrices $\{\Gamma_{(i,2)}\}_{i\in[m]}$, $\{\Gamma_{(i,3)}\}_{i\in[m]}$ and $\{\Gamma_{(i,4)}\}_{i\in[m]}$ that: (a) have i.i.d. standard Gaussian entries; (b) are independent of each other; and (c) are independent of all the random variable that appeared so far. Since, for $i \in [m]$, we have $\mathbb{E}\Gamma_{(i,2)} = \mathbb{E}\Gamma_{(i,3)} = \mathbb{E}\Gamma_{(i,4)} = 0$, we express Term III (cf. (54)) as follows.

$$\begin{aligned}
&\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\zeta_i\cdot\big(\big\langle P^{\perp}\Gamma_{(i,1)}P+P\mathbb{E}\Gamma_{(i,2)}P+P^{\perp}\mathbb{E}\Gamma_{(i,3)}P^{\perp}+P\mathbb{E}\Gamma_{(i,4)}P^{\perp},V\big\rangle\\
&\qquad\qquad+\big\langle(P^{\perp}\Gamma_{(i,1)}P+P\mathbb{E}\Gamma_{(i,2)}P+P^{\perp}\mathbb{E}\Gamma_{(i,3)}P^{\perp}+P\mathbb{E}\Gamma_{(i,4)}P^{\perp})^T,V\big\rangle\big)\\
&\overset{(i)}{\leq}\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\zeta_i\cdot\big(\big\langle P^{\perp}\Gamma_{(i,1)}P+P\Gamma_{(i,2)}P+P^{\perp}\Gamma_{(i,3)}P^{\perp}+P\Gamma_{(i,4)}P^{\perp},V\big\rangle\\
&\qquad\qquad+\big\langle(P^{\perp}\Gamma_{(i,1)}P+P\Gamma_{(i,2)}P+P^{\perp}\Gamma_{(i,3)}P^{\perp}+P\Gamma_{(i,4)}P^{\perp})^T,V\big\rangle\big)\\
&\overset{(ii)}{\leq}\mathbb{E}_{\tilde{G}_i,\zeta_i}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\zeta_i\cdot\big\langle\tilde{G}_i+\tilde{G}_i^T,V\big\rangle=\sqrt{2}\cdot\mathbb{E}_{G_i,\zeta_i}\sup_{V\in\mathcal{E}}\big\langle\sum_{i\in[m]}\zeta_i\cdot G_i,V\big\rangle,
\end{aligned}\tag{55}$$

where: (i) follows from Jensen's inequality; (ii) uses Lemma B.3 that shows $P^{\perp}\Gamma_{(i,1)}P + P\Gamma_{(i,2)}P + P^{\perp}\Gamma_{(i,3)}P^{\perp} + P\Gamma_{(i,4)}P^{\perp} \sim \tilde{G}_i$ for a matrix $\tilde{G}_i$ with entries iid standard Gaussian. Finally, in the right hand side of the last equality $G_i = (\tilde{G}_i + \tilde{G}_i^T)/\sqrt{2}$ is a matrix from the Gaussian orthogonal ensemble (GOE).

Lemma B.3. *Consider $A, B, C, D \in \mathbb{R}^{n\times n}$ that have entries iid standard Gaussian and are independent of each other. Let $P, P^{\perp}$ be orthogonal projections with $P + P^{\perp} = \mathbf{I}$. Then, the matrix*

$$X = P^{\perp}AP + PBP^{\perp} + P^{\perp}CP^{\perp} + PDP^{\perp}$$

has entries iid standard Gaussian.

Proof. We can write $X = X_1P + X_2P^{\perp}$, with $X_1 = P^{\perp}A + PB$ and $X_2 = P^{\perp}C + PD$. We show that X_1, X_2 are independent with entries iid standard Gaussian each. Let $\mathbf{y}^i$ denote the i^{th} column of a matrix Y. Clearly, $\mathbf{x}_1^i$ is a Gaussian vector with mean zero entries. Also,

$$\begin{aligned}
\mathbb{E}[\mathbf{x}_1^i(\mathbf{x}_1^i)^T] &= P^{\perp}\mathbb{E}[\mathbf{a}_1^i(\mathbf{a}_1^i)]^TP^{\perp} + P\mathbb{E}[\mathbf{b}_1^i(\mathbf{b}_1^i)^T]P + P\mathbb{E}[\mathbf{b}_1^i(\mathbf{a}_1^i)^T]P^{\perp}] + P^{\perp}\mathbb{E}[\mathbf{a}_1^i(\mathbf{b}_1^i)^T]P^T\\
&= P^{\perp} + P = \mathbf{I}.
\end{aligned}$$

Thus, $\mathbf{x}_1^i \sim \mathcal{N}(0, \mathbf{I})$. Moreover, $\mathbf{x}_1^i$ is independent of $\mathbf{x}_1^j$ for $i \neq j$. This shows that X_1 has entries iid standard Gaussian. Of course, the same argument shows that this is also true for X_2. Clearly, X_1 is independent of X_2. With these, and repeating the argument above for X_1, it is easy to show that $X = X_1P + X_2P^{\perp}$ has entries iid Gaussian. □

