Optimal Convex Lifted Sparse Phase Retrieval and PCA With an Atomic Matrix Norm Regularizer

Andrew D. McRae, *Member, IEEE*, Justin Romberg, *Fellow, IEEE*, and Mark A. Davenport, *Senior Member, IEEE*

***Abstract*— We present novel analysis and algorithms for solving sparse phase retrieval and sparse principal component analysis (PCA) with convex lifted matrix formulations. The key innovation is a new mixed atomic matrix norm that, when used as regularization, promotes low-rank matrices with sparse factors. We show that convex programs with this atomic norm as a regularizer provide near-optimal sample complexity and error rate guarantees for sparse phase retrieval and sparse PCA. While we do not know how to solve the convex programs exactly with an efficient algorithm, for the phase retrieval case we carefully analyze the program and its dual and thereby derive a practical heuristic algorithm. We show empirically that this practical algorithm performs similarly to existing state-of-the-art algorithms.**

***Index Terms*— Sparse phase retrieval, sparse PCA, convex relaxation, atomic norm.**

I. Introduction

A. Sparsity, Phase Retrieval, and PCA

CONSIDER the standard linear regression problem in which we make observations of the form $y_i = \langle x_i, \beta^* \rangle + \xi_i$, $i = 1, \dots, n$, where $\beta^* \in \mathbf{R}^p$ is a vector we want to estimate, $x_1, \dots, x_n \in \mathbf{R}^p$ are measurement vectors, and $\xi_1, \dots, \xi_n$ represent noise or other error. If the x_i's are chosen randomly and independently (e.g., i.i.d. Gaussian), and the noise is zero-mean and independent with $\mathrm{var}(\xi_i) \leq \sigma^2$, it is well-known that in general, we need[1] $n \gtrsim p$ measurements to estimate β^* meaningfully, and the best possible error we can obtain is $\|\hat{\beta} - \beta^*\|_2 \lesssim \sigma\sqrt{p/n}$.

We can potentially do much better if we exploit *sparsity* in the vector β^*. If β^* has (at most) s nonzero entries, the standard LASSO algorithm, which requires solving an ℓ_1-regularized least-squares optimization problem, yields an estimator $\hat{\beta}$ satisfying $\|\hat{\beta} - \beta^*\|_2 \lesssim \sigma\sqrt{(s/n)\log(p/s)}$ as long as the number of measurements satisfies $n \gtrsim s\log(p/s)$ (see, e.g., [1, Chapter 10]). Thus by using a convex regularized optimization problem we can exploit sparsity to reduce the number of measurements n and the estimation error proportionally to sparsity level (i.e., the number of nonzero entries in β^*). In this paper, we seek to extend this phenomenon to two problems: *phase retrieval* and *principal component analysis* (PCA). To introduce our main results, we briefly describe phase retrieval and PCA and their sparse variants. We focus on the formulations most relevant to our results. More complete background and related literature can be found in Sections I-B and I-C.

In phase retrieval, we seek to estimate a vector β^* from n noisy *quadratic* observations of the form $y_i = |\langle x_i, \beta^* \rangle|^2 + \xi_i$. The nonlinearity in the measurement model makes estimation and analysis more complicated than if our measurements are linear. To get around this, a common approach is to note that for any $x, \beta \in \mathbf{R}^p$, $|\langle x, \beta \rangle|^2 = \langle X, B \rangle_{\mathrm{HS}}$, where $X = x \otimes x$ and $B = \beta \otimes \beta$ are rank-1 positive semidefinite (PSD) matrices, and $\langle \cdot, \cdot \rangle_{\mathrm{HS}}$ denotes the Hilbert-Schmidt (Frobenius) matrix inner product. We can then write our observations as the *linear* measurements $y_i = \langle X_i, B^* \rangle_{\mathrm{HS}} + \xi_i$, where $B^* = \beta^* \otimes \beta^*$ and $X_i = x_i \otimes x_i$. This is often called a "lifted" formulation, since we are mapping the parameter of interest from $\mathbf{R}^p$ to the larger space of $p \times p$ PSD matrices. If the x_i's are randomly chosen (say, Gaussian), and we solve the semidefinite program

$$\widehat{B} = \arg\min_{B \succeq 0} \ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle X_i, B \rangle_{\mathrm{HS}})^2,$$

we can bound $\|\widehat{B} - B^*\|_{\mathrm{HS}} \lesssim \sigma\sqrt{p/n}$ as long as $n \gtrsim p$, where σ is the standard deviation of the ξ_i's. (As shown in [2], this implies that the leading eigenvector of $\widehat{B}$ is close to β^* up to its sign.) Both the sample complexity and the error rate are comparable to those in ordinary linear regression.

