The PhaseLift for Non-quadratic Gaussian Measurements

Christos Thrampoulidis[1] and Ankit Singh Rawat[1,2]

[1]Research Laboratory of Electronics, MIT, Cambridge, MA 02139, USA.
[2]College of Information and Computer Science, University of Massachusetts, Amherst, MA 01003, USA.
E-mail: {cthrampo, asrawat}@mit.edu.

December 10, 2017

Abstract

We study the problem of recovering a structured signal $\mathbf{x}_0$ from high-dimensional measurements of the form $y = f(\mathbf{a}^T\mathbf{x}_0)$ for some nonlinear function f. When the measurement vector $\mathbf{a}$ is iid Gaussian, Brillinger observed in his 1982 paper that $\mu_\ell \cdot \mathbf{x}_0 = \min_{\mathbf{x}} \mathbb{E}(y - \langle \mathbf{a}, \mathbf{x}\rangle)^2$, where $\mu_\ell = \mathbb{E}_\gamma[\gamma f(\gamma)]$ with γ being a standard Gaussian random variable. Based on this simple observation, he showed that, in the classical statistical setting, the least-squares method is consistent. More recently, Plan & Vershynin extended this result to the high-dimensional setting and derived 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 appropriate generic semidefinite-optimization based method. In a nutshell, our idea is to treat such link functions as if they were linear in a lifted space of higher-dimension. An appealing feature of our error analysis is that it captures the effect of the nonlinearity in a few simple summary parameters, which can be particularly useful in system design.

1 Introduction

1.1 Motivation

We consider the problem of estimating an unknown signal vector $\mathbf{x}_0 \in \mathbb{R}^n$ from a vector $\mathbf{y} = (y_1, y_2, \ldots, y_m)^T$ of m *generalized linear measurements*[1] taking the following form:

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

Here, each $\mathbf{a}_i \in \mathbb{R}^m$ represents a (known) measurement vector, the f_i's are independent copies of a (possibly random) link function f. The model (1) includes many commonly encountered instances of signal recovery problems as special cases. A few examples include: (a) $f_i(x) = x + z_i$, with say z_i being normally distributed, for standard linear regression setup with gaussian noise; (b) $f(x) = \text{sign}(x)$ (or, $\text{sign}(x + z_i)$), for 1-bit quantized (noisy) measurements; (c) $f(x) = |x|^2 + z_i$, for quadratic (noisy) measurements, and so on. Recovery problems from quadratic measurements arise in a variety of problems in optics and are referred to as phase-retrieval (PR) problems. In fact, in various imaging applications, a physical model that more accurately describes the variation in the number of photons detected by the optical sensor is as follows: $y_i \sim \text{Poisson}(|\mathbf{a}_i^T\mathbf{x}_0|^2)$, $i = 1, 2, \ldots, m,$. To allow for statistical models of the measurements of this latter type, we also consider a slight generalization of (1) where the measurements

[1]In the statistics and econometrics literature, the measurement model in (1) is popular under the name *single-index model*; it can also be regarded as a special case of what is known as *sufficient dimension reduction* problem.

are drawn independently according to a conditional distribution of a (known) probability density function as follows:

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

Our focus is on the high-dimensional regime, where both the number of measurements m and the dimension of the unknown signal n are large. Also, we aim to exploit a-priori available structural information about the unknown signal $\mathbf{x}_0$. To exploit this information it is typical to associate with the structure of $\mathbf{x}_0$ a properly chosen function $\mathcal{R} : \mathbb{R}^n \to \mathbb{R}$, which we refer to as the *regularizer*. Of particular interest are *convex* and non-smooth regularizers, such as the ℓ_1-norm for sparse signals, the $\ell_{1,2}$ for group-sparse ones, and so on.

1.2 Prior art

In the simplest case where the link function is *linear*, i.e., $f_i(x) = x + z_i$, perhaps the most popular way 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, $\mathcal{K}_{\mathcal{R}} \subset \mathbb{R}^n$ is a convex set that encodes the available information about $\mathbf{x}_0$. For instance,

$$\mathcal{K}_{\mathcal{R}} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathcal{R}(\mathbf{x}) \leq K\},$$

for some (tuning) parameter $K > 0$. The generalized Lasso enjoys much success in practice and comes with provable performance guarantees under general assumptions on the choice of the regularizer function $\mathcal{R}$ 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 even in the case of *non*-linear observations $y_i = f(\mathbf{a}_i^T \mathbf{x}_0)$ with f being a non-linear function. Naturally, this approach gives rise to the following question:

When (if ever) is the solution $\hat{\mathbf{x}}$ of the Lasso still a good estimate of $\mathbf{x}_0$?

This question has been recently addressed in a quantitative way by Plan & Vershynin, under the generic assumption that the measurement vectors are independent Gaussians [1]. Naturally, the answer depends both on: (i) the specific non-linearity f in (1); (ii) on the structure of $\mathbf{x}_0$ and the choice of the regularizer. It turns out that the dependence on these factors can be summarized in terms of a few key parameters that are easy-to-compute. In particular, the effect of the nonlinearity is entirely captured by the following two parameters:

$$\mu_\ell := \mathbb{E}[\gamma f(\gamma)], \quad \text{and} \quad \tau_\ell^2 := \mathbb{E}[(f(\gamma) - \mu\gamma)^2], \qquad \text{for } \gamma \sim \mathcal{N}(0, 1), \tag{4}$$

where the expectations are over γ and the (possibly) random link function f. With these and assuming[2] $\|\mathbf{x}_0\|_2 = 1$ the error performance of the Lasso, as derived in [1, Thm. 1.4], is as follows.

For illustration, let us focus on the case of sparse recovery: suppose $\mathbf{x}_0$ is k-sparse and $\hat{\mathbf{x}}$ is the solution to (3) with ℓ_1-regularization Then, if $m \gtrsim k \log(n/k)$ and n is large enough, then with high probability[3]:

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

[2] Information about the magnitude of $\mathbf{x}_0$ might be in general lost because of the non-linearity. To keep things general, we are interested in an estimate that has high correlation with the true signal (also called weak recovery, e.g.[2]). Hence, here the assumption $\|\mathbf{x}_0\|_2 = 1$ is made without loss of generality.

[3] 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).

The fact that just μ_ℓ and τ_ℓ entirely capture the effect of the nonlinearity f can be particularly useful. This is the case in systems where certain parameters of f can be directly designed (e.g., think of quantized measurements where one chooses the levels and thresholds of quantization) [3]. But also, when f is entirely determined by nature, one can use (5) to design a *pre-processing function* h such that feeding $\mathbf{z}_i = h(\mathbf{y}_i)$ to the Lasso reduces the error. In both cases, according to (5), the optimal design is the one that minimizes the *effective noise parameter* τ_ℓ^2/μ_ℓ^2; see also [3].

What if the Lasso fails? The shortcoming of (5), or rather of the lasso estimator itself, is that it obviously fails to produce a good estimate whenever $\mu_\ell = 0$. For example, this is the case under quadratic measurements, since $\mu_\ell = \mathbb{E}[\gamma^3] = 0$. But, it also happens for all even link functions. In this paper, we provide an affirmative answer to the following natural questions:

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?

The term *generic* above refers to a method that can adapt to different nonlinearities with minimal changes, such as appropriate tuning of involved regularization parameters.

In a nutshell, our idea is as follows. Rather than pretending that f is linear (as the Lasso does), which can be a bad approximation for (say) even link functions, we first *lift the measurements* to a higher-dimensional space and then we perform the linear approximation at the new space.

1.3 Contribution

When measurements are quadratic we can apply a *lifting method* that effectively makes the measurements linear in a higher-dimensional space. Observe that,

$$\mathbf{y}_i = (\mathbf{a}_i^T\mathbf{x}_0)^2 + z_i = \mathrm{tr}(\mathbf{a}_\mathrm{i}\mathbf{a}_\mathrm{i}^\mathrm{T}\cdot\mathbf{x}_0\mathbf{x}_0^\mathrm{T}) + z_i. \tag{6}$$

Thus, the measurements are linear functions of the rank-one matrix $X_0 = \mathbf{x}_0\mathbf{x}_0^T \in \mathbb{R}^{n\times n}$:

$$\mathbf{y}_i = \mathrm{tr}(\mathbf{a}_\mathrm{i}\mathbf{a}_\mathrm{i}^\mathrm{T}\cdot\mathrm{X}_0) + \mathbf{z}_i.$$

We can now attempt to reconstruct the unknown matrix X_0 by searching for a positive-semidefinite and low-rank matrix X that minimizes the residual between $\mathbf{y}_i$ and $\mathrm{tr}(\mathbf{a}_\mathrm{i}\mathbf{a}_\mathrm{i}^\mathrm{T}\cdot\mathrm{X})$. Further replacing the rank constraint with trace constraint, we arrive at the following convex method:

$$\widehat{X} = \arg\min_{X\succeq 0} \sum_{i=1}^{m}(\mathbf{y}_i - \mathrm{tr}(\mathbf{a}_\mathrm{i}\mathbf{a}_\mathrm{i}^\mathrm{T}\cdot\mathrm{X}))^2 + \lambda\cdot\mathrm{tr}(\mathrm{X}), \tag{7}$$

where $\lambda > 0$ is a regularization parameter. Then, return the leading eigenvector of $\widehat{X}$ as a final estimate for $\mathbf{x}_0$. Such formulations of the PR problem that are based on semidefinite programming relaxation are by now well studied in the literature, e.g., [4, 5, 6, 7]. They often go by the name "PhaseLift" that was coined in [4]. PhaseLift in (7) can be naturally adapted to account for a-priori available structural information on $\mathbf{x}_0$. For instance, if $\mathbf{x}_0$ is k-sparse, then X_0 is k^2-sparse and one further constraints the feasible set in (7) such that $X \in \mathcal{K}_{\ell_1} := \{W \mid \|W\|_1 \le K\}$, where $K > 0$ is a regularization parameter and $\|W\|_1 = \sum_{i=1}^n\sum_{j=1}^n |W_{ij}|$.

PhaseLift beyond quadratics. In this paper, we go beyond the quadratic measurement model (6). Our goal is to obtain an estimate of $\mathbf{x}_0$ from (1). In order to do this, we combine two tricks that we have already seen. First, we lift to a higher-dimensional space. In contrast to (6) where the lifted measurements are linear in the new space, this is of course not true in general. Then, we borrow the idea of [1]: we just pretend that the measurements are linear and use one of the already available recovery algorithms for linear problems. We accompany the methodology with guarantees similar in nature to [1] by quantifying

the recovery performance as a function of the specific nonlinearity and of the signal structure. Our goal is to obtain estimate $\hat{\mathbf{x}}$, with high correlation value: $\frac{\hat{\mathbf{x}}^T \mathbf{x}_0}{\|\mathbf{x}_0\|_2 \|\hat{\mathbf{x}}\|_2}$. We assume onwards that $\|\mathbf{x}_0\|_2 = 1$ and that measurement vectors $\mathbf{a}_i$ are standard (real) Gaussian.

The new algorithm for nonlinear measurements is efficient for link functions f for which the following parameter is nonzero:

$$\mu_q := \frac{1}{2}\mathbb{E}[(\gamma^2 - 1)f(\gamma)], \qquad \text{for } \gamma \sim \mathcal{N}(0,1). \tag{8}$$

By simple integration by parts, note that this is equivalent to $\mu_q := \frac{1}{2}\mathbb{E}[f''(\gamma)]$ for two-times differentiable functions. For example, $\mu_q = 1 > 0$ for quadratic measurements. Applying the lifting trick as previously described, instead of directly searching for $\mathbf{x}_0$ in the natural domain, we search for the rank-1 matrix $X_0 = \mathbf{x}_0 \mathbf{x}_0^T$ in the lifted domain. For this task, we solve the following semidefinite optimization program:

$$\widehat{X} = \arg\min_{X \succeq 0} \sum_{i=1}^{m} (\mathbf{y}_i - \mathrm{tr}(\mathbf{a}_i \mathbf{a}_i^{\mathrm{T}} \cdot \mathrm{X}))^2 \quad \text{s.t.} \qquad \mathrm{tr}(\mathrm{X}) = \mu_q \quad \text{and} \quad X \in \mathcal{K}_{\mathcal{R}}. \tag{9}$$

Finally, we return the leading eigenvector of $\widehat{X}$ as an estimate of $\mathbf{x}_0$. We prove guarantees on the performance of (9) similar in nature to (5). In particular, the error bounds summarize both the problem geometry and the specific nonlinearity in (separate) summary parameters that are easy to compute. For example, for sparse recovery of a k-sparse signal we show the following about (9) with ℓ_1-regularization. If $m \gtrsim k^2 \log(n/k)$ and n sufficiently large then it holds with high probability that

$$\|\hat{X} - \mu_q X_0\|_F \lesssim \tau_q \cdot \frac{k \log(n/k)}{\sqrt{m}}, \tag{10}$$

where the parameter τ_q is defined as follows for $\gamma \sim \mathcal{N}(0,1)$:

$$\tau_q^2 := \mathbb{E}\left[\left(f(\gamma) - \mu \cdot \gamma^2\right)^2\right]. \tag{11}$$

Observe the resemblance between (5) and our result in (9). On the one hand, in terms of sample complexity, (8) pays a penalty of $\mathcal{O}\left(k^2 \log(n/k)\right)$ rather than $\mathcal{O}\left(k \log(n/k)\right)$, but this gap also appears with currently known algorithms for even just quadratic measurements (e.g., [8]). On the other hand, the Lasso-related parameters μ_ℓ and τ_ℓ in (4) are replaced by the new parameters μ_q and τ_q, respectively. Hence, just like it is the case with (5), our result suggests means of optimal design of either f or a pre-processing function h.