Now condition on $\{\zeta_i\}$. By rotational invariance of the Gaussian measure, $\sum_{i\in[m]}\zeta_i \cdot G_i$ is distributed as $\left(\sum_{i\in[m]}\zeta_i^2\right)^{1/2} G$, where G is a GOE matrix. Thus,

$$(55) \leq \sqrt{2}\cdot\mathbb{E}[(\sum_{i\in[m]}\zeta_i^2)^{1/2}]\cdot\mathbb{E}\sup_{V\in\mathcal{E}}\langle G,V\rangle \leq \sqrt{2}\cdot\underbrace{\sqrt{\mathbb{E}[\zeta_1^2]}}_{=\upsilon}\cdot\sqrt{m}\cdot\underbrace{\mathbb{E}\sup_{V\in\mathcal{E}}\langle G,V\rangle}_{\omega_g(\mathcal{E})},\tag{56}$$

where the last inequality follows from Jensen.

Combining (45) with (49), (55), and (56) we conclude with Lemma 3.2.

B.3 The expected excess loss

It is instructive to see why (9) encourages small $\|\widehat{V}\|_F = \|\widehat{X} - X_0\|_F$ by establishing the following result about the expected excess loss function.

Lemma B.4 (Expected excess loss)**.** *The excess loss of* (9) *satisfies:*

$$\mathbb{E}[\mathcal{L}(\mu X_0) - \mathcal{L}(X)] = \frac{m}{2}\|X - \mu X_0\|_F^2. \tag{57}$$

The lemma implies that μX_0 minimizes the expected loss of (9) among all feasible solutions. In the rest of this section, we prove the lemma by computing the expectation of the two terms in (25). On the one hand, for any $V \in \mathbb{S}_n$:

$$\begin{aligned}\frac{1}{2}\mathbb{E}\|\mathcal{A}(V)\|_2^2 &= \frac{1}{2}\sum_{i=1}^{m}\mathbb{E}\left(\mathbf{a}_i^T V \mathbf{a}_i - \mathrm{tr(V)}\right)^2 \\ &= \frac{m}{2}\cdot\left(3\cdot\sum_{i\in[n]}V_{i,i}^2 + 2\cdot\sum_{i,j\in[n]\,:\,i\neq j}V_{i,j}^2 + \sum_{i,j\in[n]\,:\,i\neq j}V_{i,i}V_{j,j}\right) - \frac{m}{2}(\mathrm{tr(V)})^2 \\ &= m\cdot\|V\|_F^2. \end{aligned}\tag{58}$$

On the other hand, we will show later that

$$\mathbb{E}\langle \mathbf{y} - \mu\cdot\mathcal{A}(X_0), \mathcal{A}(V)\rangle = 0. \tag{59}$$

Of course, (58) and (59) imply the desired.

In the remaining, we show (45). Recall the decomposition in (45). We compute the expectation of each term separately.

1. $\mathbb{E}[\text{Term I}]$ **:** Clearly, by (30): $\mathbb{E}[\text{Term I}] = \sum_{i\in[m]}(\mathbf{x}_0^T V\mathbf{x}_0)\mathbb{E}[\xi_i] = 0$.

2. $\mathbb{E}[\text{Term II}]$ **:** Working as in (50) (introducing independent sets of measurement vectors $\{\mathbf{a}_i\}_{i\in[m]}$ and $\{\widetilde{\mathbf{a}}_i\}_{i\in[m]}$) we compute that

$$\begin{aligned}\mathbb{E}[\text{Term II}] &= \mathbb{E}\sum_{i\in[m]}\left((P^{\perp}\widetilde{\mathbf{a}}_i)^T V(P^{\perp}\widetilde{\mathbf{a}}_i) + \mathbf{x}_0^T V\mathbf{x}_0 - \mathrm{tr(V)}\right)\cdot\left(f(\gamma_i) - \mu\cdot(\gamma_i^2 - 1)\right) \\ &\overset{(i)}{=} \sum_{i\in[m]}\left(\mathbb{E}[(P^{\perp}\widetilde{\mathbf{a}}_i)^T V(P^{\perp}\widetilde{\mathbf{a}}_i)] + \mathbf{x}_0^T V\mathbf{x}_0 - \mathrm{tr(V)}\right)\cdot\mathbb{E}[f(\gamma_i) - \mu\cdot(\gamma_i^2-1)] \\ &= 0.\end{aligned}\tag{60}$$

   where, *(i)* follows by independence of $\{\mathbf{a}_i\}_{i\in[m]}$ and $\{\widetilde{\mathbf{a}}_i\}_{i\in[m]}$.