In PCA, we observe n i.i.d. random vectors $\{x_i\}_{i=1}^n$, and we want to estimate the leading eigenvector v_1 of the covariance matrix $\Sigma = \mathbf{E}(x_1 \otimes x_1)$. Again, this can be solved in a lifted manner with a semidefinite program, noting that

$$P_1 := v_1 \otimes v_1 = \arg\max_{P \in \mathbf{R}^{p \times p}} \ \langle \Sigma, P \rangle_{\mathrm{HS}} \text{ s.t. } \|P\|_* \leq 1.$$

Manuscript received 8 November 2021; revised 26 September 2022; accepted 5 December 2022. Date of publication 14 December 2022; date of current version 16 February 2023. This work was supported in part by NSF under Grant CCF-1718771 and Grant CCF-2107455. *(Corresponding author: Andrew D. McRae.)*

Andrew D. McRae is with the Institute of Mathematics, EPFL, 1015 Lausanne, Switzerland (e-mail: andrew.mcrae@epfl.ch).

Justin Romberg and Mark A. Davenport are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: jrom@ece.gatech.edu; mdav@gatech.edu).

Communicated by S. Jalali, Associate Editor for Machine Learning and Statistics.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3228508.

Digital Object Identifier 10.1109/TIT.2022.3228508

[1]Here and throughout the paper, $\lesssim$ and $\gtrsim$ denote, respectively, $\leq$ and $\geq$ within absolute constants.

An estimator $\widehat{P}$ of P_1 is obtained[2] by replacing Σ with the empirical covariance $\widehat{\Sigma}$. Again, if $n \gtrsim p$, we can recover P_1 within error proportional to $\sqrt{p/n}$ (where the constants depend on the gap between the first and second leading eigenvalues of Σ).

Sparse phase retrieval seeks to combine phase retrieval with sparse recovery. If β^* is s-sparse, and we observe $y_i = |\langle x_i, \beta^* \rangle|^2 + \xi_i$ for $i \in \{1, \dots, n\}$, can we recover β^* with a similar sample complexity and error as in linear sparse recovery? Similarly, the question we consider in *sparse PCA* is whether, if the leading eigenvector v_1 is s-sparse, we can recover it with a similar sample complexity and error as in linear recovery.

Our main contributions are the following:

- We present novel convex relaxations of the sparse phase retrieval and sparse PCA problems that use both a lifted formulation and a sparsity-inducing regularization, and we prove that for both problems, an estimator computed via a convex program achieves an $O(s \log(p/s))$ sample complexity as in linear sparse recovery. Furthermore, in both problems, the estimators achieve the optimal $O(\sqrt{(s/n)\log(p/s)})$ error rate (with the caveat, for the sparse phase retrieval problem with unbounded noise, that n may need to be larger than the minimum sample complexity to obtain this optimal rate).
- Although we do not know how to compute the convex programs exactly (we suspect they may, in fact, be computationally intractable), we present a heuristic motivated by a careful analysis of the dual problem and the problem's optimality conditions, and we show that in the case of sparse phase retrieval, the resulting algorithm achieves nearly identical empirical performance to existing state-of-the-art sparse phase retrieval algorithms.

In the following sections, we describe the sparse phase retrieval and sparse PCA problems in more detail, and we review the related literature.

B. Sparse Phase Retrieval

Phase retrieval in p dimensions with (sub-)Gaussian measurements is by now well-studied. If we have n observations of the form $y_i \approx |\langle x_i, \beta^* \rangle|^2$, we can solve the optimization problem

$$\hat{\beta} = \arg\min_{\beta \in \mathbf{R}^p} \sum_{i=1}^{n} (y_i - |\langle x_i, \beta^* \rangle|^2)^2. \tag{1}$$

Unfortunately, this is a nonconvex problem, so there is no immediately obvious way to solve it efficiently. (A similar optimization problem and similar nonconvexity appear if we instead write our measurements without the square, i.e., our observations are $\approx |\langle x_i, \beta^* \rangle|$.)

Most approaches to this algorithmic difficulty fall into one of two categories. One method is to optimize a nonconvex loss function such as (1) directly (and iteratively) with a suitable initialization (e.g., [3]). The other is the lifted semidefinite approach outlined in Section I-A. For example, Candés and Li [4] show that if the design vectors x_i are Gaussian, $y_i = |\langle x_i, \beta^* \rangle|^2 + \xi_i$, and we have $n \gtrsim p$ measurements, solving

$$\widehat{B} = \arg\min_{B \succeq 0} \sum_{i=1}^{n} |y_i - \langle X_i, B \rangle_{\mathrm{HS}}|$$

achieves $\|\widehat{B} - B^*\|_F \lesssim \frac{1}{n}\sum_{i=1}^{n} |\xi_i|$ with high probability. In the case of zero-mean random noise with standard deviation σ, we can, by using a squared loss, improve this to $\|\widehat{B} - B^*\|_F \lesssim \sigma\sqrt{p/n}$ (see [5]). Thus we can solve the phase retrieval problem with a sample complexity and susceptibility to noise proportional to the dimension p; this is the same complexity as ordinary linear regression.