Remarks. We note that the program in (9) is a variant of the PhaseLift program in (7). Thus, for convenience, we henceforth refer to the method in (9) as PhaseLift. We further note that the PhaseLift (9) is not entirely agnostic to the link function f, since it requires knowledge of the parameter μ_q. However, it is generic in the sense that no other modifications are needed for different f's aside from adjusting the value of the trace constraint. In particular, the same solver can be used to obtain $\widehat{X}$ for different link functions. Similar to the Lasso in (3) the constraint $X \in \mathcal{K}_{\mathcal{R}}$ is meant to promote structural information on $\mathbf{x}_0$. Moreover, we remark the following property regarding μ_q. In the common case of additive independent noise to a deterministic link function f (i.e., $y_i = f(\mathbf{a}_i^T \mathbf{x}_0) + z_i$ with z_i independent of $\mathbf{a}_i, \mathbf{x}_0$), computing μ_q (consequently, tuning of PhaseLift (9)) does *not* require knowledge of the statistics of the noise since, $2\mu_q = \mathbb{E}[(\gamma^2 - 1)(f(\gamma) + z)] = \mathbb{E}[(\gamma^2 - 1)f(\gamma)]$.

In order to motivate the trace constraint in (7), recall the assumption that $\|\mathbf{x}_0\|_2 = 1$, which is equivalent to $\mathrm{tr}(\mathrm{X}_0) = 1$. Also, observe that for quadratic measurements, $\mu_q = 1$; thus, the trace constraint becomes $\mathrm{tr}(\mathrm{X}) = 1 = \mathrm{tr}(\mathrm{X}_0)$. Inherent in the formulation of (8) is the condition $\mu_q > 0$ (since, $X \succeq 0$ implies that $tr(X) \geq 0$). The method returns a trivial non-informative estimate $X = 0$, when

$\mu_q = 0$. On the other hand, if $\mu_q < 0$, then the same method works only by replacing the constraint $X \succeq 0$ with $X \preceq 0$.

It is easy to see that for odd link functions $\mu_q = 0$, in which case the PhaseLift algorithm in (9) fails. Of course, for such functions one may use the Lasso. On the other hand, the Lasso fails for even functions and the PhaseLift can be used instead. For other nonlinearities, such as ReLU(x) both the Lasso and PhaseLift are applicable. Of course, the latter is computationally heavier. However, it remains an interesting question which one is statistically favorable, or, if some natural combination of the two methods can achieve the best of both worlds.

1.4 Relevant literature and outlook

Over the last decade or so, with the advent of compressed sensing theory, there has been significant progress related to recovery of structured signals from high-dimensional measurements, both in theory and in applications. The majority of the works assume a (noisy) linear measurement model and a recovery method that solves a convex program, which minimizes a weighted combination of a loss function and a (typically non-smooth) regularizer. The prototypical example of this approach is sparse recovery with ℓ_1-regularized least squares (aka, Lasso). Other notable examples include low-rank matrix recovery with nuclear-norm minimization, group-sparse recovery with $\ell_{1/2}$-minimization, and so on. By now, performance guarantees for these algorithms are available for wide range of assumptions on the distribution of the measurement matrix (Gaussian, sub-Gaussian, random Fourier, etc,) [9]. Specifically for iid Gaussian matrices, there is a complete and precise theory under general assumptions on the noise distribution, the signal-structure, the loss-function and the regularizer (see references in [10, Sec. 7]).

Extensions of those results to nonlinear link functions, have been only more recently considered in the high-dimensional setting. The core observation goes back to the work by Brillinger [11], who observed that for Gaussian measurement vectors it holds $\mu_\ell \mathbf{x}_0 = \min_{\mathbf{x}} \mathbb{E}(f(\langle \mathbf{a}, \mathbf{x}_0 \rangle) - \langle \mathbf{a}, \mathbf{x} \rangle)^2$. After Brillinger's result, Li and Duan [12] generalized the result to elliptically symmetric distributions. Plan and Vershynin [1] (see also [13]) extended Birllinger's result to the high-dimensional setting by considering structured signal recovery with the constrained Lasso. Thrampoulidis et al. further extend the result to the regularized Lasso in [14], where they pinpoint the exact constants in the analysis and use the results to design optimal quantization schemes. More recent works extend the results to other loss-functions beyond least squares [15], to elliptically symmetric distributions [16], to projected gradient-descent [17] and to signal-demixing applications [18]. See also [19] for use of these ideas to obtain computational speedups compared to maximum-likelihood estimation. Finally, very recently [20] have appropriately modified and extended Brillinger's original observation to sub-Gaussian vectors. Based on that, they propose and analyze generic convex solvers for recovery from nonlinear link function with sub-Gaussian design.

Unfortunately, all these works starting with the original result by Brillinger assume that the link function satisfies $\mu_\ell \neq 0$. In particular, this excludes all even functions. Instead, our method and analysis works for these; and more generally for link functions satisfying $\mu_q \neq 0$. In that sense, this paper is a direct counterpart of [1] for "even-like" nonlinearities. Specifically, when applied to the classical statistical setting of "large m, but fixed n", it becomes the counterpart to Brillinger's result [11]. It is natural to seek further correspondences with the results in [20] by extending our results to the regularized version of (9), to other loss functions, etc. Also, a particularly interesting research question that is raised is related to unifying the effectiveness of the Lasso and of (9) in a single generic algorithm that would combine the best of the two worlds and would apply for nonlinearities satisfying either $\mu_\ell \neq 0$ or $\mu_q \neq 0$.

Out of all the nonlinear link functions, the quadratic deserves special attention since it corresponds to the fundamental task of phase-retrieval. There has been an explosion in the development of methods for this case over the past few years, with a particular focus towards rigorous recovery guarantees. See [4, 5, 6] and many references therein.

Among the most well-established methods, are those based on semidefinite relaxation (*e.g.*, [21, 4]). Importantly, such solution methods were chronologically the first to enjoy rigorous recovery guarantees.

They operate by lifting the original n-dimensional natural parameter space to a higher dimensional matrix space. In this paper, we extend the lifting idea to general link functions beyond quadratics and obtain rigorous recovery guarantees for this general setting. The drawback of lifting techniques is the increase in the dimensionality, which introduces challenges in computational complexity and memory requirement for the resulting algorithms. To overcome these issues, subsequent works on the phase-retrieval problem develop nonconvex formulations of the phase retrieval problem and solution algorithms that start with a careful spectral initialization (see [22, Sec. 6] for references), which is then iteratively refined by a gradient-descent-like scheme of low computational complexity. See also [23, 24, 25] for a, more recent, alternative *convex* formulation of the problem in the original n-dimensional parameter space. We speculate that many of these solution methods can be combined with the ideas introduced in this paper to extend their reach to non-quadratic link functions. This is a possibly interesting research direction.

Finally, this paper is also very closely related to the line of work in [26, 27, 2], which studies the performance of spectral initialization for measurements of the form $\mathbf{y}_i = f(|\langle \mathbf{a}_i, \mathbf{x}_0 \rangle|)^2$. However, we note that [27, 2] only study the problem of perfect recovery and the results are asymptotic. Also, compared to [26] our error analysis is tight with respect to the link function, i.e., the error bounds accurately capture its effect in a few summary parameters. Such results can be directly useful in signal-processing applications where the engineer has control over some parameters of the nonlinearity f [14], and also to the design of pre-processing functions h (see [27, 2] for a successful application of this idea). We plan to investigate such directions in future work.

1.5 Notation and organization

We use $\mathbb{S}_n$ to denote the set of real $n \times n$ symmetric matrices. $\langle G, V \rangle = \mathrm{tr}(\mathrm{GV})$ denotes the standard inner product in $\mathbb{S}_n$. For a matrix $A \in \mathbb{R}^{m \times n}$, $\|\cdot\|_2$ and $\|\cdot\|_F$ denote its spectral norm and Frobenius norm, respectively. For a vector $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|_p$ denotes its ℓ_p norm, where $p \geq 0$. For a matrix $A \in \mathbb{R}^{m \times n}$, we use $\|A\|_0$ and $\|A\|_1$ to denote the ℓ_0 norm and ℓ_1 norm, respectively, of the mn-length vector obtained by stacking the entries of A as a vector.

2 Error bounds for nonlinear measurements

2.1 An example: Sparse recovery

Here, we assume that the true signal $\mathbf{x}_0$ is sparse; let us denote the number of its non-zero entries by k. Then, the matrix $X_0 = \mathbf{x}_0 \mathbf{x}_0^T$ is also sparse with at most k^2 non-zero entries. In order to promote its sparsity we solve (9) with an ℓ_1-norm constraint. Theorem 2.1 below provides a bound on the performance of the algorithm. Before that, recall the definitions of μ_q and τ_q in (8) and (11), respectively:

$$\mu := \frac{1}{2}\mathbb{E}[(\gamma^2 - 1) f(\gamma)], \quad \text{and} \quad \tau^2 := \mathbb{E}\left[\left(f(\gamma) - \mu \cdot \gamma^2\right)^2\right], \qquad \text{for } \gamma \sim \mathcal{N}(0,1), \tag{12}$$

where we have dropped the subscript q for convenience.

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 of the PhaseLift (9) with $\mathcal{K}_{\mathcal{R}} = \{X \ : \ \|X\|_1 \leq \mu \|X_0\|_1\}$. There exist universal constants $c, C > 0$ such that, if the number of observations obeys*

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

then, for sufficiently large n, $\widehat{X}$ satisfies

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

with high probability.

Recall that a statement is said to hold with high probability if it holds with probability at least 0.99 (say). Also, the constants $c, C > 0$ in the statement of the theorem may only depend on the probability of success. We defer the proof to Section C.1.

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. [4]). In particular, letting λ_1 and $\mathbf{v}_1$ denote the maximum eigenvalue and the principle eigenvector of $\hat{X}$ respectively, we obtain $\hat{\mathbf{x}} = \sqrt{\lambda_1}\mathbf{v}_1$. Then, (20) guarantees that

$$\|\hat{\mathbf{x}} - \sqrt{\mu}\mathbf{x}_0\|_2 \leq C \cdot \min\left\{\sqrt{\mu}, \frac{E}{\sqrt{\mu}}\right\}$$

where E denotes the expression in the right hand side of (14). The proof is same as in [4, Sec. 6] and is thus omitted.

The number of measurements needed by Theorem 2.1 is $\mathcal{O}\left(k^2 \log(n/k)\right)$. Notably, this is the same as the guarantees of [4] in the special case of quadratic measurements. Moreover, for quadratic measurements, it is a well-known issue that this is an order of magnitude larger compared to the minimum number of samples required for sparse recovery from linear measurements, which is $\mathcal{O}\left(k \log(n/k)\right)$. The same k^2-barrier appears in most of the algorithms that have been thus far proposed for sparse recovery from quadratic measurements (e.g., see [22, Sec. 6] and references therein). In that sense, the sample complexity of Theorem 2.1 is same as state of the art. Furthermore, it applies beyond quadratic measurements and clearly captures the effect of the non-linearity in terms of the parameters μ and τ.

2.2 General result

The main result of this section Theorem 2.2 characterizes the error performance of (9) for general signal structure and choice of the regularizer. Our bounds are given in terms of specific summary parameters that we define in this section. We distinguish among two sets of parameters:

- *Geometric parameters* that capture the effectiveness of the imposed geometric constraints in (9) for the purpose of promoting solutions of desired structure (positive semidefinite, low-rank, sparse, etc.)
- *Model parameters* that capture the specific link-function f of the generalized linear measurement model in (1).

2.2.1 Geometric parameters

First, we need the notion of tangent cone.

Definition 2.1 (Tangent cone)**.** The tangent cone of a set $\mathcal{K} \subset \mathbb{S}_n$ at $X \in \mathbb{S}_n$ is defined as

$$\mathcal{D}(\mathcal{K}, X) := \{\tau V : \tau \geq 0, V \in \mathcal{K} - X\}.$$

Second, we need to recall the notion of (conic) Gaussian width.