3. $\mathbb{E}[\text{Term III}]$ **:** Similar for third term, we have:

$$\begin{aligned}\mathbb{E}[\text{Term III}] &= \mathbb{E}\sum_{i\in[m]}\left(\mathrm{tr}((\mathrm{P}^{\perp}\widetilde{\mathbf{a}}_\mathrm{i})\mathbf{x}_0^{\mathrm{T}}\mathrm{V}) + \mathrm{tr}(\mathbf{x}_0(\mathrm{P}^{\perp}\widetilde{\mathbf{a}}_\mathrm{i})^{\mathrm{T}}\mathrm{V})\right)\cdot\left(\gamma_i\cdot(f(\gamma_i) - \mu\cdot(\gamma_i^2-1))\right), \\ &= \sum_{i\in[m]}\left(\mathrm{tr}((\mathrm{P}^{\perp}\mathbb{E}[\widetilde{\mathbf{a}}_\mathrm{i})]\mathbf{x}_0^{\mathrm{T}}\mathrm{V}) + \mathrm{tr}(\mathbf{x}_0(\mathrm{P}^{\perp}\mathbb{E}[\widetilde{\mathbf{a}}_\mathrm{i}])^{\mathrm{T}}\mathrm{V})\right)\cdot\mathbb{E}[\gamma_i\cdot(f(\gamma_i) - \mu\cdot(\gamma_i^2-1))] = 0,\end{aligned}\tag{61}$$

   where, the last equality follows because $\widetilde{\mathbf{a}}_i$'s are centered.

B.4 Proof of Theorem 3.2

The proof is similar to that of Theorem 3.1 and follows the steps in the proof of Theorem [22, Thm. 1.9]; thus, we only highlight the appropriate modifications needed. Fix any $t > 0$. If $\|\widehat{V}\|_F \leq t$, then the error bound holds trivially. Thus, assume that $\alpha := \|\widehat{V}\|_F > t$.

Our starting point is again (26), where we need a lower (upper) bound for the left (right) -hand side of the inequality. First, we show in Lemma B.5 below that under (35) the following event holds with probability at least 0.995:

$$\inf_{V\in\mathcal{K}_{\mathbf{x}_0}\cap t\mathcal{B}_{F,n}^c}\frac{1}{\sqrt{m}}\frac{\|\mathcal{A}(V)\|_2}{\|V\|_F}\geq c_1. \tag{62}$$

Combining this with (26) shows that

$$\begin{aligned}c_1^2\alpha &\leq \frac{2}{m}\,\alpha^{-1}\,\langle \mathbf{y}-\mu\cdot\mathcal{A}(X_0),\mathcal{A}(\widehat{V})\rangle\\ &\leq \frac{2}{m}\sup_{V\in\alpha^{-1}\mathcal{K}_{\mathbf{x}_0}\cap\mathcal{S}_{F,n}}\langle \mathbf{y}-\mu\cdot\mathcal{A}(X_0),\mathcal{A}(\widehat{V})\rangle,\end{aligned} \tag{63}$$

where, we used the homogeneity of the RHS, and the facts that $\widehat{V}\in\mathcal{K}_{\mathbf{x}_0}$ and $\alpha^{-1}\widehat{V}\in\mathcal{S}_{F,n}$. Now, let V_1 be optimal in the maximization in (63). Since $\mathcal{K}_{\mathbf{x}_0}$ is star-shaped and $\alpha>t$, we have that $\alpha V_1\in\mathcal{K}_{\mathbf{x}_0}\implies tV_1\in\mathcal{K}_{\mathbf{x}_0}$. Thus,

$$(63)\leq \frac{2}{m}\sup_{V\in\mathcal{K}_{\mathbf{x}_0}\cap t\mathcal{S}_{F,n}}\langle \mathbf{y}-\mu\cdot\mathcal{A}(X_0),\mathcal{A}(\widehat{V})\rangle. \tag{64}$$

Then, it only takes a slight modification of Lemma 3.2 (details are omitted for brevity) to show that with probability at least 0.995, the expression above is upper bounded by

$$(64)\leq \frac{2}{\sqrt{m}}\big(t\,\chi+2\,\mathbb{E}\big[\omega_{e,t}(\mathcal{K}_{\mathbf{x}_0};\boldsymbol{\eta})\big]+\sqrt{2}\,\upsilon\,\omega_{g,t}(\mathcal{K}_{\mathbf{x}_0})\big).$$

Tracing back the inequalities to (63) shows the desired bound on α and completes the proof of the Theorem 3.2.

Lemma B.5. *Let $\mathcal{C}\subset\mathbb{S}_n$ be a star shaped set, i.e., for each $\lambda\in(0,1)$, we have $\lambda\mathcal{C}\subset\mathcal{C}$. Let $t>0$ and*

$$m\geq c\cdot\big(\omega_{e,t}(\mathcal{C})\big)^2/t^2.$$

Then, with probability at least 0.995, we have

$$\inf_{V\in\mathcal{K}\cap t\mathcal{B}_{F,n}^c}\frac{1}{\sqrt{m}}\frac{\|\mathcal{A}(V)\|_2}{\|V\|_F}\geq c',$$

where $\mathcal{B}_{F,n}^c:=\{U\in\mathbb{S}_n\mid\|U\|_F\geq 1\}$. Here, $c,c'>0$ are absolute constants.

Proof. Since $\mathcal{C}$ is star-shaped, it holds

$$\inf_{V\in\mathcal{K}\cap t\mathcal{B}_{F,n}^c}\frac{\|\mathcal{A}(V)\|_2}{\|V\|_F}=\inf_{V\in\mathcal{K}\cap t\mathcal{S}_{F,n}}\frac{\|\mathcal{A}(V)\|_2}{t}.$$

Thus, we can apply Lemma B.1 and finish the proof by using the assumption $m\geq c\cdot\big(\omega_{e,t}(\mathcal{C})\big)^2/t^2$. □

C Controlling the weighted empirical width via generic chaining

C.1 Background

For a set $\mathcal{T}$, we say that a sequence $\{\mathcal{A}_n\}$ of partitions of $\mathcal{T}$ is *increasing* if every set of $\mathcal{A}_{n+1}$ is contained in a set of $\mathcal{A}_n$.