Several results have been published on how to adapt iterative nonconvex phase retrieval algorithms to the sparse setting [6], [7], [8], [9], [10], [11]. Some [7], [10] do indeed achieve $O(\sigma\sqrt{(s/n)\log p})$ error bounds with zero-mean noise—this is very close to the optimal rate in linear sparse recovery (the rest do not analyze theoretically the noisy case). However, the theory in this literature requires $n \gtrsim s^2 \log p$, which, unless s is very small, is much larger than what is required in linear sparse recovery. As Soltanolkotabi [12] points out, the key difficulty is finding a good initialization for the algorithms—once we are close enough to β^*, we only need[3] $n \gtrsim_{\log} s$ measurements to converge to a correct estimate. In practice, the first initialization step is often to estimate the support of β^*; the best known methods require $n \gtrsim_{\log} s^2$ measurements. We compare several of these algorithms (in addition to that of the purely algorithmic/empirical work [13]) to ours empirically in Section V-C, and we see that all of them appear empirically to have *linear* sample complexity in s. Another similar iterative algorithm is given in [14]; it has similar sample complexity requirements but, interestingly, it is derived from a more abstract compression-based algorithm that, though not practically computable, does obtain optimal $O(s)$ sample complexity.

We see qualitatively similar sample complexity requirements in the works [15] and [16], which extend to the sparse case the convex PhaseMax framework [17], [18]. Both results only require $n \gtrsim_{\log} s$ measurements if we already have an "anchor" vector $\beta_0 \in \mathbf{R}^p$ that has significant correlation with β^*. However, it is not known how to find such a β_0 (in a computationally efficient manner) without $n \gtrsim_{\log} s^2$ measurements.

More related to our results are methods to adapt the lifted convex phase retrieval approach to the sparse setting. The foundational theoretical work in this area is by Li and Voroninski [19], although some work (mostly empirical) appeared in [20] and [21]. The key idea is that if $\beta^* \in \mathbf{R}^p$ is s-sparse, the lifted version $B^* = \beta^* \otimes \beta^*$ is both rank-1 and at most s^2-sparse. In the noiseless case, they solve the optimization

[2]It would be computationally suboptimal in practice to compute the leading eigenvector of $\widehat{\Sigma}$ with a semidefinite program, but this formulation helps motivate our approach to the sparse case.

[3]Here and hereafter, $\gtrsim_{\log}$ ($\lesssim_{\log}$) will denote "greater (less) than within a logarithmic factor."

problem

$$\widehat{B} = \arg\min_{B \succeq 0} \ \lambda_1 \operatorname{tr}(B) + \lambda_2 \|B\|_{1,1} \quad \text{s.t. } \langle X_i, B\rangle_{\mathrm{HS}} = y_i, \ i = 1, \dots, n, \quad (2)$$

where $\|\cdot\|_{1,1}$ denotes the elementwise ℓ_1 norm of a matrix. The trace regularization term promotes low rank, while the ℓ_1 norm promotes sparsity. As with the nonconvex methods, their theory requires $n \gtrsim s^2 \log p$ measurements to get exact recovery. The result of [5], when specialized to sparse phase retrieval, extends this approach to the noisy case, getting, within log factors, the same $O(s^2)$ sample and noise complexity.

Finally, we note that although we are primarily concerned with generic measurement vectors x_i (e.g., sub-Gaussian), one can obtain better theoretically guaranteed sample complexity with practical algorithms if we have complete control over how the measurements are chosen; see, for example, [22], [23].

C. Sparse PCA

PCA is a well-established technique with which, given points $x_1, \dots, x_n \in \mathbf{R}^p$, we try to find a low-dimensional linear (or affine) subspace that contains most of the energy in the data. If $x_1, \dots, x_n$ have zero empirical mean (e.g., after centering), the closest r-dimensional subspace to the points (in mean square ℓ_2 distance) is the space spanned by the top r eigenvectors of the empirical covariance matrix $\widehat{\sum} = \frac{1}{n}\sum_{i=1}^n x_i \otimes x_i$.