Definition 2.2 (Gaussian width)**.** Let $\mathcal{K} \subset \mathbb{S}_n$ be a cone. Then, its conic Gaussian width $\omega_g(\mathcal{C})$ is defined as

$$\omega_g(\mathcal{C}) := \mathbb{E}_G\Big[\sup_{\substack{V \in \mathcal{C} \\ \|V\|_F = 1}} \langle G, V\rangle\Big], \tag{15}$$

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

The Gaussian width plays a central role in asymptotic convex geometry: its square $\omega_g(\mathcal{C})^2$ can be formally described as as a measure of the effective dimension of $\mathcal{C}$ [28, 29]. It also appears as a key quantity in the study of random linear inverse problems. 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 ambient dimension, sparsity level). The methods and ideas have been developed in recent line of work and are based on polarity arguments [30, 31, 29]; see also Appendix E. We make use of these ideas in the proof of Theorem 2.2.

Finally, we need two more geometric parameters: Talagrand's γ_1- and γ_2-functionals [32, Defn. 1.2.5]. To streamline the presentation, we defer the formal definitions of these parameters to Appendix D. We mention here that they are generally both defined with respect to an arbitrary set $\mathcal{T}$ and arbitrary metric d. For our purposes, we choose the metric as the Frobenius norm and the spectral norm for the γ_2 and γ_1 functionals. Thus, for a set $\mathcal{T} \subset \mathbb{S}_n$ we write

$$\gamma_1(\mathcal{T}, \|\cdot\|_2) \qquad \text{and} \qquad \gamma_2(\mathcal{T}, \|\cdot\|_F).$$

The γ-functionals are fundamental in the study of suprema of random processes and specifically in the theory of generic chaining [32]. In general, explicit calculation of the γ-functionals can be challenging depending on the specific set $\mathcal{T}$. However, it is often possible to control them in a sufficient (for our purposes) way. Below, we briefly discuss two such approaches.

Remark 1 (Metric entropy). A simple and popular approach to obtain approximations is 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

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

Please refer to [32, Sec. 1.2] for a proof for the case $\alpha = 2$; the case $\alpha = 1$ follows along the same lines. While this approach is not tight in the general case ([32]; also see [33] for recent refinements), it often results in satisfactory estimates. In particular it has been used successfully in compressed sensing applications [9].

Remark 2 (Gaussian width). 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 [32, Thm. 2.1.1]:

$$\gamma_2(\mathcal{T}; \|\cdot\|_F) \leq C \cdot \omega_g(\mathcal{T})$$

For example, this is particularly useful when $\mathcal{T}$ can be expressed as the intersection of a cone with the sphere (as is the case in Theorem 3.1). This is useful, since as discussed previously, the Gaussian width can be often well approximated.

2.2.2 Model parameters

First, recall the definition of μ in (8)[4]:

$$\mu := \frac{1}{2}\mathbb{E}[(\gamma^2 - 1)f(\gamma)], \qquad \text{for } \gamma \sim \mathcal{N}(0, 1). \tag{16}$$

When f is two times differentiable and f'' denotes its second derivative, then by integration by parts it holds that

$$\mu = \frac{1}{2}\mathbb{E}[f''(\gamma)].$$

[4]Note that we have dropped the subscript q for the sake of simplicity.

We further need to define the following parameters. For $\gamma \sim \mathcal{N}(0,1)$ let,

$$\begin{aligned}\tau^2 &:= \mathbb{E}\left[\left(f(\gamma) - \mu\cdot\gamma^2\right)^2\right],\\ \upsilon^2 &:= \mathbb{E}\left[\gamma^2\cdot\left(f(\gamma) - \mu\cdot\gamma^2\right)^2\right],\\ \chi^2 &:= \mathbb{E}\left[(\gamma^2-1)^2\cdot\left(f(\gamma) - \mu\cdot\gamma^2\right)^2\right].\end{aligned} \tag{17}$$

All expectations are taken over γ and the possibly random link function f.

Remark 3. The results extend to the more general model of (2) with a natural modification in the definitions (8) and (17). In particular, $f(\gamma)$ is substituted by a (random variable) $y \sim p(y|\gamma)$, and the expectation is over γ and the conditional distribution $p(y|\gamma)$. We calculate the explicit values of these model parameters for some standard link functions in Appendix A.

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 generalized linear model of* (1)*. Recall the definitions of* μ *and of* τ,υ,χ *in* (16) *and in* (17)*, respectively. Assume that* $\mu > 0$ *and that* $\mu X_0 \in \mathcal{K}_{\mathcal{R}}$*, where* $X_0 = \mathbf{x}_0\mathbf{x}_0^T$*. Let* $\mathcal{S}$ *denote the unit sphere in* $\mathbb{S}_n$ *and call*

$$\Gamma := \min\left\{\ \sqrt{n}\ ,\ \gamma_2\Big(\mathcal{D}(\mathcal{K}_{\mathcal{R}},\mu X_0)\cap\mathcal{S},\|\cdot\|_F\Big) + \gamma_1\Big(\mathcal{D}(\mathcal{K}_{\mathcal{R}},\mu X_0)\cap\mathcal{S},\|\cdot\|_2\Big)\ \right\}. \tag{18}$$

There exist universal constants $c, C > 0$ *such that, if the number of observations obeys*

$$m \geq c\cdot\Gamma^2, \tag{19}$$

then, the solution $\widehat{X}$ *of the PhaseLift* (9) *satisfies*

$$\|\widehat{X} - \mu X_0\|_F \leq C\cdot\frac{\tau\cdot\Gamma + \upsilon\cdot\min\left\{\sqrt{n},\omega_g\big(\mathcal{D}(\mathcal{K}_{\mathcal{R}},\mu X_0)\big)\right\} + \chi}{\sqrt{m}}, \tag{20}$$

with high probability.

As in Theorem 2.2, the constants may only depend on the probability of success. Also, notice that Theorem 2.2 holds for all values of n. Typically, for large enough n, the first term in the right hand side (RHS) of (20), i.e. $\tau\cdot\frac{\Gamma}{\sqrt{m}}$, becomes the dominant term.

For a simple illustration of the theorem, consider the case of a generic true signal $\mathbf{x}_0$ in the sense that no prior information is available. In this case, we solve the PhaseLift (9) with no additional constraints other than $X \succeq 0$ and $\mathrm{tr}(\mathrm{X}) = \mu$. Hence, $\Gamma \leq \sqrt{n}$ and from (39) the sample requirement is $m \geq c'\cdot n$.

3 Technical results and proofs

3.1 Preliminaries

We begin this section by introducing the necessary notation. Let $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ be the vector of m observations: $y_i = f_i(\mathbf{a}_i^T\mathbf{x}_0)\ \ \forall\ i \in [m]$. Additionally, we use X_0 and $\hat{X}$ to denote $\mathbf{x}_0\mathbf{x}_0^T$ and the solution to the PhaseLift program in (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:

$$\mathcal{A}(X) := \left(\mathbf{a}_1^T X \mathbf{a}_1 \ , \ \mathbf{a}_2^T X \mathbf{a}_2 \ , \ \ldots \ , \ \mathbf{a}_m^T X \mathbf{a}_m\right)^T .$$

Recall that $\|\mathbf{x}_0\| = 1 \Leftrightarrow \mathrm{tr}(\mathrm{X}_0) = 1$ and also $\mu > 0$, so the matrix μX_0 is feasible in (9). We define the "error matrix" as

$$\widehat{V} = \widehat{X} - \mu X_0.$$

Note that, in order to establish Theorem 2.2, we need to upper bound $\|\widehat{V}\|_F$. Towards this direction, let us consider the excess loss function which is defined as follows.

$$\begin{aligned}\mathcal{L}(\mu X_0) - \mathcal{L}(\widehat{X}) &= \|\mathbf{y} - \mathcal{A}(\mu X_0)\|_2^2 - \|\mathbf{y} - \mathcal{A}(\mu X_0 + \widehat{V})\|_2^2 \\ &= -\frac{\|\mathcal{A}(\widehat{V})\|_2^2}{2} + \langle \mathbf{y} - \mathcal{A}(X_0), \mathcal{A}(\widehat{V})\rangle. \end{aligned} \tag{21}$$

3.2 The expected excess loss

Before we present a detailed proof of Theorem 2.2, it is instructive to see why the PhaseLift (9) encourages small $\|\widehat{V}\|_F$ by establishing the following result about the expected excess loss function.

Lemma 3.1 (Expected excess loss of PhaseLift)**.** *Let X be feasible in (9). In particular, X satisfies* $\mathrm{tr}(\mathrm{X}) = \mu$. *Then,*

$$\mathbb{E}[\mathcal{L}(\mu X_0) - \mathcal{L}(X)] = \frac{m}{2}\|X - \mu X_0\|_F^2. \tag{22}$$

The lemma implies that μX_0 minimizes the expected loss of (9) among all feasible solutions.

In the rest of this section, we prove lemma 3.1. We compute the expectation of the two terms in (21). On the one hand, for any $V \in \mathbb{S}_n$:

$$\begin{aligned}\frac{1}{2}\mathbb{E}\|\mathcal{A}(V)\|_2^2 &= \frac{1}{2}\sum_{i=1}^m \mathbb{E}\left(\mathbf{a}_i^T V \mathbf{a}_i\right)^2 \\ &= \frac{m}{2}\cdot\Big(3\cdot\sum_{i\in[n]} V_{i,i}^2 + 2\cdot\sum_{i,j\in[n]\ :\ i\neq j} V_{i,j}^2 + \sum_{i,j\in[n]\ :\ i\neq j} V_{i,i}V_{j,j}\Big) \\ &= m\cdot\|V\|_F^2 + \frac{m}{2}\cdot\left(\mathrm{tr}(V)\right)^2. \end{aligned} \tag{23}$$

On the other hand, we will show later that

$$\mathbb{E}\langle \mathbf{y} - \mu\cdot\mathcal{A}(X_0), \mathcal{A}(V)\rangle = m\cdot(\mathbb{E}[f(\gamma)] - \mu)\cdot\mathrm{tr}(\mathrm{V})\,. \tag{24}$$

Before that observe that (23) and (24), yields for any $V \in \mathbb{S}_n$:

$$\mathbb{E}[\mathcal{L}(\mu X_0) - \mathcal{L}(X)] = -m\cdot\|V\|_F^2 - \frac{m}{2}\cdot\left(\mathrm{tr}(V)\right)^2 + m\cdot(\mathbb{E}[f(\gamma)] - \mu)\cdot\mathrm{tr}(\mathrm{V})\,.$$

Specifically, if $V \in \mathbb{S}_n$ further satisfies $\mathrm{tr}(\mathrm{V}) = 0$ then (22) is true as desired.

In the remaining, we show (24). To begin with, for any $i \in [m]$, we can rewrite $\mathbf{a}_i$ as

$$\mathbf{a}_i = \mathbf{x}_0\mathbf{x}_0^T\mathbf{a}_i + \left(\mathbf{I} - \mathbf{x}_0\mathbf{x}_0^T\right)\mathbf{a}_i =: P\mathbf{a}_i + P^\perp\mathbf{a}_i, \tag{25}$$

where P denotes the projection operator in the direction of $\mathbf{x}_0$, and $P^\perp$ denotes projection to the space orthogonal to the subspace spanned by x_0. With this representation, and defining

$$\gamma_i := \mathbf{a}_i^T\mathbf{x}_0,$$

for convenience, we have that

$$\begin{aligned}
\langle \mathbf{y} - \mu \cdot \mathcal{A}(X_0), \mathcal{A}(V) \rangle &= \sum_{i\in[m]} (\mathbf{a}_i^T V \mathbf{a}_i) \cdot \big(f(\mathbf{a}_i^T x_0) - \mu \cdot (\mathbf{a}_i^T \mathbf{x}_0)^2\big) \\
&= \sum_{i\in[m]} \big((P\mathbf{a}_i)^T V (P\mathbf{a}_i) + (P^\perp \mathbf{a}_i)^T V (P^\perp \mathbf{a}_i) + 2 \cdot (P\mathbf{a}_i)^T V (P^\perp \mathbf{a}_i)\big) \cdot \big(f(\mathbf{a}_i^T x_0) - \mu \cdot (\mathbf{a}_i^T \mathbf{x}_0)^2\big) \\
&= \sum_{i\in[m]} \big(\gamma_i^2 \cdot \mathbf{x}_0^T V \mathbf{x}_0 + (P^\perp \mathbf{a}_i)^T V (P^\perp \mathbf{a}_i) + \gamma_i \cdot \mathbf{x}_0^T V (P^\perp \mathbf{a}_i) + \gamma_i \cdot (P^\perp \mathbf{a}_i)^T V \mathbf{x}_0\big) \cdot \big(f(\gamma_i) - \mu \cdot \gamma_i^2\big), \\
&= (\text{Term I}) + (\text{Term II}) + (\text{Term III}) \qquad (26)
\end{aligned}$$

where,

$$\text{Term I} := \sum_{i\in[m]} \big(\mathbf{x}_0^T V \mathbf{x}_0\big) \cdot (\gamma_i^2 - 1) \cdot \big(f(\gamma_i) - \mu \cdot \gamma_i^2\big), \qquad (27a)$$

$$\text{Term II} := \sum_{i\in[m]} \big((P^\perp \mathbf{a}_i)^T V (P^\perp \mathbf{a}_i) + \mathbf{x}_0^T V \mathbf{x}_0\big) \cdot \big(f(\gamma_i) - \mu \cdot \gamma_i^2\big), \qquad (27b)$$

$$\text{Term III} := \sum_{i\in[m]} \big(\mathbf{x}_0^T V (P^\perp \mathbf{a}_i) + (P^\perp \mathbf{a}_i)^T V \mathbf{x}_0\big) \cdot \gamma_i \cdot \big(f(\gamma_i) - \mu \cdot \gamma_i^2\big). \qquad (27c)$$

Thus, it remains to compute the expectation of the three terms above.