Definition 4 (Admissible sequence)**.** *Given a set $\mathcal{T}$, an admissible sequence is an increasing sequence $\{\mathcal{A}_n\}$ of partitions of $\mathcal{T}$ such that* $\mathrm{card}(\mathcal{A}_n)\leq N_n:=2^{2^n}$.

Given a partition $\mathcal{A}_n$ of $\mathcal{T}$ and $\mathbf{t}\in\mathcal{T}$, we use $A_n(\mathbf{t})$ to denote the set in $\mathcal{A}_n$ that contains $\mathbf{t}$. With this notation in place, we now define a useful geometric quantity for the metric space $(\mathcal{T},d)$.

Definition 5. *Given $\alpha>0$ and a metric space $(\mathcal{T},d)$, we define*

$$\gamma_\alpha(\mathcal{T},d)=\inf\sup_{\mathbf{t}}\sum_{n\geq 0}2^{n/\alpha}\Delta(A_n(\mathbf{t})), \tag{65}$$

where $\Delta(A_n(\mathbf{t}))$ denotes the diameter of the set $A_n(\mathbf{t})$. The infimum in (65) *is taken over all admissible sequences.*

A popular approach is via Dudley's bound that is expressed in terms of the metric entropy of the set. For a metric space $(\mathcal{T}, d)$ let $\mathcal{N}(\mathcal{T}, d, \epsilon)$ denote the covering number of $\mathcal{T}$ with balls of radius ϵ in metric d. For $\alpha = 1, 2$, there exist universal constant $C(\alpha) > 0$ such that [27, Sec. 1.2]

$$\gamma_\alpha(\mathcal{T}, d) \leq C(\alpha) \cdot \int_0^\infty \big(\log(\mathcal{N}(\mathcal{T}, d, \epsilon))\big)^{1/\alpha} \mathrm{d}\epsilon. \tag{66}$$

Proposition C.1 ([27, Theorem 1.2.7]). *For a set $\mathcal{T}$ with two distances d_1 and d_2, consider a process $\{X_{\mathbf{t}}\}_{\mathbf{t}\in\mathcal{T}}$ such that $\mathbb{E}X_{\mathbf{t}} = 0$ and $\forall\, \mathbf{s}, \mathbf{t} \in \mathcal{T},\ \forall\, u > 0$,*

$$\mathbb{P}\left(|X_{\mathbf{s}} - X_{\mathbf{t}}| \geq u\right) \leq 2e^{-\min\left(\frac{u^2}{d_2(\mathbf{s},\mathbf{t})^2}, \frac{u}{d_1(\mathbf{s},\mathbf{t})}\right)}. \tag{67}$$

Then,

$$\mathbb{E}\sup_{\mathbf{s},\mathbf{t}\in\mathcal{T}} |X_{\mathbf{s}} - X_{\mathbf{t}}| \leq L \cdot \big(\gamma_1(\mathcal{T}, d_1) + \gamma_2(\mathcal{T}, d_2)\big). \tag{68}$$

We consider the following zero-mean process $X_V := \langle \sqrt{m} H_{\mathbf{p}}, V\rangle$ indexed by $V \in \mathbb{S}_n$. We show in Lemma E.2 that the process satisfies:

$$\mathbb{P}\left(\ |X_V - X_U| \geq t\ \right) \leq 2e^{-\min\left\{\frac{t^2}{4\|\mathbf{p}\|_2^2\|V-U\|_F^2},\ \frac{t}{\|\mathbf{p}\|_\infty\|V-U\|_2}\right\}}, \tag{69}$$

for all $t \geq 0$ and $V, U \in \mathbb{S}_n$. Therefore, we can employ Proposition C.1 with distances $d_2(T, U) = \|\mathbf{p}\|_2\|T - U\|_F$ and $d_1(T, U) = \|\mathbf{p}\|_\infty\|T - U\|_2$ to obtain the desired:

$$\begin{aligned}\mathbb{E}\sup_{V\in\mathcal{C}} X_V &\leq L \cdot \big(\gamma_1(\mathcal{C}, \|\mathbf{p}\|_\infty\|\cdot\|_2) + \gamma_2(\mathcal{C}, \|\mathbf{p}\|_2\|\cdot\|_F)\big)\\ &\leq L \cdot \big(\|\mathbf{p}\|_\infty \cdot \gamma_1(\mathcal{C}, \|\cdot\|_2) + \|\mathbf{p}\|_2 \cdot \gamma_2(\mathcal{C}, \|\cdot\|_F)\big).\end{aligned}$$

D Computing the weighted empirical quadratic Gaussian width via polarity

Here, we prove Lemmas 3.4 and 3.5.