For simplicity, take $r = 1$. Suppose the x_i's are i.i.d. copies of a random variable x with true covariance Σ with eigenvalue decomposition $\Sigma = \sum_\ell \sigma_\ell v_\ell \otimes v_\ell$, where $\sigma_1 > \sigma_2 \geq \cdots \geq \sigma_p$. If x is Gaussian, and $\sigma_2 \gtrsim \frac{\sigma_1}{p-1}$, then, with high probability [24],

$$\|\widehat{\sum} - \Sigma\|_2 \lesssim \sqrt{\sigma_1 \frac{\sigma_1 + (p-1)\sigma_2}{n}} \lesssim \sqrt{\sigma_1 \sigma_2 \frac{p}{n}}.$$

Then, if $\hat{v}_1$ is the leading eigenvector of $\widehat{\sum}$, the Davis-Kahan $\sin\Theta$ theorem gives

$$\|\hat{v}_1 \otimes \hat{v}_1 - v_1 \otimes v_1\|_2 \lesssim \frac{\sqrt{\sigma_1\sigma_2}}{\sigma_1 - \sigma_2}\sqrt{\frac{p}{n}}.$$

This rate is minimax-optimal over general covariance matrices with the given σ_1, σ_2 (see [25]).

When p is large compared to n, we need to impose more structure on Σ to recover the leading eigenvector(s) accurately. In sparse PCA, we consider the case in which the eigenvector(s) of interest are *sparse*. This problem has been extensively studied in the past decade: see [26] for a recent review.

In the single-eigenvector recovery case ($r = 1$), Cai et al. [27] show that if the leading eigenvector v_1 is s-sparse, the minimax rate for all estimators $\hat{v}_1$ of v_1 over the simple class $\{\Sigma = \sigma_2 I_p + (\sigma_1 - \sigma_2) v_1 \otimes v_1 : v_1 \ s\text{-sparse}, \|v_1\|_2 = 1\}$ is

$$\|\hat{v}_1 \otimes \hat{v}_1 - v_1 \otimes v_1\|_2 \approx \frac{\sqrt{\sigma_1\sigma_2}}{\sigma_1 - \sigma_2}\sqrt{\frac{s\log(p/s)}{n}}.$$

While this theoretical result is clean and achieves our desire to bring sparse-recovery sample complexity and error to the PCA problem, one practical problem remains: how do we *compute* an estimator $\hat{v}_1$ that achieves these theoretical properties? The optimal estimator proposed in [27] is, to quote that paper "computationally intensive." As with sparse phase retrieval, the best theoretical results for computationally efficient algorithms require $n \gtrsim_{\log} s^2$ to guarantee accurate recovery (see, e.g., [27], [28]). Once again, proper initialization (often by estimating the support of v_1) is the key difficulty.

There is strong evidence to suggest that this s^2 barrier may be inescapable for computationally efficient algorithms. Recent results suggest that any statistically optimal estimator that requires fewer measurements must be NP-hard to compute. Berthet and Rigollet [29] showed that if a certain testing problem in random graph theory (the *planted clique problem*) is NP-hard to compute in certain regimes (which is widely believed although so-far unproved in standard computational models), then accurately *testing for the existence of* a sparse leading eigenvector when $n \lesssim_{\log} s^2$ is NP-hard. Berthet, and Samworth [30], [31] further refine this by showing that, under a similar assumption, there is no efficiently computable consistent estimator of v_1 when $n \lesssim_{\log} s^2$.

II. Key Tool: A Sparsity-and-Low-Rank–Inducing Atomic Norm

To motivate our approach, consider the optimization problem (2) from [19] for sparse phase retrieval or its least-squares version

$$\widehat{B} = \arg\min_{B \succeq 0} \ \frac{1}{2n}\sum_{i=1}^n (y_i - \langle X_i, B\rangle_{\mathrm{HS}})^2 + \lambda_1 \operatorname{tr}(B) + \lambda_2 \|B\|_{1,1}. \quad (3)$$

It turns out that quadratic (in sparsity) $O(s^2)$ complexity is a fundamental performance bound for this class of methods. Our target matrix B^* has two kinds of structure: it is rank-1 and s^2-sparse. The trace regularization in our estimator encourages low rank, while the ℓ_1 regularization encourages sparsity. However, recent work [32] and [33] has shown it is impossible to take advantage of both kinds of structure simultaneously with a regularizer that is merely a convex combination of the two structure-inducing regularizers; the best we can do is exploit either the low rank as in non-sparse phase retrieval, in which case we get $O(p)$ complexity, or the s^2-sparsity, in which case we get $O(s^2)$ complexity.