1. $\mathbb{E}[\text{Term I}]$ **:** For convenience, denote

$$\xi_i := (\gamma_i^2 - 1)(f(\gamma_i) - \mu \cdot \gamma_i^2). \qquad (28)$$

By definition of μ in (8) observe that

$$\mathbb{E}[\xi_i] = 0 \;\forall\; i \in [m]. \qquad (29)$$

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

2. $\mathbb{E}[\text{Term II}]$ **:** 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,

$$\mathbb{E} P^\perp \mathbf{a}_i (\mathbf{a}_i^T \mathbf{x}_0) = P^\perp \mathbb{E} \mathbf{a}_i \mathbf{a}_i^T \mathbf{x}_0 = P^\perp \mathbf{x}_0 = \mathbf{0}. \qquad (30)$$

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 (27b) as follows.

$$\begin{aligned}
\mathbb{E}[\text{Term II}] &= \mathbb{E} \sum_{i\in[m]} \big((P^\perp \widetilde{\mathbf{a}}_i)^T V (P^\perp \widetilde{\mathbf{a}}_i) + \mathbf{x}_0^T V \mathbf{x}_0\big) \cdot \big(f(\gamma_i) - \mu \cdot \gamma_i^2\big) \qquad (31) \\
&\overset{(i)}{=} \sum_{i\in[m]} \big(\mathbb{E}[(P^\perp \widetilde{\mathbf{a}}_i)^T V (P^\perp \widetilde{\mathbf{a}}_i)] + \mathbf{x}_0^T V \mathbf{x}_0\big) \cdot \mathbb{E}[f(\gamma_i) - \mu \cdot \gamma_i^2] \\
&\overset{(ii)}{=} m \cdot \big(\mathbb{E}[f(\gamma)] - \mu\big) \cdot \mathrm{tr(V)}\,.
\end{aligned}$$

where, *(i)* follows by independence of $\{\mathbf{a}_i\}_{i\in[m]}$ and $\{\widetilde{\mathbf{a}}_i\}_{i\in[m]}$; and, *(ii)* follows since $\mathbb{E}[\gamma^2] = 1$ and

$$\mathbb{E}[(P^\perp \widetilde{\mathbf{a}}_i)^T V (P^\perp \widetilde{\mathbf{a}}_i)] = \mathrm{tr(P^\perp V P^\perp \mathbb{E}[\widetilde{\mathbf{a}}_i \widetilde{\mathbf{a}}_i^T])} = \mathrm{tr(P^\perp V P^\perp)} = \mathrm{tr(V)} - \mathbf{x}_0^T V \mathbf{x}_0.$$

3. $\mathbb{E}$[Term III] **:** Repeating the argument that led to (31), we can take the expectation in $\mathbb{E}$[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

$$\begin{aligned}\mathbb{E}[\text{Term III}] &= \mathbb{E}\sum_{i\in[m]} \big(\operatorname{tr}((\mathrm{P}^{\perp}\widetilde{\mathbf{a}}_{\mathrm{i}})\mathbf{x}_0^{\mathrm{T}}\mathrm{V}) + \operatorname{tr}(\mathbf{x}_0(\mathrm{P}^{\perp}\widetilde{\mathbf{a}}_{\mathrm{i}})^{\mathrm{T}}\mathrm{V})\big)\cdot\big(\gamma_i\cdot(f(\gamma_i)-\mu\cdot\gamma_i^2)\big), \\ &= \sum_{i\in[m]} \big(\operatorname{tr}((\mathrm{P}^{\perp}\mathbb{E}[\widetilde{\mathbf{a}}_{\mathrm{i}}])\mathbf{x}_0^{\mathrm{T}}\mathrm{V}) + \operatorname{tr}(\mathbf{x}_0(\mathrm{P}^{\perp}\mathbb{E}[\widetilde{\mathbf{a}}_{\mathrm{i}}])^{\mathrm{T}}\mathrm{V})\big)\cdot\mathbb{E}[\gamma_i\cdot(f(\gamma_i)-\mu\cdot\gamma_i^2)] = 0,\end{aligned} \tag{32}$$

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

3.3 Main technical result

It follows from the optimality of $\widehat{X}$ and feasibility of μX_0 for the program in (9) that

$$\mathcal{L}(\mu X_0) - \mathcal{L}(\widehat{X}) \geq 0. \tag{33}$$

By combining this observation with (21), we obtain

$$\frac{\|\mathcal{A}(\widehat{V})\|_2^2}{2} \leq \langle \mathbf{y} - \mu\cdot\mathcal{A}(X_0), \mathcal{A}(\widehat{V})\rangle. \tag{34}$$

On the one hand, recall from (9) that $\widehat{V}$ satisfies $\operatorname{tr}(V) = 0$, $\mu X_0 + V \succeq 0$ and $\mu X_0 + V \in \mathcal{K}_{\mathcal{R}}$. Thus, $\widehat{V}$ belongs to the following convex cone

$$\mathcal{C}_0 := \mathcal{C}_+ \cap \mathcal{C}_{\mathcal{R}}, \tag{35}$$

where the cones $\mathcal{C}_+$ and $\mathcal{C}_{\mathcal{R}}$ are defined as follows:

$$\mathcal{C}_+ := \{V \ : \ \mu X_0 + V \succeq 0, \text{ and } \operatorname{tr}(V) \leq 0\ \} \quad \text{and} \quad \mathcal{C}_{\mathcal{R}} := \mathcal{D}(\mathcal{K}_{\mathcal{R}}; \mu X_0). \tag{36}$$

From this, $\widehat{V}/\|\widehat{V}\|_F$ belongs to the spherical part of $\mathcal{C}_0$:

$$\mathcal{E} := \{V \ : V \in \mathcal{C}_0 \text{ and } \|V\|_F = 1\}.$$

On the other hand, observe in (34) that the LHS (resp., RHS) is homogeneous of degree 2 (resp., 1). With these, it follows from (34) that

$$\|\widehat{V}\|_F \cdot \inf_{V\in\mathcal{E}} \frac{\|\mathcal{A}(V)\|_2^2}{2} \leq \sup_{V\in\mathcal{E}} \langle \mathcal{A}(V), \mathbf{y} - \mu\cdot\mathcal{A}(X_0)\rangle. \tag{37}$$

From (37), it suffices to obtain high-probability lower and upper bounds on the quantities on the LHS and on the RHS, respectively. This will immediately lead to an upper bound on $\|\widehat{V}\|_F$.

We present the bound in Theorem 3.1 below, which is the main technical result of this paper. Combining this with the methodologies of Section 3.4 on controlling the weighted empirical width (cf. Definition 3.1) leads to Theorems 2.2 and 2.1. We defer this to Appendices C.2 and C.1.

Before the statement of the theorem, we define an appropriate generalization of the Gaussian width to our setting.

Definition 3.1 (Weighted empirical quadratic Gaussian 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 $\epsilon_1, \epsilon_2, \ldots, \epsilon_m$ be independent Rademacher random variables. For a set $\mathcal{C} \subset \mathbb{S}_n$ and a vector $\mathbf{p} := (p_1, \ldots, p_m)$ define the weighted empirical quadratic Gaussian width $\omega_e(\mathcal{C}; \mathbf{p})$ as follows:

$$\omega_e(\mathcal{C}; \mathbf{p}) := \mathbb{E}\big[\sup_{V \in \mathcal{C}} \langle V, H_{\mathbf{p}} \rangle\big], \qquad \text{where} \qquad H_{\mathbf{p}} := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} p_i \cdot \varepsilon_i \cdot \mathbf{a}_i \mathbf{a}_i^T, \tag{38}$$

and the expectation is over the randomness of $\{\mathbf{a}_i\}$. In particular, when $\mathbf{p} = \mathbf{1}$ we write

$$\omega_e(\mathcal{C}) := \omega_e(\mathcal{C}; \mathbf{1}),$$

and call this the empirical quadratic Gaussian width.

The terminology *weighted empirical quadratic Gaussian width* precisely reflects the facts that compared to the Gaussian width in (15), $H_{\mathbf{p}}$ involves a normalized sum (cf., "empirical") of weighted (by p_i's) outer products (cf., "quadratic") of Gaussian random vectors. The role of the Rademacher random variables is simply to guarantee that $H_{\mathbf{p}}$ is centered. For convenience, we onwards refer to $\omega_e(\mathcal{C}; \mathbf{p})$ and $\omega_e(\mathcal{C})$ simply as the *weighted empirical width* and *empirical width*, respectively. In Section 3.4 we discuss efficient ways to control the weighted empirical width.

Theorem 3.1 (Main technical result)**.** *Suppose that $\mathbf{a}_i \sim \mathcal{N}(0, \mathbf{I})$, and that $\mathbf{y}$ follows the generalized linear model of (1). Recall the definitions of μ and of υ, χ in (16) and in (17), respectively. Assume that $\mu > 0$ and $\mu X_0 \in \mathcal{K}_{\mathcal{R}}$, where $X_0 = \mathbf{x}_0 \mathbf{x}_0^T$. Let $\mathcal{C}_0$ be as in (35) and recall Definition 3.1. Finally, define $\boldsymbol{\eta} := (\eta_1, \ldots, \eta_m)$, where $\eta_i := \big(f(\gamma_i) - \mu \cdot \gamma_i^2\big)$, $i \in [m]$ for $\gamma_i \overset{iid}{\sim} \mathcal{N}(0, 1)$.*

There exist constants $c, C > 0$ such that, if the number of observations obeys

$$m \geq c \cdot \left(\omega_e(\mathcal{C}_0; \mathbf{1})\right)^2, \tag{39}$$

then, the solution $\widehat{X}$ of the PhaseLift (9) satisfies

$$\|\widehat{X} - \mu X_0\|_F \leq C \cdot \frac{\mathbb{E}_{\boldsymbol{\eta}}\big[\omega_e(\mathcal{C}_0; \boldsymbol{\eta})\big] + \upsilon\sqrt{2} \cdot \omega_g(\mathcal{C}_0) + \chi}{\sqrt{m}}, \tag{40}$$

with high probability.

3.3.1 Proof of Theorem 3.1

In the rest of this section, we prove Theorem 3.1.

Lower bound: First, we lower bound the LHS of (37). This can be done based on Mendelson's *Small Ball method*, a powerful strategy developed for lower bounding nonnegative empirical processes [34], which can be thought of as a generalization of Gordon's escape through a mesh lemma beyond Gaussians (cf. Section 2.2.1). We summarize the result in Lemma 3.2 below and defer details to Appendix B.

Lemma 3.2 (Lower bound)**.** *There exists positive absolute constant $c > 0$ such that for any $t > 0$ it holds*

$$\inf_{V \in \mathcal{E}} \|\mathcal{A}(V)\|_2 \geq c \cdot \sqrt{m} - 2 \cdot \omega_e(\mathcal{C}_0; \mathbf{1}) - t, \tag{41}$$

with probability at least $1 - e^{-t^2/4}$.

Upper bound: Next, we upper bound the expression on the RHS of (37). From (26),

$$\sup_{V\in\mathcal{E}}\langle \mathbf{y}-\mu\cdot\mathcal{A}(X_0),\mathcal{A}(V)\rangle \le \sup_{V\in\mathcal{E}}(\text{Term I})+\sup_{V\in\mathcal{E}}(\text{Term II})+\sup_{V\in\mathcal{E}}(\text{Term III}), \tag{42}$$

where we recall the definitions in (27). We separately upper bound each one of these three terms. This leads to the following lemma, the proof of which is deferred to Section 3.3.2.