D.1 Proof of Lemma 3.4

For an illustration of the applicability of Proposition 3.1 we use it here to control the empirical quadratic Gaussian width of the cone

$$\mathcal{C}_+ = \{V : \mu X_0 + V \succeq 0 \text{ and } \mathrm{tr(V)} \leq 0\}.$$

Recall that the set of feasible directions in (9) is a subset of $\mathcal{C}_+$. In order to put this into the language of Proposition 3.1, note that $\mathcal{C}_+ = \mathcal{D}(\widetilde{\mathcal{K}}_{\mathcal{R}}, \mu X_0)$ for the following convex function $\mathcal{R} : \mathbb{S}_n \to \mathbb{R}$

$$\mathcal{R}(X) := \mathrm{tr(X)} + \begin{cases} 0, & X \succeq 0,\\ +\infty, & \text{else,}\end{cases}$$

where, we also used the fact that $X_0 \succeq 0$ and $\mathrm{tr}(\mu \mathrm{X}_0) = \mu$. The proof of the lemma follows rather standard arguments. Similar calculation are performed in [30, Sec. 8.6.2]. In particular, the second statement of the lemma can also be found in [30].

Proof. By rotational invariance of the Gaussian distribution, we may assume without loss of generality that

$$X_0 = \mathbf{x}_0\mathbf{x}_0^T = \begin{bmatrix}\|\mathbf{x}_0\|_2^2 & \mathbf{0}^T\\ \mathbf{0} & 0\end{bmatrix}.$$

Also, partition H as follows

$$H = \begin{bmatrix}h_{11} & \mathbf{h}_{12}^T\\ \mathbf{h}_{12} & H_{22}\end{bmatrix}.$$

Let

$$\lambda = \lambda_{\max}(H_{22}) \tag{70}$$

the maximum eigenvalue of H_{22}. From Proposition 3.1,

$$\begin{aligned}\omega_e^2(\mathcal{C}_+;\mathbf{p}) &\le \mathbb{E}\big[\inf_{V\in\partial\mathcal{R}(X_0)}\|H-\lambda\cdot V\|_F^2\big]\\ &\overset{(i)}{=} \mathbb{E}\big[(h_{11}-\lambda)^2\big]+2\mathbb{E}\big[\|\mathbf{h}_{12}\|_2^2\big]+\\ &\qquad \mathbb{E}\big[\inf_{S:\ \lambda_{\max}(S)\le 1}\|H_{22}-\lambda S\|_F^2\big]\\ &\overset{(ii)}{=} \mathbb{E}\big[(h_{11}-\lambda)^2\big]+2\cdot\mathbb{E}\big[\|\mathbf{h}_{12}\|_2^2\big].\end{aligned} \tag{71}$$

where: *(i)* uses the fact that (e.g., [30, Sec. 8.6.1])

$$\partial\mathcal{R}(X_0)=\left\{\begin{bmatrix}1 & \mathbf{0}^T\\ \mathbf{0} & S\end{bmatrix}\ :\ \lambda_{\max}(S)\le 1\right\};$$

(ii) uses (70). It remains to compute the two terms in (71). On the one hand, we have

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{h}_{12}\|_2^2\big] &= \frac{1}{m}\mathbb{E}\big[\sum_{\ell=2}^{n}\big(\sum_{i\in[m]}p_i\varepsilon_i\mathbf{a}_{i,1}\mathbf{a}_{i,\ell}\big)^2\big]\\ &= \frac{1}{m}\big[\sum_{\ell=2}^{n}\sum_{i\in[m]}p_i^2\big]=\frac{\|\mathbf{p}\|_2^2}{m}(n-1).\end{aligned} \tag{72}$$

On the other hand, by interlacing of eigenvalues

$$\lambda=\lambda_{\max}(H_{22})\le\lambda_{\max}(H),$$

and, as shown in Lemma E.1:

$$\lambda_{\max}(H)\le\frac{C_1}{\sqrt{m}}\cdot(\|\mathbf{p}\|_2\sqrt{n}+\|\mathbf{p}\|_\infty n).$$

Thus,

$$\begin{aligned}\mathbb{E}\big[(h_{11}-\lambda)^2\big] &= \frac{1}{m}\mathbb{E}\big[\big(\sum_{i\in[m]}p_i\varepsilon_i\mathbf{a}_{i,1}^2\big)^2\big]+\lambda^2\\ &\le 3\frac{\|\mathbf{p}\|_2^2}{m}+\lambda^2.\end{aligned} \tag{73}$$

Putting together (72) and (73) in (71), it follows that

$$\omega_e(\mathcal{C}_+;\mathbf{p})\le\frac{C_2}{\sqrt{m}}\cdot(\|\mathbf{p}\|_2\sqrt{n}+\|\mathbf{p}\|_\infty n),$$

as desired. To prove the remaining statement, let $\mathbf{p}=\mathbf{1}$ above. Note that $\|\mathbf{1}\|_2=\sqrt{m}$ and $\|\mathbf{1}\|_\infty=1$. Thus, provided that $m\ge c\cdot n$: $\omega_e(\mathcal{C}_+;\mathbf{p})\le C_3\sqrt{n}$, as desired. □

D.2 Proof of Lemma 3.5

Here, we prove Lemma 3.5. We employ Proposition 3.1 to control the quadratic width. Note that $\mathcal{C}_{\text{sparse}}$ is the descent cone of the ℓ_1-norm at X_0, i.e., $\mathcal{C}_{\text{sparse}}=\mathcal{D}(\mathcal{K}_{\mathcal{R}},\mu X_0)$ for $\mathcal{R}:\mathbb{S}_n\to\mathbb{R}$ such that $\mathcal{R}(X):=\|X\|_1$.