To see intuitively why we have this problem, note that the nuclear norm and elementwise ℓ_1 norm are both examples of *projective tensor norms* [34]. For matrix A of any size,

$$\|A\|_* = \inf \ \{\sum \|u_k\|_2 \|v_k\|_2 : A = \sum u_k \otimes v_k\}$$

and

$$\|A\|_{1,1} = \inf \ \{\sum \|u_k\|_1 \|v_k\|_1 : A = \sum u_k \otimes v_k\}$$

Equivalently, these norms are atomic norms [35] where the atoms are rank-1 matrices with unit ℓ_2 or ℓ_1 norms. For a PSD matrix, the trace is the nuclear norm, so the regularizer

in (3) can be expressed as

$$\lambda_1 \operatorname{tr}(B) + \lambda_2 \|B\|_{1,1} = \lambda_1 \inf \{\sum \|u_k\|_2 \|v_k\|_2 : B = \sum u_k \otimes v_k\} + \lambda_2 \inf \{\sum \|w_k\|_1 \|z_k\|_1 : B = \sum w_k \otimes z_k\}. \quad (4)$$

A key feature of $B^* = \beta^* \otimes \beta^*$ is that the factors of its rank-1 decomposition have a certain ℓ_2 norm *and* are sparse. Because the two infima in (4) are separate, the regularizer promotes matrices with two *separate* atomic decompositions of low ℓ_2 and ℓ_1 norm respectively. It does not encourage a decomposition into low-rank matrices with factors that have *simultaneously* low ℓ_2 norm and low ℓ_1 norm.

Inspired by the framework of Haeffele and Vidal [36], we propose the following regularizer:

$$\|B\|_{*,s} := \inf \{\sum \theta_s(u_k, v_k) : B = \sum u_k \otimes v_k\}, \quad (5)$$

where

$$\theta_s(u,v) = (\|u\|_2 + \frac{1}{\sqrt{s}}\|u\|_1)\left(\|v\|_2 + \frac{1}{\sqrt{s}}\|v\|_1\right),$$

and $s > 0$ is a parameter that represents the sparsity (or an approximation thereof) of the vector we are interested in recovering. For some intuition on this choice of regularizer, note that

$$\{A : \|A\|_{*,s} \le 1\} \approx \operatorname{conv}\{u \otimes v : \|u\|_2 = \|v\|_2 = 1, u, v \text{ are } s\text{-sparse}\},$$

by which we mean that either is contained within a modest scaled version of the other. One direction is a simple consequence of the fact that for an s-sparse vector u, $\|u\|_1 \le \sqrt{s}\|u\|_2$. The other direction is provided by Lemma 6 in Section A. Thus $\|\cdot\|_{*,s}$ is (equivalent to) an atomic norm whose atoms are precisely the type of matrix we expect B^* to be.[4] Similar notions of atomic norms that promote simultaneous low rank and sparsity have appeared in [37] and [33].

We will show in the next section that using $\|\cdot\|_{*,s}$ as a regularizer in lifted formulations of sparse phase retrieval and PCA gives sample complexity and error bounds nearly identical to the linear regression case.

III. Theoretical Guarantees for Atomic-Norm Regularized Estimators

In this section, we state precisely our main problems, assumptions, abstract convex optimization algorithm, and theoretical guarantees.

A. Sparse Phase Retrieval

Suppose $\beta^* \in \mathbf{R}^p$ is an s-sparse vector. Let x be a random vector in $\mathbf{R}^p$. We observe n i.i.d. copies $(x_1, y_1), \dots, (x_n, y_n)$ of the random couple (x, y), where y is a real random variable whose distribution conditioned on x depends only on $\langle x, \beta^* \rangle^2$ (i.e., $y \sim p_y(y \mid \langle x, \beta^* \rangle^2)$). Let $\xi := y - \langle x, \beta^* \rangle^2$ denote the "noise." We make the following assumptions:

[4] If we "guess wrongly" the sparsity of β^*, we can still get similar results with different constants of equivalence.

Assumption 1 (Sub-Gaussian Measurements): The entries $(x^{(1)}, \dots, x^{(p)})$ of x are i.i.d. real random variables with $\mathbf{E}\, x^{(\ell)} = 0$, $\mathbf{E}(x^{(\ell)})^2 = 1$, $\mathbf{E}(x^{(\ell)})^4 > 1$, and sub-Gaussian norm $\|x^{(\ell)}\|_{\psi_2} \le K$ for some $K > 0$.

Note that the fourth-moment assumption excludes Rademacher random variables. In what follows, for simplicity of presentation, all dependence on K and the difference $\mathbf{E}(x^{(\ell)})^4 - 1$ will be subsumed into unspecified constants.