Lemma 3.3 (Upper bound)**.** *Let υ,χ and $\boldsymbol{\eta}$ be defined as in the statement of Theorem 3.1. Then,*

$$\mathbb{E}\sup_{V\in\mathcal{E}}\langle \mathbf{y}-\mu\cdot\mathcal{A}(X_0),\mathcal{A}(V)\rangle \le \sqrt{m}\cdot\big(\chi+\sqrt{2}\cdot\upsilon\cdot\omega_g(\mathcal{C}_0)+2\cdot\mathbb{E}\big[\omega_e(\mathcal{C}_0;\boldsymbol{\eta})\big]\big) \tag{43}$$

Putting things together: In the rest of the argument, c, C refer to numerical constants, whose value may differ from instance to instance. First, Lemma 3.2 and (39) imply that there exists constant $c>0$ such that the following event holds with probability at least 0.995:

$$\inf_{V\in\mathcal{E}}\frac{1}{m}\|\mathcal{A}(V)\|_2^2\ge c.$$

Second, Lemma 3.3 combined with Markov's inequality imply that the the following event also holds with probability at least 0.995:

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

By the union bound, both events hold with probability 0.99. Conditioned on them, the desired follows immediately from (37).

3.3.2 Proof of Lemma 3.3

Here, we present the proof of Lemma 3.3. We separately upper bound each one of the three terms in the RHS of (42).

1. **Bounding** Term I **:** Recalling the definition of ξ_i in (28), note that

$$\begin{aligned}\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\xi_i\cdot\mathbf{x}_0^TV\mathbf{x}_0 \le \mathbb{E}\sup_{V\in\mathcal{E}}\|V\|_F\cdot\Big|\sum_{i\in[m]}\xi_i\Big| &\overset{(i)}{\le} \mathbb{E}\Big|\sum_{i\in[m]}\xi_i\Big| = \mathbb{E}\sqrt{\Big(\sum_{i\in[m]}\xi_i\Big)^2}\\ &\overset{(ii)}{\le}\sqrt{\mathbb{E}\Big(\sum_{i\in[m]}\xi_i\Big)^2}\\ &\overset{(iii)}{=}\sqrt{\sum_{i\in[m]}\mathbb{E}\xi_i^2}\\ &=\sqrt{m}\cdot\underbrace{\sqrt{\mathbb{E}\xi_1^2}}_{=\chi},\end{aligned} \tag{44}$$

   where: (i) follows from $\|\mathbf{x}_0\|_2=1$ and $V\in\mathcal{E}\implies\|V\|_F=1\implies\sup_{V\in\mathcal{E}}\mathbf{x}_0^TV\mathbf{x}_0$; (ii) follows from Jensen's inequality; and, (iii) by independence of the ξ_i's and (29).

2. **Bounding** Term II **:** Recall from (45) that we can rewrite Term II as follows:

$$\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\big((P^{\perp}\widetilde{\mathbf{a}}_i)^T V(P^{\perp}\widetilde{\mathbf{a}}_i)+\mathbf{x}_0^T V\mathbf{x}_0\big)\cdot\big(f(\gamma_i)-\mu\cdot\gamma_i^2\big)$$
$$=\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\big((P^{\perp}\widetilde{\mathbf{a}}_i)^T V(P^{\perp}\widetilde{\mathbf{a}}_i)+\mathbf{x}_0^T V\mathbf{x}_0\big)\cdot\eta_i, \tag{45}$$

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

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

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 (45) in the following manner.

$$\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\big((P^{\perp}\widetilde{\mathbf{a}}_i)^T V(P^{\perp}\widetilde{\mathbf{a}}_i)+\big(\mathbb{E}(\mathbf{g}_i^T\mathbf{x}_0)^2\big)\cdot\mathbf{x}_0^T V\mathbf{x}_0+\mathbb{E}(\mathbf{g}_i^T\mathbf{x}_0)\cdot\widetilde{\mathbf{a}}_i^T P^{\perp}V\mathbf{x}_0\big)\cdot\eta_i$$
$$\overset{(i)}{=}\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\mathbb{E}\left(\big((P^{\perp}\widetilde{\mathbf{a}}_i)^T V(P^{\perp}\widetilde{\mathbf{a}}_i)+(\mathbf{g}_i^T\mathbf{x}_0)^2\cdot\mathbf{x}_0^T V\mathbf{x}_0+\mathbf{g}_i^T\mathbf{x}_0\cdot\widetilde{\mathbf{a}}_i^T P^{\perp}V\mathbf{x}_0\big)\cdot\eta_i\mid\{\gamma_i,P^{\perp}\widetilde{\mathbf{a}}_i\}_{i\in[m]}\right)$$
$$\overset{(ii)}{=}\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\mathbb{E}\left(\big((P^{\perp}\widetilde{\mathbf{a}}_i)^T V(P^{\perp}\widetilde{\mathbf{a}}_i)+(\widetilde{\mathbf{a}}_i^T\mathbf{x}_0)^2\cdot\mathbf{x}_0^T V\mathbf{x}_0+\widetilde{\mathbf{a}}_i^T\mathbf{x}_0\cdot\widetilde{\mathbf{a}}_i^T P^{\perp}V\mathbf{x}_0\big)\cdot\eta_i\mid\{\gamma_i,P^{\perp}\widetilde{\mathbf{a}}_i\}_{i\in[m]}\right)$$
$$\overset{(iii)}{\leq}\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\big((P^{\perp}\widetilde{\mathbf{a}}_i)^T V(P^{\perp}\widetilde{\mathbf{a}}_i)+(\widetilde{\mathbf{a}}_i^T\mathbf{x}_0)^2\cdot\mathbf{x}_0^T V\mathbf{x}_0+\widetilde{\mathbf{a}}_i^T\mathbf{x}_0\cdot\widetilde{\mathbf{a}}_i^T P^{\perp}V\mathbf{x}_0\big)\cdot\eta_i$$
$$\overset{(iv)}{\leq}\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\big(\widetilde{\mathbf{a}}_i^T V\widetilde{\mathbf{a}}_i\big)\cdot\eta_i=\mathbb{E}\sup_{V\in\mathcal{E}}\Big\langle\sum_{i\in[m]}\eta_i\cdot\widetilde{\mathbf{a}}_i\widetilde{\mathbf{a}}_i^T,V\Big\rangle, \tag{46}$$

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. (30)), we replace $\{\mathbf{g}_i^T\mathbf{x}_0\}$ with $\{\widetilde{\mathbf{a}}_i^T\mathbf{x}_0\}$ in order to obtain (ii). The terms (iii) and (iv) follow from Jensen's inequality and the fact that

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

Next, we continue from (46) using the fact that $V\in\mathcal{E}\implies \mathrm{tr(V)}=0$ as follows:

$$\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\eta_i\langle\widetilde{\mathbf{a}}_i\widetilde{\mathbf{a}}_i^T,V\rangle=\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\eta_i\big(\langle\widetilde{\mathbf{a}}_i\widetilde{\mathbf{a}}_i^T,V\rangle-\mathrm{tr(V)}\big)$$
$$=\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\eta_i\langle\widetilde{\mathbf{a}}_i\widetilde{\mathbf{a}}_i^T-\mathbf{I},V\rangle$$
$$\leq 2\cdot\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\eta_i\langle\varepsilon_i\widetilde{\mathbf{a}}_i\widetilde{\mathbf{a}}_i^T,V\rangle=2\sqrt{m}\cdot\mathbb{E}_{\boldsymbol{\eta}}\big[\omega_e(\mathcal{C}_0;\boldsymbol{\eta})\big] \tag{47}$$

where, the last inequality follows from standard symmetrization argument, and $\varepsilon_1,\ldots,\varepsilon_m$ are independent Rademacher random variables.

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

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

As in (32), we can we can rewrite Term III as follows

$$\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\zeta_i\cdot\big(\,\mathrm{tr}((\mathrm{P}^{\perp}\widetilde{\mathbf{a}}_{\mathrm{i}})\mathbf{x}_0^{\mathrm{T}}\mathrm{V})+\mathrm{tr}(\mathbf{x}_0(\mathrm{P}^{\perp}\widetilde{\mathbf{a}}_{\mathrm{i}})^{\mathrm{T}}\mathrm{V})\,\big). \tag{48}$$

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 Term III (cf.(48)) as follows.

$$\begin{aligned}&\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\zeta_i\big(\,\mathrm{tr}(\mathrm{P}^{\perp}\Gamma_{(\mathrm{i},1)}\mathbf{x}_0\mathbf{x}_0^{\mathrm{T}}\mathrm{V})+\mathrm{tr}(\mathbf{x}_0\mathbf{x}_0^{\mathrm{T}}\Gamma_{(\mathrm{i},1)}^{\mathrm{T}}\mathrm{P}^{\perp}\mathrm{V})\,\big)\\&\quad=\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\zeta_i\cdot\big(\,\mathrm{tr}((\mathrm{P}^{\perp}\Gamma_{(\mathrm{i},1)}\mathrm{P}+(\mathrm{P}^{\perp}\Gamma_{(\mathrm{i},1)}\mathrm{P})^{\mathrm{T}}\mathrm{V})\,\big),\end{aligned} \tag{49}$$

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. (49)) as follows.

$$\begin{aligned}&\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\zeta_i\cdot\big(\langle P^{\perp}\Gamma_{(i,1)}P+P\mathbb{E}\Gamma_{(i,2)}P+P^{\perp}\mathbb{E}\Gamma_{(i,3)}P^{\perp}+P\mathbb{E}\Gamma_{(i,4)}P^{\perp},V\rangle\\&\qquad\qquad+\langle(P^{\perp}\Gamma_{(i,1)}P+P\mathbb{E}\Gamma_{(i,2)}P+P^{\perp}\mathbb{E}\Gamma_{(i,3)}P^{\perp}+P\mathbb{E}\Gamma_{(i,4)}P^{\perp})^T,V\rangle\big)\\&\overset{(i)}{\le}\mathbb{E}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\zeta_i\cdot\big(\langle P^{\perp}\Gamma_{(i,1)}P+P\Gamma_{(i,2)}P+P^{\perp}\Gamma_{(i,3)}P^{\perp}+P\Gamma_{(i,4)}P^{\perp},V\rangle\\&\qquad\qquad+\langle(P^{\perp}\Gamma_{(i,1)}P+P\Gamma_{(i,2)}P+P^{\perp}\Gamma_{(i,3)}P^{\perp}+P\Gamma_{(i,4)}P^{\perp})^T,V\rangle\big)\\&\overset{(ii)}{\le}\mathbb{E}_{\tilde{G}_i,\zeta_i}\sup_{V\in\mathcal{E}}\sum_{i\in[m]}\zeta_i\cdot\langle\tilde{G}_i+\tilde{G}_i^T,V\rangle=\sqrt{2}\cdot\mathbb{E}_{G_i,\zeta_i}\sup_{V\in\mathcal{E}}\big\langle\sum_{i\in[m]}\zeta_i\cdot G_i,V\big\rangle,\end{aligned} \tag{50}$$

where: (i) follows from Jensen's inequality; (ii) uses Lemma 3.4 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 3.4. *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,

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

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 $X = X_1 P + X_2 P^\perp$ 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,

$$(50) \leq \sqrt{2} \cdot \mathbb{E}[\big(\sum_{i\in[m]} \zeta_i^2\big)^{1/2}] \cdot \mathbb{E} \sup_{V\in\mathcal{E}} \langle G, V\rangle \leq \sqrt{2} \cdot \underbrace{\sqrt{\mathbb{E}[\zeta_1^2]}}_{=\upsilon} \cdot \sqrt{m} \cdot \underbrace{\mathbb{E} \sup_{V\in\mathcal{E}} \langle G, V\rangle}_{\omega_g(\mathcal{E})}, \tag{51}$$

where the last inequality follows from Jensen.

Combining (42) with (44), (50), and (51) we conclude with Lemma 3.3.

3.4 Controlling the weighted empirical width

The weighted quadratic Gaussian width $\omega_e(\mathcal{C}; \mathbf{p})$ depends both on the geometry of the cone $\mathcal{C}$ and on the weights p_i. A simple way to isolate the dependence of $\omega_e(\mathcal{C}; \boldsymbol{\eta})$ on the weights and on the geometry is through contraction [35, Thm. 4.4] as follows:

$$\begin{aligned}\omega_e(\mathcal{C}; \mathbf{p}) &= \mathbb{E}_{\widetilde{\mathbf{a}}_i}\left[\mathbb{E}_{\varepsilon_i} \sup_{V\in\mathcal{C}} \sum_{i\in[m]} \mathbf{p}_i \varepsilon_i \langle \widetilde{\mathbf{a}}_i \widetilde{\mathbf{a}}_i^T, V\rangle\right] \\ &\leq \mathbb{E}_{\widetilde{\mathbf{a}}_i}\left[\big(\max_{i\in[m]} \mathbf{p}_i\big) \cdot \mathbb{E}_{\varepsilon_i} \sup_{V\in\mathcal{C}} \sum_{i\in[m]} \varepsilon_i \langle \widetilde{\mathbf{a}}_i \widetilde{\mathbf{a}}_i^T, V\rangle\right] = \|\mathbf{p}\|_\infty \cdot \omega_e(\mathcal{C}).\end{aligned} \tag{52}$$

However, this may not always give the right statistical rates. We present here two alternative ways of controlling $\omega_e(\mathcal{C}; \mathbf{p})$.