Proof. Start by recalling the following characterization of the subgradient set of $\|X\|_1$:

$$S\in\partial\mathcal{R}(X)\subset\mathbb{R}^{n\times n}$$

iff, for $i, j \in [n]$,

$$S_{i,j} = \begin{cases} 1 & \text{if } X_{i,j} > 0, \\ -1 & \text{if } X_{i,j} < 0, \\ \varsigma_{i,j} \in [-1,1] & \text{if } X_{i,j} = 0. \end{cases} \tag{74}$$

Now using the rotational invariance of the Gaussian distribution, we may assume without loss of generality that $\mathbf{x}_0 = (x_1^0, \dots, x_n^0)$ be such that $\mathbf{x}_{0,i} > 0$ for $i = 1, \dots, k$ and $\mathbf{x}_{0,i} = 0$ for $i = k+1, \dots, n$. This further implies that all the entries in the k-th order sub-matrix of X_0 are positive and the rest of X_0 contains zero entries. Therefore, $S \in \partial\mathcal{R}(X_0)$, iff

$$S_{i,j} = \begin{cases} 1 & \text{if } i, j \in [k], \\ \varsigma_{i,j} \in [-1,1] & \text{otherwise.} \end{cases} \tag{75}$$

By employing Proposition 3.1, we get

$$\begin{aligned} \omega_e^2(\mathcal{C}_{\text{sparse}}; \mathbf{p}) &\leq \mathbb{E}\big[\inf_{S \in \partial\mathcal{R}(X_0)} \|H - \lambda \cdot S\|_F^2 \big] \\ &\overset{(i)}{=} \mathbb{E}\Big[\sum_{(i,j) \in [k]^2} (h_{ij} - 1)^2 + \sum_{(i,j) \notin [k]^2} \text{st}(h_{ij}; \lambda)^2 \Big], \end{aligned} \tag{76}$$

where (i) follows from (75). Note that $\text{st}(\cdot; \lambda)$ denotes the soft thresholding function, which is defined as

$$\text{st}(h; \lambda) = \begin{cases} \frac{h}{|h|} \cdot (|h| - \lambda) & \text{if } |h| \geq \lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{77}$$

Next, we separately bound the two terms appearing in (76). On the one hand,

$$\begin{aligned} \mathbb{E}\Big[\sum_{(i,j) \in [k]^2} (h_{i,j} - 1)^2 \Big] &= \mathbb{E}\Big[\sum_{i=1}^{k} (h_{ii} - 1)^2 \Big] + \mathbb{E}\Big[\sum_{(i \neq j) \in [k]^2} (h_{ij} - 1)^2 \Big] \\ &= k\Big(\lambda^2 + \mathbb{E}[h_{11}^2]\Big) + k(k-1)\Big(\lambda^2 + \mathbb{E}[h_{12}^2]\Big) \\ &\overset{(i)}{=} k\Big(\lambda^2 + 3\frac{\|\mathbf{p}\|_2^2}{m}\Big) + k(k-1)\Big(\lambda^2 + \frac{\|\mathbf{p}\|_2^2}{m}\Big), \end{aligned} \tag{78}$$

where (i) follows from the following observations:

$$\mathbb{E}[h_{11}^2] = \frac{1}{m} \cdot \mathbb{E}\big[\big(\sum_{i \in [m]} p_i \varepsilon_i a_{i,1}^2 \big)^2\big] = \frac{1}{m} \cdot \sum_{i \in [m]} p_i^2 \mathbb{E}[a_{i,1}^4] = 3\frac{\|\mathbf{p}\|_2^2}{m} \tag{79}$$

and

$$\mathbb{E}[h_{12}^2] = \frac{1}{m} \cdot \mathbb{E}\big[\big(\sum_{i \in [m]} p_i \varepsilon_i a_{i,1} a_{i,2} \big)^2\big] = \frac{1}{m} \cdot \sum_{i \in [m]} p_i^2 \mathbb{E}[a_{i,1}^2 a_{i,2}^2] = \frac{\|\mathbf{p}\|_2^2}{m}. \tag{80}$$

On the other hand, the second term in (76) gives

$$\mathbb{E}\Big[\sum_{(i,j) \notin [k]^2} \text{st}(h_{ij}; \lambda)^2 \Big] = (n-k)\mathbb{E}\big[\text{st}(h_{k+1,k+1}; \lambda)^2\big] + (n-k)(n+k-1)\mathbb{E}\big[\text{st}(h_{1,k+1}; \lambda)^2\big]. \tag{81}$$