Assumption 2 (Zero-Mean, Bounded-Moment Noise): $\mathbf{E}[\xi \mid x] = 0$ almost surely, and, for all $u \in \mathbf{R}^p$ such that $\|u\|_2 \le 1$,

$$\mathbf{E}\, \xi^2 \langle x, u \rangle^4 \le \sigma^2(\beta^*),$$

where $\sigma^2(\beta^*)$ is a quantity that possibly depends on the vector β^*, the distribution of x, and the conditional distribution of y. Furthermore, there are $M, \eta \ge 0$ such that

$$\|\xi \langle x, u \rangle^2\|_\alpha \le M \alpha^{\eta+1}$$

for $\alpha \ge 3$ and all $u \in \mathbf{R}^p$ such that $\|u\|_2 \le 1$ (where $\|Z\|_\alpha := (\mathbf{E}|Z|^\alpha)^{1/\alpha}$ for any random variable Z).

Our two working examples are the following:

- Independent additive noise: ξ is independent of all other quantities, in which case we can take $\sigma^2(\beta) \approx \operatorname{var}(\xi)$, and M and η depend on the moments of ξ.
- Poisson noise: $y \sim \operatorname{Poisson}(\langle x, \beta^* \rangle^2)$ conditioned on x. In this case, under Assumption 1, we can take $\sigma^2(\beta^*) \approx \|\beta^*\|_2^2$, $M \approx \|\beta^*\|_2 + 1$, and $\eta = 1$ (we prove this in Section D).

As before, we lift the problem into the space of PSD matrices by setting $B^* = \beta^* \otimes \beta^*$ and $X = x \otimes x$. We then choose a regularization parameter $\lambda \ge 0$ and compute our estimate by the following optimization problem:

$$\widehat{B} = \underset{B \in \mathbf{R}^{p \times p}}{\arg\min} \frac{1}{2n} \sum_{i=1}^n (y_i - \langle X_i, B \rangle_{\mathrm{HS}})^2 + \lambda \|B\|_{*,s}. \quad (6)$$

We then have the following guarantee for sample complexity and error, proved in Section IV-A:

Theorem 1: Suppose Assumption 1 and 2 hold. Suppose β^* is s-sparse and that the number of measurements n satisfies $n \gtrsim s \log(ep/s)$. If the regularization parameter satisfies

$$\lambda \gtrsim \sqrt{\frac{s \log(ep/s)}{n} \sigma^2(\beta^*)} + \frac{M}{n^{1-c}} \left(s \log \frac{ep}{s}\right)^{\eta+1},$$

where $c \approx (s \log(ep/s))^{-1}$, then, with probability at least $1 - e^{-bn} - e^{-s}(s/p)^s$ (where $b > 0$ is a constant), the estimator $\widehat{B}$ from (6) satisfies

$$\|\widehat{B} - B^*\|_* \lesssim \lambda.$$

Remark 1: For simplicity of presentation, we assume that the sparsity level s used in the regularizer is in fact (an upper bound on) the sparsity of β^*. We could easily extend our results to the "misspecified" case $\|\beta^*\|_0 = s_0 > s$.

Remark 2: By a standard argument (found, e.g., in [2]), if $\hat{\beta} \otimes \hat{\beta}$ is the closest rank-1 approximation to $\widehat{B}$, then $\hat{\beta}$ satisfies

$$\min\{\|\hat{\beta} - \beta^*\|_2, \|\hat{\beta} + \beta^*\|_2\} \lesssim \frac{\lambda}{\|\beta^*\|_2}.$$

Remark 3: The required sample complexity $s\log(ep/s)$ is precisely the optimal sample complexity from traditional linear sparse recovery. For large n, the noise error rate (with appropriately chosen λ) is also the optimal $\sqrt{(s/n)\log(ep/s)}$, but, if $\eta > 0$, achieving this rate may require n to be significantly larger than $s\log(ep/s)$. More precisely, the first term containing the optimal rate will dominate if and only if

$$n^{1-2c} \gtrsim \frac{M^2}{\sigma^2(\beta^*)}\left(s\log\frac{ep}{s}\right)^{1+2\eta}.$$

If the noise ξ is bounded, we can take $\eta = 0$, and we only need $n^{1-2c} \gtrsim s\log\frac{ep}{s}$ to obtain the optimal error rate. For most interesting cases (where c is very small), this is negligibly different from the sample complexity requirement. If ξ is (conditionally) sub-Gaussian, we can take $\eta = 1/2$, in which case we need $n^{1-2c} \gtrsim_{\log} s^2$. If ξ is (conditionally) sub-exponential, as in the Poisson noise case, we need $n^{1-2c} \gtrsim_{\log} s^3$. The need for larger n comes (in our proof) from concentration inequalities for sums of terms of the form $\xi\langle x,u\rangle^2$ for arbitrary vectors u; these terms have larger moments than the $\xi\langle x,u\rangle$ terms we would typically see in linear settings. This could perhaps be improved with judicious truncation as in, for example, [38].