3.4.1 First bound: generic chaining

The first one is very general and provides a bound on $\omega_e(\mathcal{C}; \mathbf{p})$ in terms of the following. On the one hand, the aspects of the geometry of $\mathcal{C}$ are captured by Talagrand's γ_1- and γ_2- functionals with respect to appropriate metrics. On the other hand, the role of the weights is captured by the ℓ_∞- and ℓ_2-norm of $\mathbf{p}$.

Lemma 3.5 (Generic chaining bound)**.** *For a cone $\mathcal{C} \subset \mathbb{S}_n$ and $\mathbf{p} \in \mathbb{R}^m$, recall the definition of the weighted empirical width $\omega_e(\mathcal{C}; \mathbf{p})$ in (38). There exists universal constant $C > 0$ for which:*

$$\omega_e(\mathcal{C}; \mathbf{p}) \leq C \cdot \frac{\|\mathbf{p}\|_2 \cdot \gamma_2(\mathcal{C}; \|\cdot\|_F) + \|\mathbf{p}\|_\infty \cdot \gamma_1(\mathcal{C}; \|\cdot\|_2)}{\sqrt{m}}. \tag{53}$$

In particular,

$$\omega_e(\mathcal{C}) = \omega_e(\mathcal{C}; \mathbf{1}) \leq C \cdot \left(\gamma_2(\mathcal{C}; \|\cdot\|_F) + \frac{\gamma_1(\mathcal{C}; \|\cdot\|_2)}{\sqrt{m}}\right). \tag{54}$$

We defer the proof of the lemma to Appendix D. Note from (53) that further using the crude bound $\|\mathbf{p}\|_\infty \leq \|\mathbf{p}\|_2$ leads to the following simplified expression:

$$\omega_e(\mathcal{C}; \mathbf{p}) \leq C \cdot \frac{\|\mathbf{p}\|_2}{\sqrt{m}} \cdot \left(\gamma_2(\mathcal{C}; \|\cdot\|_F) + \gamma_1(\mathcal{C}; \|\cdot\|_2)\right). \tag{55}$$

3.4.2 Second bound: polarity

Alternatively, it is possible to directly apply polarity arguments, which have been recently developed in the compressed sensing literature for the Gaussian width, but extend to more general notions such as the weighted empirical width. The idea is as follows. It can be shown using polarity that $\omega_e(\mathcal{C};\mathbf{p})$ is upper bounded by the expected distance of $H_{\mathbf{p}}$ to the polar cone $\mathcal{C}^\circ$, which is useful when a convenient description of $\mathcal{C}^\circ$ is available. For example, this is the case when $\mathcal{C}$ is the tangent cone of some convex proper function (say) $\mathcal{R}$. Then, from standard results in convexity, $\mathcal{C}^\circ$ is the cone of subdifferential of $\mathcal{R}$ [36]. This is summarized in Proposition 3.1 below.

Definition 3.2. The descent cone of a proper convex function $\mathcal{R}:\mathbb{S}_n\to\mathbb{R}$ at a point $X_0\in\mathbb{S}_n$ is defined as

$$\mathcal{D}(\mathcal{R},X_0):=\{\lambda V:\lambda\geq 0,\mathcal{R}(X_0+V)\leq\mathcal{R}(X_0)\}.$$

The following proposition is based on the principles developed in [30, 31]. Here, we follow the exposition in [37, Prop. 7.1].

Proposition 3.1 (Polarity bound)**.** *Let $\mathcal{R}:\mathbb{S}_n\to\mathbb{R}$ be a proper convex function, and fix $X_0\in\mathbb{S}_n$. 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:*

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

For specific choices of $\mathcal{R}$, it is possible to upper bound the RHS of (3.1) on a case-by-case basis. Many such examples have been worked out in the literature for the case of the Gaussian width (i.e., an iid Gaussian matrix instead of $H_{\mathbf{p}}$ in (3.1)), e.g. [29] and references therein. For illustration, in Lemmas 3.6 and 3.7 below we apply proposition 3.1 to bound the weighted empirical width of the following two cones: (a) C_+; and, (b) $C_{\mathcal{R}}$ for sparse recovery. We defer the proofs of both of these results to Appendix E.

Lemma 3.6 (Weighted empirical width of $\mathcal{C}_+$)**.** *For $\mathbf{p}:=(p_1,\ldots,p_m)$ and the cone $\mathcal{C}_+=\{V:\mu X_0+V\succeq 0,\ \text{and}\ \mathrm{tr}(V)\leq 0\}$, there exists universal constant $C>0$ such that*

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

In particular,

$$\omega_e(\mathcal{C}_+;\mathbf{1})\leq C\cdot\sqrt{n},\qquad\text{provided that } m\geq c\cdot n \text{ for some constant } c>0. \tag{56}$$

Lemma 3.7 (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 universal constant $C>0$ such that*

$$\omega_e(\mathcal{C}_{\text{sparse}};\mathbf{p})\leq\frac{C}{\sqrt{m}}\cdot k\cdot\sqrt{\log(n/k)}\left(\|\mathbf{p}\|_2+\|\mathbf{p}\|_\infty\sqrt{2\log(n/k)}\right).$$

In particular,

$$\omega_e(\mathcal{C}_{\text{sparse}},\mathbf{1})\leq C\cdot k\sqrt{\log(n/k)},\qquad\text{provided that } m\geq c\cdot k^2\log(n/k)\ \text{ for some constant } c>0.$$

References

[1] Y. Plan and R. Vershynin, "The generalized lasso with non-linear observations," *arXiv preprint arXiv:1502.04071*, 2015.

[2] M. Mondelli and A. Montanari, "Fundamental limits of weak recovery with applications to phase retrieval," *arXiv preprint arXiv:1708.05932*, 2017.

[3] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Lasso with non-linear measurements is equivalent to one with linear measurements," in *Advances in Neural Information Processing Systems*, 2015, pp. 3420–3428.

[4] E. J. Candes, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1241–1274, 2013.

[5] Y. Chen and E. J. Candes, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *Communications on Pure and Applied Mathematics*, vol. 70, pp. 0822–0883, 2017.

[6] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 4034–4059, 2015.

[7] K. Jaganathan, Y. C. Eldar, and B. Hassibi, "Stft phase retrieval: Uniqueness guarantees and recovery algorithms," *IEEE Journal of selected topics in signal processing*, vol. 10, no. 4, pp. 770–781, 2016.

[8] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2886–2908, 2015.

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

[10] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Precise error analysis of regularized M-estimators in high-dimensions," *arXiv preprint*, 2016.

[11] D. R. Brillinger, "A generalized linear model with" gaussian" regressor variables," *A Festschrift For Erich L. Lehmann*, p. 97, 1982.

[12] K.-C. Li and N. Duan, "Regression analysis under link violation," *The Annals of Statistics*, pp. 1009–1052, 1989.

[13] Y. Plan, R. Vershynin, and E. Yudovina, "High-dimensional estimation with geometric constraints," *arXiv preprint arXiv:1404.3749*, 2014.

[14] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "LASSO with non-linear measurements is equivalent to one with linear measurements," in *Advances in Neural Information Processing Systems*, 2015, pp. 3420–3428.

[15] M. Genzel, "High-dimensional estimation of structured signals from non-linear observations with general convex loss functions," *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1601–1619, 2017.

[16] L. Goldstein, S. Minsker, and X. Wei, "Structured signal recovery from non-linear and heavy-tailed measurements," *arXiv preprint arXiv:1609.01025*.

[17] S. Oymak and M. Soltanolkotabi, "Fast and reliable parameter estimation from nonlinear observations," *arXiv preprint arXiv:1610.07108*, 2016.

[18] M. Soltani and C. Hegde, "Fast algorithms for demixing sparse signals from nonlinear observations," *IEEE Transactions on Signal Processing*, 2017.

[19] M. A. Erdogdu, L. H. Dicker, and M. Bayati, "Scaled least squares estimator for glms in large-scale problems," in *Advances in Neural Information Processing Systems*, 2016, pp. 3324–3332.

[20] Z. Yang, K. Balasubramanian, and H. Liu, "High-dimensional non-gaussian single index models via thresholded score function estimation," in *International Conference on Machine Learning*, 2017, pp. 3851–3860.

[21] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin, "Painless reconstruction from magnitudes of frame coefficients," *Journal of Fourier Analysis and Applications*, vol. 15, no. 4, pp. 488–501, 2009.

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

[23] S. Bahmani and J. Romberg, "Phase Retrieval Meets Statistical Learning Theory: A Flexible Convex Relaxation," *arXiv preprint arXiv:1610.04210*, 2016.

[24] T. Goldstein and C. Studer, "PhaseMax: Convex Phase Retrieval via Basis Pursuit," *arXiv preprint arXiv:1610.07531*, 2016.

[25] O. Dhifallah, C. Thrampoulidis, and Y. M. Lu, "Phase retrieval via linear programming: Fundamental limits and algorithmic improvements," in *55th Annual Allerton Conference on Communication, Control, and Computing*, 2017.

[26] M. Neykov, Z. Wang, and H. Liu, "Agnostic estimation for misspecified phase retrieval models," in *Advances in Neural Information Processing Systems*, 2016, pp. 4089–4097.

[27] Y. M. Lu and G. Li, "Phase transitions of spectral initialization for high-dimensional nonconvex estimation," *arXiv preprint arXiv:1702.06435*, 2017.

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

[29] D. Amelunxen, M. Lotz, M. B. Mccoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Information and Inference*, vol. 3, no. 3, pp. 224–294, Sep. 2014.

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

[31] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

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

[33] R. Van Handel, "Chaining, interpolation, and convexity," *arXiv preprint arXiv:1508.05906*, 2015.

[34] S. Mendelson, "Learning without concentration," *Journal of the ACM (JACM)*, vol. 62, no. 3, p. 21, 2015.

[35] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991, vol. 23.

[36] R. T. Rockafellar, *Convex analysis*. Princeton university press, 1997, vol. 28.

[37] J. A. Tropp, "Convex recovery of a structured signal from independent random linear measurements," in *Sampling Theory, a Renaissance*. Springer, 2015, pp. 67–101.

[38] D. Hsu, S. Kakade, T. Zhang, *et al.*, "A tail inequality for quadratic forms of subgaussian random vectors," *Electronic Communications in Probability*, vol. 17, 2012.

A Model parameters for standard link functions

Here, we evaluate the model parameters defined in Section 2.2.2 for some standard link functions. In what follows, as defined in Section 2.2.2, γ denotes a standard Gaussian random variable.

Quadratic with additive bounded noise. Let $\mathbf{y}_i = f\big(\mathbf{a}_i^T\mathbf{x}_0\big) = |\mathbf{a}_i^T\mathbf{x}_0|^2 + \mathbf{z}_i$, with $|\mathbf{z}_i| \leq \delta$ and independent of everything else. In this case, we have

$$\mu = \frac{1}{2}\mathbb{E}[f''(\gamma)] = \frac{1}{2}\cdot 2 = 1.$$

Now the model parameter τ can be evaluated as

$$\tau^2 = \mathbb{E}[\big(f\big(\gamma\big) - \gamma^2\big)^2\big] = \mathbb{E}\big[z^2\big] \leq \delta^2$$

or $\tau \leq \delta$. Similarly, one can easily verify that, for the underlying link function, we have $\upsilon \leq \delta$ and $\chi \leq \delta\sqrt{2}$.

Quadratic with additive gaussian noise. Let $\mathbf{y}_i = f\big(\mathbf{a}_i^T\mathbf{x}_0\big) = |\mathbf{a}_i^T\mathbf{x}_0|^2 + \mathbf{z}_i$, with the noise $\mathbf{z}_i \sim \mathcal{N}(0, s^2)$ and independent of everything else. Again, it holds that $\mu = 1$. Now, we have

$$\tau^2 = \mathbb{E}[\big(f\big(\gamma\big) - \gamma^2\big)^2\big] = \mathbb{E}\big[z^2\big] = s^2$$

or $\tau = s$. Moreover, it can be easily verified that $\upsilon = s$. Furthermore, we have

$$\chi = \mathbb{E}\big[(\gamma^2-1)^2\cdot\big(f(\gamma) - \gamma^2\big)^2\big] = \mathbb{E}\big[(\gamma^2-1)^2\cdot z^2\big] = \mathbb{E}[\gamma^4 + 1 - 2\gamma^2]\cdot\mathbb{E}[z^2] = 2s^2$$

or $\chi = s\sqrt{2}$.