Let's consider a function $g : \mathbb{R}^+ \to \mathbb{R}$ such that

$$g(x) = \begin{cases} (|x| - \lambda)^2 & \text{if } |x| \geq \lambda, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\mathrm{st}(h_{ij};\lambda)^2 = g(|h_{ij}|)$. Moreover, using integration by parts, the following identity holds for a non-negative random variable U:

$$\mathbb{E}[g(U)] = g(0) + \int_0^\infty g'(t)\mathbb{P}[U > t]\mathrm{d}t.$$

From these, we get that

$$\begin{aligned}
&\mathbb{E}\big[\mathrm{st}(h_{1,k+1};\lambda)^2\big] \\
&= \int_\lambda^\infty 2(t-\lambda)\mathbb{P}[|h_{1,k+1}| > t]\mathrm{d}t \\
&\overset{(i)}{\leq} \int_\lambda^\infty 4(t-\lambda)e^{\max\left\{-\frac{mt^2}{\psi_1^2\|\mathbf{p}\|_2^2}, -\frac{\sqrt{m}t}{\psi_1\|\mathbf{p}\|_\infty}\right\}}\mathrm{d}t \\
&= 4\max\Bigg\{\int_\lambda^\infty (t-\lambda)e^{-\frac{mt^2}{\psi_1^2\|\mathbf{p}\|_2^2}}\mathrm{d}t, \\
&\qquad\qquad \int_\lambda^\infty (t-\lambda)e^{-\frac{\sqrt{m}t}{\psi_1\|\mathbf{p}\|_\infty}}\mathrm{d}t\Bigg\} \\
&\leq \max\Bigg\{\frac{\|\mathbf{p}\|_2^2\psi_1^2}{m}\left(\frac{\lambda\sqrt{2}\sqrt{m}}{\|\mathbf{p}\|_2\psi_1} - 1\right)e^{-\frac{m\lambda^2}{\psi_1^2\|\mathbf{p}\|_2^2}}, \\
&\qquad\qquad 4\frac{\|\mathbf{p}\|_\infty^2\psi_1^2}{m}e^{-\frac{\lambda\sqrt{m}}{\psi_1\|\mathbf{p}\|_\infty}}\Bigg\},
\end{aligned} \tag{82}$$

where in the last line we performed integration by parts and further used the known bound $Q(x) \leq \frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-x^2/2}$ on the Gaussian Q-function. The inequality in (i) follows form Bernstein's inequality (e.g., [31, Thm. 2.8.2]) and ψ_1 is an absolute constant that denotes the sub-exponential norm of $\varepsilon_i a_{i,1}a_{i,k+1}$ [31, Sec. 2.7]. Similarly, using the fact that $\varepsilon_i a_{i,k+1}^2$ is also a sub-exponential random variable, say with sub-exponential parameters ψ_2, we get that

$$\begin{aligned}
\mathbb{E}\big[\mathrm{st}(h_{k+1,k+1};\lambda)^2\big] &= \int_\lambda^\infty 2(t-\lambda)\mathbb{P}[h_{k+1,k+1} > t]\mathrm{d}t \\
&= \max\left\{\frac{\|\mathbf{p}\|_2^2\psi_2^2}{m}\left(\frac{\lambda\sqrt{2}\sqrt{m}}{\|\mathbf{p}\|_2\psi_2} - 1\right)e^{-\frac{m\lambda^2}{\psi_2^2\|\mathbf{p}\|_2^2}}, \; 4\frac{\|\mathbf{p}\|_\infty^2\psi_2^2}{m}e^{-\frac{\lambda\sqrt{m}}{\psi_2\|\mathbf{p}\|_\infty}}\right\}
\end{aligned} \tag{83}$$

Call $\psi := \max\{\psi_1, \psi_2\}$ and set

$$\lambda = \frac{\psi}{\sqrt{m}}\left(\|\mathbf{p}\|_2\sqrt{\log\left(\frac{n^2}{k^2}\right)} + \|\mathbf{p}\|_\infty\log\left(\frac{n^2}{k^2}\right)\right)$$

Combining (78), (81), (82), (83) with (76), for this choise of λ we obtain that

$$\begin{aligned}
\omega_e^2(\mathcal{C}_{\mathrm{sparse}};\mathbf{p}) &\leq 3k^2\left(\lambda^2 + \frac{\|\mathbf{p}\|_2^2}{m}\right) + (n^2-k^2)\frac{k^2}{n^2}4\lambda^2 \\
&\leq \frac{C}{m}k^2\left(\|\mathbf{p}\|_2\sqrt{\log\left(\frac{n^2}{k^2}\right)} + \|\mathbf{p}\|_\infty\log\left(\frac{n^2}{k^2}\right)\right)^2,
\end{aligned}$$

for some sufficiently large absolute constant $C > 0$.

□

E Spectral norm of weighted sum of Gaussian outer products

Lemma E.1 delivers an upper bound on the spectral norm of a weighted sum of outer products of Gaussians $\sum_{i=1}^m p_i\varepsilon_i\mathbf{a}_i\mathbf{a}_i^T$. The proof uses an ϵ-net argument and Hanson-Wright inequality for Gaussians (Lemma E.2).