Remark 4: In the independent additive noise case, one can check that our proof gives a high-probability bound uniform over s-sparse β^*. If $\mathrm{var}(\xi) = \sigma^2$, we get, for appropriately chosen λ,

$$\|\widehat{B} - B^*\|_* \lesssim \sqrt{\frac{s\log(ep/s)}{n}}\sigma + \frac{M}{n^{1-c}}\left(s\log\frac{ep}{s}\right)^{\eta+1}.$$

Remark 5: In the Poisson observation case, we obtain, for appropriately chosen λ,

$$\|\widehat{B} - B^*\|_* \lesssim \sqrt{\frac{s\log(ep/s)}{n}}\|\beta^*\|_2 + \frac{\|\beta^*\|_2+1}{n^{1-c}}\left(s\log\frac{ep}{s}\right)^2.$$

When $\beta^* \neq 0$, and n is large enough that the first error term dominates, we have, up to a sign, that

$$\|\hat{\beta} - \beta^*\|_2 \lesssim \sqrt{\frac{s\log(ep/s)}{n}},$$

where $\hat{\beta}$ is the appropriately-scaled leading eigenvector of $\widehat{B}$. Thus we get an error bound does that not depend on $\|\beta^*\|_2$.

Remark 6: If there is no noise ($\xi = 0$), our analysis could easily be adapted to study the problem

$$\min_B \ \|B\|_{*,s} \text{ s.t. } \langle X_i, B\rangle_{\mathrm{HS}} = y_i, i = 1,\dots,n.$$

To understand how to use our proof techniques, note that any solution $\widehat{B}$ to the above problem satisfies $\sum_{i=1}^n \langle X_i, H\rangle_{\mathrm{HS}}^2 = 0$ and

$$0 \geq \|\widehat{B}\|_{*,s} - \|B^*\|_{*,s} \geq \langle W_{B^*}, H\rangle_{\mathrm{HS}},$$

for any subgradient $W_{B^*} \in \partial\|B^*\|_{*,s}$, where $H = \widehat{B} - B$.

B. Sparse PCA

We can apply the atomic regularizer to the sparse PCA problem via another standard lifted formulation:

Theorem 2: Suppose we observe n i.i.d. copies of the p-dimensional vector $x \sim \mathcal{N}(\mu, \Sigma)$, where $\Sigma = \sigma_1 v_1 \otimes v_1 + \Sigma_2$, v_1 is s-sparse and unit-norm, $\sigma_1 > \|\Sigma_2\| =: \sigma_2$, and $\Sigma_2 v_1 = 0$. Choose

$$\lambda \gtrsim \sqrt{\sigma_1\sigma_2}\sqrt{\frac{s\log(ep/s)}{n}}$$

and let

$$\widehat{P} = \underset{P\in\mathbf{R}^{p\times p}}{\arg\min} -\langle\widehat{\sum}, P\rangle_{\mathrm{HS}} + \lambda\|P\|_{*,s} \text{ s.t. } \|P\|_* \leq 1, \quad (7)$$

where

$$\widehat{\sum} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})\otimes(x_i - \bar{x}) = \left(\frac{1}{n}\sum_{i=1}^n x_i\otimes x_i\right) - \bar{x}\otimes\bar{x}$$

is the empirical covariance of $x_1,\dots,x_n$ ($\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$).

For $t > 0$, if $n \gtrsim \max\left\{s\log\frac{ep}{s}, \left(\frac{\sigma_1}{\sigma_1-\sigma_2}\right)^2 t\right\}$, then, with probability at least $1 - e^{-t} - 3\,e^{-s}(s/p)^s$,

$$\|\widehat{P} - P_1\|_F \lesssim \frac{\lambda}{\sigma_1 - \sigma_2},$$

where $P_1 = v_1 \otimes v_1$.

We prove this result fully in Section C. A sketch of the proof is provided in Section IV-B.

Remark 7: The assumption that x is Gaussian could easily be relaxed to $x = \Sigma^{1/2}z$, where z is a sub-Gaussian random vector, as in, for example, [25].

Remark 8: For properly chosen λ the resulting error rate

$$\|\widehat{P} - P_1\|_F \lesssim \frac{\sqrt{\sigma_1\sigma_2}}{\sigma_1 - \sigma_2}\sqrt{\frac{s\log(ep/s)}{n}}$$

matches the minimax lower bounds in [25] and [27].

C. PSD Constraints and Another Regularizer

For phase retrieval and PCA, it is natural to restrict our estimators to be PSD. All of our theoretical results hold if we add a $B \succeq 0$ constraint to (6) or a $P \succeq 0$ constraint to (7).