Quadratic with Poisson noise. Let $\mathbf{y}_i \sim \text{Poisson}\big((\mathbf{a}_i^T\mathbf{x}_0)^2\big)$. Towards computing the summary parameters, let Y denote a random variable with distribution $Y \sim \text{Poisson}\big(\gamma^2\big)$. Then, by conditioning on γ, we can calculate the value of μ as follows.

$$\mu = \frac{1}{2}\mathbb{E}_{\gamma,Y}\big[(\gamma^2-1)Y\big] = \frac{1}{2}\mathbb{E}_{\gamma}\Big[(\gamma^2-1)\mathbb{E}\big[Y \mid \gamma\big]\Big] = \frac{1}{2}\mathbb{E}_{\gamma,Y}\big[(\gamma^2-1)\gamma^2\big] = 1.$$

Next, we utilize the value of μ to compute the parameter τ as

$$\tau^2 = \mathbb{E}_{\gamma,Y}\big[\big(Y-\gamma^2\big)^2\big] = \mathbb{E}_{\gamma}\Big[\mathbb{E}_Y\big[\big(Y-\gamma^2\big)^2\mid\gamma\big]\Big] = \mathbb{E}_{\gamma}\big[\gamma^2\big] = 1.$$

Similarly,

$$\upsilon^2 = \mathbb{E}_{\gamma,Y}\big[\gamma^2\cdot\big(Y-\gamma^2\big)^2\big] = \mathbb{E}_{\gamma}\Big[\gamma^2\cdot\mathbb{E}_Y\big[\big(Y-\gamma^2\big)^2\mid\gamma\big]\Big] = \mathbb{E}_{\gamma}\big[\gamma^4\big] = 3,$$

and

$$\chi^2 = \mathbb{E}_{\gamma,Y}\big[(\gamma^2-1)^2\cdot\big(Y-\gamma^2\big)^2\big] = \mathbb{E}_{\gamma}\Big[(\gamma^2-1)^2\cdot\mathbb{E}_Y\big[\big(Y-\gamma^2\big)^2\mid\gamma\big]\Big] = \mathbb{E}_{\gamma}\big[\gamma^6+\gamma^2-2\gamma^4\big] = 10.$$

B Mendelson's lower bound on nonnegative empirical process

In this section we deduce Lemma 3.2 from [34, Thm. 5.4]. We start with the following result, which is an immediate application of [34, Thm. 5.4] to our setting. In particular, we follow here the exposition in [37, Prop. 5.1].

Lemma B.1 (Lower bound for a nonnegative empirical process [34])**.** *Fix a set $\mathcal{C} \in \mathbb{S}_n$. Let $\mathbf{a} \sim \mathcal{N}(0, \mathbf{I})$, and let $\mathbf{a}_1, \dots, \mathbf{a}_m$ be independent copies of $\mathbf{a}$. Introduce the marginal tail function*

$$Q_\xi(\mathcal{C}; \mathbf{a}) := \inf_{V \in \mathcal{C}} \mathbb{P}(\, | \operatorname{tr}(\mathrm{V}\mathbf{a}_\mathrm{i}\mathbf{a}_\mathrm{i}^\mathrm{T}) | \geq \xi \,), \quad \text{where } \xi \geq 0.$$

Let $\varepsilon_1, \dots, \varepsilon_m$ be independent Rademacher random variables, independent from everything else, and define the empirical mean width of the set:

$$W_m(\mathcal{C}; \mathbf{a}) := \mathbb{E} \sup_{V \in \mathcal{C}} \operatorname{tr}(\mathrm{VH}) \qquad \text{where} \quad H := \frac{1}{\sqrt{m}} \sum_{i=1}^m \varepsilon_i \mathbf{a}_i \mathbf{a}_i^T.$$

Then, for any $\xi > 0$ and $t > 0$, with probability at least $1 - e^{-t^2/2}$ it holds,

$$\inf_{V \in \mathcal{C}} \left(\sum_{i=1}^m | \operatorname{tr}(\mathrm{V}\mathbf{a}_\mathrm{i}\mathbf{a}_\mathrm{i}^\mathrm{T}) |^2 \right)^{1/2} \geq \xi \sqrt{m} \cdot Q_{2\xi}(\mathcal{C}; \mathbf{a}) - 2W_m(\mathcal{C}; \mathbf{a}) - \xi t.$$

It is further shown in [37, Sec. 8.5] that if $\mathcal{C}$ is a subset of the sphere, i.e., if

$$V \in \mathcal{C} \implies \|V\|_F = 1, \tag{57}$$

then, for some absolute constant $c_0 > 0$ it holds

$$Q_1(\mathcal{C}; \mathbf{a}) \geq c_0.$$

Also, if (57) holds then note that the definition of $W_m(\mathcal{C}; \mathbf{a})$ coincides with the definition of $\omega_e(\mathcal{C})$ in Definition 3.1. Combining these with Lemma B.1, it follows that

$$\inf_{V \in \mathcal{C}} \|\mathcal{A}(V)\|_2 = \inf_{V \in \mathcal{C}} \left(\sum_{i=1}^m | \operatorname{tr}(\mathrm{V}\mathbf{a}_\mathrm{i}\mathbf{a}_\mathrm{i}^\mathrm{T}) |^2 \right)^{1/2} \geq c_0 \cdot \sqrt{m} - 2\omega_e(\mathcal{C}) - t. \tag{58}$$

with probability at least $1 - e^{-t^2/4}$. The set $\mathcal{E}$ of Lemma 3.2 clearly satisfies (57). Hence, this concludes the proof.

C Proofs for Section 2

C.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 3.4.2. Specifically, we use Lemma 3.7.

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 X_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

$$\mathbb{E}_{\boldsymbol{\eta}}[\omega_e(\mathcal{C}_0; \boldsymbol{\eta})] \leq C \cdot k \cdot \sqrt{\log(n/k)} \cdot \left(\frac{\mathbb{E}_{\boldsymbol{\eta}} \|\boldsymbol{\eta}\|_2}{\sqrt{m}} + \frac{\mathbb{E}_{\boldsymbol{\eta}} \|\boldsymbol{\eta}\|_\infty}{\sqrt{m}} \sqrt{2 \log(n/k)} \right). \tag{59}$$

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

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

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

$$\mathbb{E}_{\boldsymbol{\eta}}[\omega_e(\mathcal{C}_0;\boldsymbol{\eta})] \le C\cdot\tau\cdot k\cdot\sqrt{\log(n/k)}\left(1+\sqrt{2\log(n/k)}\right). \tag{60}$$

It only remains to compute the Gaussian width of $\mathcal{C}_{\text{sparse}}$. It is know that (e.g., [31, Prop. 3.10])

$$\omega_g(\mathcal{C}_{\text{sparse}}) \le 3\sqrt{2}k\sqrt{\log(n/k)}. \tag{61}$$

Putting (60) and (61) in (40), we have shown that if $m \ge ck^2\log(n/k)$ it holds with high probability:

$$\|\widehat{X}-\mu X_0\|_F \le C_1\cdot\frac{C_2\cdot\tau\cdot k\sqrt{\log(n/k)}\left(1+\sqrt{2\log(n/k)}\right)+6\upsilon\cdot k\sqrt{\log(n/k)}+\chi}{\sqrt{m}}. \tag{62}$$

Then, the statement of Theorem 2.1 follow from (62) by taking n (in fact n/k) large enough (which will depend on C_2, τ, ρ and χ).

C.2 Proof of Theorem 2.2

We apply Theorem 3.1, but we need to control the weighted empirical width. Recall from (35) that $\mathcal{C}_0 = \mathcal{C}_+ \cap \mathcal{C}_{\mathcal{R}}$, hence

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

First, we prove that

$$(19) \Rightarrow (39). \tag{63}$$

On the one hand, if $\sqrt{m} \ge c\sqrt{n}$, then from polarity arguments in Lemma 3.6 (specifically, Equation (56)) we have that $\omega_e(\mathcal{C}_+;\mathbf{1}) \le C\sqrt{n} \le C'\sqrt{m}$. On the other hand, if

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

then by (54):

$$\omega_e(\mathcal{C}_{\mathcal{R}}\cap\mathcal{S}) \le C\cdot\left(\gamma_2(\mathcal{C}_{\mathcal{R}}\cap\mathcal{S};\|\cdot\|_F)+\frac{\gamma_1(\mathcal{C}_{\mathcal{R}}\cap\mathcal{S};\|\cdot\|_2)}{\sqrt{m}}\right) \le C'\sqrt{n}.$$

Thus, we have shown (63).

Next, we show that

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

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 3.6:

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

On the other hand, from Eqn. (55),

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

Thus, we have shown (C.2).

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

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

This follows using $\omega_g(\mathcal{C}_0) \leq \min\{\omega_e(\mathcal{C}_+), \omega_e(\mathcal{C}_{\mathcal{R}})\}$ and the well-known bound $\omega_g(\mathcal{C}_+) \leq 6\sqrt{n}$ (e.g. [31]).

D Controlling the weighted empirical width via generic chaining

D.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 D.1 (Admissible sequence)**.** Given a set $\mathcal{T}$, an admissible sequence is an increasing sequence $\{\mathcal{A}_n\}$ of partitions of $\mathcal{T}$ such that $\text{card}(\mathcal{A}_n) \leq N_n := 2^{2^n}$.

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 D.2. Given $\alpha > 0$ and a metric space $(\mathcal{T}, d)$, we define

$$\gamma_\alpha(\mathcal{T}, d) = \inf \sup_{\mathbf{t}} \sum_{n \geq 0} 2^{n/\alpha} \Delta(A_n(\mathbf{t})), \tag{64}$$

where $\Delta(A_n(\mathbf{t}))$ denotes the diameter of the set $A_n(\mathbf{t})$. The infimum in (64) is taken over all admissible sequences.

Proposition D.1 ([32, 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 $\forall\ \mathbf{s}, \mathbf{t} \in \mathcal{T},\ \forall\ u > 0$,*

$$\mathbb{P}\left(|X_{\mathbf{s}} - X_{\mathbf{t}}| \geq u\right) \leq 2\exp\left(-\min\left(\frac{u^2}{d_2(\mathbf{s},\mathbf{t})^2}, \frac{u}{d_1(\mathbf{s},\mathbf{t})}\right)\right). \tag{65}$$

Then,

$$\mathbb{E}\sup_{\mathbf{s},\mathbf{t}\in\mathcal{T}} |X_{\mathbf{s}} - X_{\mathbf{t}}| \leq L \cdot \big(\gamma_1(\mathcal{T}, d_1) + \gamma_2(\mathcal{T}, d_2)\big). \tag{66}$$

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 F.2 that the process satisfies:

$$\mathbb{P}\left(\ |X_V - X_U| \geq t\ \right) \leq 2\exp\left(-\min\left\{\frac{t^2}{4\|\mathbf{p}\|_2^2\|V-U\|_F^2}\ ,\ \frac{t}{\|\mathbf{p}\|_\infty\|V-U\|_2}\right\}\right), \tag{67}$$

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

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

E Computing the weighted empirical quadratic Gaussian width via polarity

Here, we prove Lemmas 3.6 and 3.7.

E.1 Proof of Lemma 3.6

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

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

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 $\mathcal{C}_+ = \mathcal{D}(\mathcal{R}, \mu X_0)$ for the following convex function $\mathcal{R} : \mathbb{S}_n \to \mathbb{R}$

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

where, we also used the fact that $X_0 \succeq 0$.

The proof of the lemma follows rather standard arguments. Similar calculation are performed in [37, Sec. 8.6.2]. In particular, the second statement of the lemma can also be found in [37].