Lemma E.1. *For $\mathbf{a}_1, \dots, \mathbf{a}_m \in \mathbb{R}^n$ independent copies of a standard normal vector $\mathcal{N}(0, \mathbf{I}_n)$. $\varepsilon_1, \dots, \varepsilon_m$ iid Rademacher random variables and a deterministic vector $\mathbf{p} := (p_1, \dots, p_m)$ let*

$$\tilde{H} = \sum_{i=1}^{m} p_i \cdot \varepsilon_i \mathbf{a}_i \mathbf{a}_i^T$$

Then, there exists constant $C > 0$ and sufficiently large n such that

$$\mathbb{E}\|\tilde{H}\|_2 \leq C \cdot (\|\mathbf{p}\|_2 \sqrt{n} + \|\mathbf{p}\|_\infty n).$$

Proof. (a). Fix $\mathbf{v} \in \mathcal{S}^{n-1}$ and consider

$$S := \mathbf{v}^T \tilde{H} \mathbf{v}.$$

Rewriting $S = \langle \tilde{H}, V \rangle$ for $V = \mathbf{v}\mathbf{v}^T$, we may apply Hanson-Wright inequality for Gaussians (see Lemma E.2) to find that for any $t > 0$:

$$\mathbb{P}\left(S \geq t \right) \leq \exp\left(- \min\left\{ \frac{t^2}{4\|\mathbf{p}\|_2^2} \, , \, \frac{t}{2\|\mathbf{p}\|_\infty} \right\} \right), \tag{84}$$

where we also used $\|V\|_F = \|V\|_2 = 1$.

Next, let $\mathcal{N}$ be an $1/4$-net of the sphere. By standard calculations (e.g., [31, Chapter 4] $|\mathcal{N}| \leq 9^n$ and

$$\|\tilde{H}\|_2 \leq 2 \cdot \max_{\mathbf{v} \in \mathcal{N}} \mathbf{v}^T \tilde{H} \mathbf{v}. \tag{85}$$

Also, $\mathbb{E}[\|\tilde{H}\|_2] = \int_0^\infty \mathbb{P}[\|\tilde{H}\|_2 \geq t] \mathrm{d}t$. Combine all these and choose $\delta = C_1 \cdot (\|\mathbf{p}\|_2 \sqrt{n} + \|\mathbf{p}\|_\infty n)$ to find the desired result as follows:

$$\begin{aligned}
\mathbb{E}[\|\tilde{H}\|_2] &\leq \delta + \int_\delta^\infty \mathbb{P}[\|\tilde{H}\|_2 \geq t] \mathrm{d}t \\
&\leq \delta + 2 \cdot \int_\delta^\infty \mathbb{P}\big[\max_{\mathbf{v} \in \mathcal{N}} \mathbf{v}^T \tilde{H} \mathbf{v} \geq t/2 \big] \mathrm{d}t \\
&\leq \delta + 2 \cdot 9^n \cdot \max\left\{ \int_\delta^\infty e^{-\frac{t^2}{16\|\mathbf{p}\|_2^2}} \mathrm{d}t \, , \, \int_\delta^\infty e^{-\frac{t}{4\|\mathbf{p}\|_\infty}} \mathrm{d}t \right\} \\
&\leq \delta + \max\left\{ 8\sqrt{\pi}\|\mathbf{p}\|_2 e^{-\frac{\delta^2}{16\|\mathbf{p}\|_2^2}} \, , \, 8\|\mathbf{p}\|_\infty e^{-\frac{\delta}{4\|\mathbf{p}\|_\infty}} \mathrm{d}t \right\} \\
&\leq C_2 \cdot (\|\mathbf{p}\|_2 \sqrt{n} + \|\mathbf{p}\|_\infty n),
\end{aligned} \tag{86}$$

where the last inequality holds for sufficiently large n.

□

Lemma E.2. *Let $\tilde{H}$ be defined as in Lemma E.1 and V be an $n \times n$ matrix. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\Big(\, |\langle \tilde{H}, V \rangle| \geq t \, \Big) \leq 2 \exp^{-\min\left\{ \frac{t^2}{4\|\mathbf{p}\|_2^2 \|V\|_F^2} \, , \, \frac{t}{2\|\mathbf{p}\|_\infty \|V\|_2} \right\}}.$$

Proof. Let

$$\mathbf{b} := [\mathbf{a}_1^T, \dots, \mathbf{a}_m^T] \in \mathbb{R}^{mn}$$

and

$$M := \mathrm{BlockDiag}(p_1 \varepsilon_1 V, \dots, p_m \varepsilon_m V) \in \mathbb{R}^{mn \times mn}.$$

With these, note that $\mathbf{b}^T M \mathbf{b} = \sum_{i \in [m]} p_i \varepsilon_i \, \mathrm{tr}(\mathrm{V} \mathbf{a}_\mathrm{i} \mathbf{a}_\mathrm{i}^\mathrm{T})$ and $\mathbb{E}[\mathbf{b}^T M \mathbf{b}] = \sum_{i \in [m]} p_i \varepsilon_i \, \mathrm{tr}(\mathrm{V})$. Thus,

$$\tilde{H} = \mathbf{b}^T M \mathbf{b} - \mathbb{E}[\mathbf{b}^T M \mathbf{b}].$$

Then, the lemma follows immediately by Hanson-Wright inequality for Gaussian random variables (e.g., see [13, Prop. 1.1]) and the following observations:

$$\|M\|_F^2 = \|\mathbf{p}\|_2^2 \|V\|_F^2 \;\text{ and }\; \|M\|_2 = \|\mathbf{p}\|_\infty \|V\|_2. \tag{87}$$

□