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

$$\lambda = \lambda_{\max}(H_{22}) \tag{68}$$

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

$$\begin{aligned} \omega_e^2(\mathcal{C}_+; \mathbf{p}) &\leq \mathbb{E}\big[\inf_{V \in \partial \mathcal{R}(X_0)} \|H - \lambda \cdot V\|_F^2 \big] \\ &\overset{(i)}{=} \mathbb{E}\big[(h_{11} - \lambda)^2\big] + 2\mathbb{E}\big[\|\mathbf{h}_{12}\|_2^2\big] + \mathbb{E}\big[\inf_{S: \ \lambda_{\max}(S) \leq 1} \|H_{22} - \lambda S\|_F^2 \big] \\ &\overset{(ii)}{=} \mathbb{E}\big[(h_{11} - \lambda)^2\big] + 2 \cdot \mathbb{E}\big[\|\mathbf{h}_{12}\|_2^2\big]. \end{aligned} \tag{69}$$

where: *(i)* uses the fact that (e.g., [37, Sec. 8.6.1])

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

(ii) uses (68). It remains to compute the two terms in (69). On the one hand, we have

$$\mathbb{E}\big[\|\mathbf{h}_{12}\|_2^2\big] = \frac{1}{m} \mathbb{E}\big[\sum_{\ell=2}^{n} \big(\sum_{i \in [m]} p_i \varepsilon_i \mathbf{a}_{i,1} \mathbf{a}_{i,\ell} \big)^2 \big] = \frac{1}{m} \big[\sum_{\ell=2}^{n} \sum_{i \in [m]} p_i^2 \big] = \frac{\|\mathbf{p}\|_2^2}{m} (n-1). \tag{70}$$

On the other hand, by interlacing of eigenvalues

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

and, as shown in Lemma F.1:

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

Thus,

$$\mathbb{E}\big[(h_{11} - \lambda)^2\big] = \frac{1}{m}\mathbb{E}\big[\big(\sum_{i\in[m]} p_i \varepsilon_i \mathbf{a}_{i,1}^2\big)^2\big] + \lambda^2 \leq 3\frac{\|\mathbf{p}\|_2^2}{m} + \lambda^2. \tag{71}$$

Putting together (70) and (71) in (69), it follows that

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

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 $m \geq c \cdot n$: $\omega_e(\mathcal{C}_+; \mathbf{p}) \leq C_3\sqrt{n}$, as desired. □

E.2 Proof of Lemma 3.7

Here, we prove Lemma 3.7. 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., $\mathcal{C}_{\text{sparse}} = \mathcal{D}(\mathcal{R}, \mu X_0)$ for $\mathcal{R} : \mathbb{S}_n \to \mathbb{R}$ such that

$$\mathcal{R}(X) := \|X\|_1 = \sum_{k,\ell\in[n]} X_{k\ell}.$$

By applying Propostion 3.1, we bound the empirical quadratic Gaussian width of $\mathcal{C}_{\text{sparse}}$ in the following lemma.

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 $i, j \in [n]$,

$$S_{i,j} = \begin{cases} 1 & \text{if } X_{i,j} > 0, \\ -1 & \text{if } X_{i,j} < 0, \\ \varsigma_{i,j} \in [-1, 1] & \text{if } X_{i,j} = 0. \end{cases} \tag{72}$$

Now using the rotational invariance of the Gaussian distribution, we may assume without loss of generality that $\mathbf{x}_0 = (x_1^0, \ldots, x_n^0)$ be such that $\mathbf{x}_{0,i} > 0$ for $i = 1, \ldots, k$ and $\mathbf{x}_{0,i} = 0$ for $i = k+1, \ldots, n$. 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

$$S_{i,j} = \begin{cases} 1 & \text{if } i, j \in [k], \\ \varsigma_{i,j} \in [-1, 1] & \text{otherwise.} \end{cases} \tag{73}$$

By employing Proposition 3.1, we get

$$\begin{aligned}\omega_e^2(\mathcal{C}_{\text{sparse}};\mathbf{p}) &\le \mathbb{E}\big[\inf_{S\in\partial\mathcal{R}(X_0)}\|H-\lambda\cdot S\|_F^2\big]\\ &\overset{(i)}{=}\mathbb{E}\Big[\sum_{(i,j)\in[k]\times[k]}(h_{ij}-1)^2+\sum_{(i,j)\notin[k]\times[k]}\mathrm{st}(h_{ij};\lambda)^2\Big],\end{aligned}\tag{74}$$

where (i) follows from (73). Note that $\mathrm{st}(\cdot;\lambda)$ denotes the soft thresholding function, which is defined as

$$\mathrm{st}(h;\lambda)=\begin{cases}\frac{h}{|h|}\cdot(|h|-\lambda) & \text{if } |h|\ge\lambda,\\ 0 & \text{otherwise.}\end{cases}\tag{75}$$

Next, we separately bound the two terms appearing in (74). On the one hand,

$$\begin{aligned}\mathbb{E}\Big[\sum_{(i,j)\in[k]\times[k]}(h_{i,j}-1)^2\Big] &= \mathbb{E}\Big[\sum_{i=1}^{k}(h_{ii}-1)^2\Big]+\mathbb{E}\Big[\sum_{(i\neq j)\in[k]\times[k]}(h_{ij}-1)^2\Big]\\ &= k\Big(\lambda^2+\mathbb{E}[h_{11}^2]\Big)+k(k-1)\Big(\lambda^2+\mathbb{E}[h_{12}^2]\Big)\\ &\overset{(i)}{=} k\Big(\lambda^2+3\frac{\|\mathbf{p}\|_2^2}{m}\Big)+k(k-1)\Big(\lambda^2+\frac{\|\mathbf{p}\|_2^2}{m}\Big),\end{aligned}\tag{76}$$

where (i) follows from the following observations:

$$\mathbb{E}[h_{11}^2]=\frac{1}{m}\cdot\mathbb{E}\big[\big(\sum_{i\in[m]}p_i\varepsilon_i a_{i,1}^2\big)^2\big]=\frac{1}{m}\cdot\sum_{i\in[m]}p_i^2\mathbb{E}[a_{i,1}^4]=3\frac{\|\mathbf{p}\|_2^2}{m}\tag{77}$$

and

$$\mathbb{E}[h_{12}^2]=\frac{1}{m}\cdot\mathbb{E}\big[\big(\sum_{i\in[m]}p_i\varepsilon_i a_{i,1}a_{i,2}\big)^2\big]=\frac{1}{m}\cdot\sum_{i\in[m]}p_i^2\mathbb{E}[a_{i,1}^2a_{i,2}^2]=\frac{\|\mathbf{p}\|_2^2}{m}.\tag{78}$$

On the other hand, the second term in (74) gives

$$\mathbb{E}\Big[\sum_{(i,j)\notin[k]\times[k]}\mathrm{st}(h_{ij};\lambda)^2\Big]=(n-k)\mathbb{E}\big[\mathrm{st}(h_{k+1,k+1};\lambda)^2\big]+(n-k)(n+k-1)\mathbb{E}\big[\mathrm{st}(h_{1,k+1};\lambda)^2\big].\tag{79}$$

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

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

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:

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

From these, we get that

$$\begin{aligned}\mathbb{E}\big[\mathrm{st}(h_{1,k+1};\lambda)^2\big] &= \int_{\lambda}^{\infty} 2(t-\lambda)\mathbb{P}[|h_{1,k+1}|>t]\mathrm{d}t \\ &\overset{(i)}{\leq} \int_{\lambda}^{\infty} 4(t-\lambda)e^{\max\left\{-\frac{mt^2}{\psi_1^2\|\mathbf{p}\|_2^2},-\frac{\sqrt{m}t}{\psi_1\|\mathbf{p}\|_\infty}\right\}}\mathrm{d}t \\ &= 4\max\left\{\int_{\lambda}^{\infty}(t-\lambda)e^{-\frac{mt^2}{\psi_1^2\|\mathbf{p}\|_2^2}}\mathrm{d}t, \int_{\lambda}^{\infty}(t-\lambda)e^{-\frac{\sqrt{m}t}{\psi_1\|\mathbf{p}\|_\infty}}\mathrm{d}t\right\} \\ &\leq \max\left\{\frac{\|\mathbf{p}\|_2^2\psi_1^2}{m}\left(\frac{\lambda\sqrt{2}\sqrt{m}}{\|\mathbf{p}\|_2\psi_1}-1\right)e^{-\frac{m\lambda^2}{\psi_1^2\|\mathbf{p}\|_2^2}}\ ,\ 4\frac{\|\mathbf{p}\|_\infty^2\psi_1^2}{m}e^{-\frac{\lambda\sqrt{m}}{\psi_1\|\mathbf{p}\|_\infty}}\right\}, \end{aligned} \tag{80}$$

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., [28, Thm. 2.8.2]) and ψ_1 is an absolute constant that denotes the sub-exponential norm of $\varepsilon_i a_{i,1}a_{i,k+1}$ [28, 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

$$\begin{aligned}\mathbb{E}\big[\mathrm{st}(h_{k+1,k+1};\lambda)^2\big] &= \int_{\lambda}^{\infty} 2(t-\lambda)\mathbb{P}[h_{k+1,k+1}>t]\mathrm{d}t \\ &= \max\left\{\frac{\|\mathbf{p}\|_2^2\psi_2^2}{m}\left(\frac{\lambda\sqrt{2}\sqrt{m}}{\|\mathbf{p}\|_2\psi_2}-1\right)e^{-\frac{m\lambda^2}{\psi_2^2\|\mathbf{p}\|_2^2}}\ ,\ 4\frac{\|\mathbf{p}\|_\infty^2\psi_2^2}{m}e^{-\frac{\lambda\sqrt{m}}{\psi_2\|\mathbf{p}\|_\infty}}\right\} \end{aligned} \tag{81}$$

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 (76), (79), (80), (81) with (74), for this choise of λ we obtain that

$$\begin{aligned}\omega_e^2(\mathcal{C}_{\text{sparse}};\mathbf{p}) &\leq 3k^2\left(\lambda^2+\frac{\|\mathbf{p}\|_2^2}{m}\right)+(n^2-k^2)\frac{k^2}{n^2}4\lambda^2 \\ &\leq \frac{C}{m}k^2\left(\|\mathbf{p}\|_2\sqrt{\log\left(\frac{n^2}{k^2}\right)}+\|\mathbf{p}\|_\infty\log\left(\frac{n^2}{k^2}\right)\right)^2, \end{aligned} \tag{82}$$

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

□

F Spectral norm of weighted sum of Gaussian outer products

Lemma F.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 F.2).

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

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

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 F.2) to find that for any $t > 0$:

$$\mathbb{P}\left(S \ge t \right) \le \exp\left(-\min\left\{ \frac{t^2}{4\|\mathbf{p}\|_2^2} \ , \ \frac{t}{2\|\mathbf{p}\|_\infty} \right\} \right), \tag{83}$$

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., [28, Chapter 4] $|\mathcal{N}| \le 9^n$ and

$$\|\tilde{H}\|_2 \le 2 \cdot \max_{\mathbf{v} \in \mathcal{N}} \mathbf{v}^T \tilde{H} \mathbf{v}. \tag{84}$$

Also, $\mathbb{E}[\|\tilde{H}\|_2] = \int_0^\infty \mathbb{P}[\|\tilde{H}\|_2 \ge t] \mathrm{d}t$. 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:

$$\begin{aligned}
\mathbb{E}[\|\tilde{H}\|_2] &\le \delta + \int_\delta^\infty \mathbb{P}[\|\tilde{H}\|_2 \ge t] \mathrm{d}t \\
&\le \delta + 2 \cdot \int_\delta^\infty \mathbb{P}\big[\max_{\mathbf{v} \in \mathcal{N}} \mathbf{v}^T \tilde{H} \mathbf{v} \ge t/2 \big] \mathrm{d}t \le \qquad (85) \\
&\le \delta + 2 \cdot 9^n \cdot \max\left\{ \int_\delta^\infty e^{-\frac{t^2}{16\|\mathbf{p}\|_2^2}} \mathrm{d}t \ , \ \int_\delta^\infty e^{-\frac{t}{4\|\mathbf{p}\|_\infty}} \mathrm{d}t \right\} \\
&\le \delta + \max\left\{ 8\sqrt{\pi}\|\mathbf{p}\|_2 e^{-\frac{\delta^2}{16\|\mathbf{p}\|_2^2}} \ , \ 8\|\mathbf{p}\|_\infty e^{-\frac{\delta}{4\|\mathbf{p}\|_\infty}} \mathrm{d}t \right\} \\
&\le C_2 \cdot (\|\mathbf{p}\|_2\sqrt{n} + \|\mathbf{p}\|_\infty n), \qquad \text{for sufficiently large } n. \qquad (86)
\end{aligned}$$

□

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

$$\mathbb{P}\left(\ |\langle \tilde{H}, V \rangle| \ge t \ \right) \le 2\exp\left(-\min\left\{ \frac{t^2}{4\|\mathbf{p}\|_2^2\|V\|_F^2} \ , \ \frac{t}{2\|\mathbf{p}\|_\infty\|V\|_2} \right\} \right) \tag{87}$$

Proof. Let

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

and

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

With these, note that $\mathbf{b}^T M \mathbf{b} = \sum_{i \in [m]} p_i \varepsilon_i \,\mathrm{tr}(\mathrm{V}\mathbf{a_i}\mathbf{a_i^{\mathrm{T}}})$ 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. [38, Prop. 1.1]) and the following observations:

$$\|M\|_F^2 = \|\mathbf{p}\|_2^2\|V\|_F^2 \qquad \text{and} \qquad \|M\|_2 = \|\mathbf{p}\|_\infty\|V\|_2. \tag{88}$$

□