Unlike the nuclear norm case (where the optimal decomposition is the singular value decoposition, which is identical to the eigenvalue decomposition for a PSD matrix), it is not clear whether every PSD matrix B admits a symmetric (i.e., $u_k = v_k$) optimal decomposition with regard the definition of $\|B\|_{*,s}$ in (5). Therefore, it is natural to define as a new regularizer the following gauge function/asymmetric norm on the space of PSD matrices: for $B \succeq 0$,

$$\Theta_s(B) = \inf\left\{\sum\theta_s(u_k,u_k) : B = \sum u_k\otimes u_k\right\}.$$

All of our theoretical and computational results in Section III and V can be easily extended to this choice of regularizer. This choice of regularizer is computationally convenient because if we optimize over a matrix B by optimizing over factors u_k, v_k such that $B = \sum_k u_k\otimes v_k$ (see Section V-B), we can enforce a PSD constraint simply by forcing $u_k = v_k$.

IV. Proof Highlights

In this section, we outline the proofs of Theorem 1 and 2. We fully prove Theorem 1 from some technical lemmas, while we sketch the proof of Theorem 2

A. Sparse Phase Retrieval Proof

In this section, we prove Theorem 1, which is our error bound for sparse phase retrieval. We will use the following key technical lemmas:

Lemma 1 (Subgradients of Mixed Atomic Norm): Suppose $\beta \in \mathbf{R}^p$ is s-sparse, and let $B = \beta \otimes \beta$. Then, for every matrix $A \in \mathbf{R}^{p\times p}$, there exists $W \in \partial \|B\|_{*,s}$ such that

$$\langle W, A\rangle_{\text{HS}} \geq \frac{1}{10}\|A\|_{*,s} - 5\|A\|_F.$$

Lemma 2 (Empirical Process Bound): Let $G_1, \dots, G_n$ be i.i.d. copies of a random matrix $G \in \mathbf{R}^{p\times p}$, where, for all $u, v \in \mathbf{R}^p$, $\langle Gu, v\rangle$ has zero mean,

$$\mathbf{E}\langle Gu, v\rangle^2 \leq \sigma^2 \|u\|_2^2 \|v\|_2^2,$$

and

$$\|\langle Gu, v\rangle\|_\alpha \leq M\alpha^{\eta+1}\|u\|_2\|v\|_2$$

for all $\alpha \geq 3$.

Let $Z = \frac{1}{n}\sum_{i=1}^n G_i$. For $s \geq 1$, with probability at least $1 - e^{-s}(s/p)^s$,

$$\sup_{\|A\|_{*,s}\leq 1} \langle Z, A\rangle_{\text{HS}} \lesssim \sigma\sqrt{\frac{s\log(ep/s)}{n}} + \frac{M}{n^{1-c}}\left(s\log\frac{ep}{s}\right)^{\eta+1},$$

where $c \approx \frac{1}{s\log(ep/s)}$.

Lemma 3 (Restricted Lower Isometry): Let $x_1, \dots, x_n$ be i.i.d. copies of a random vector x satisfying Assumption 1, and let $X_i = x_i \otimes x_i$. Suppose

$$n \gtrsim s\log\frac{ep}{s},$$

and let $C \geq 1$ be a fixed constant. With probability at least $1 - e^{-bn}$ (for some $b > 0$), the following event holds: For all $A \in \mathbf{R}^{p\times p}$ such that

$$\|A\|_{*,s} \leq C\|A\|_F,$$

we have

$$\frac{1}{n}\sum_{i=1}^n \langle X_i, A\rangle_{\text{HS}}^2 \gtrsim \|A\|_F^2,$$

where the constant in the lower bound depends on C.

Lemma 1 is proved in Section A. Lemma 2 and 3 are proved in Section B. With these, we can prove the sparse phase retrieval error bound:

Proof of Theorem 1: Applying Lemma 2 to the random matrices $G_i = \xi_i X_i$, we can choose λ according to the theorem statement with large enough constant so that, with probability at least $1 - e^{-s}(s/p)^s$,

$$\sup_{\|A\|_{*,s}\leq 1}\left\langle \frac{1}{n}\sum_{i=1}^n \xi_i X_i, A\right\rangle_{\text{HS}} \leq \frac{\lambda}{20}.$$

Furthermore, by Lemma 3, for $n \gtrsim s\log\frac{ep}{s}$ (with large enough constant), we have, with probability at least $1 - e^{-bn}$,

$$\frac{1}{n}\sum_{i=1}^n \langle X_i, A\rangle_{\text{HS}}^2 \gtrsim \|A\|_F^2$$

for all A satisfying $\|A\|_{*,s} \leq 100\|A\|_F$.

The intersection of these events occurs with probability at least $1 - e^{-s}(s/p)^s - e^{-bn}$. In what follows, we assume this holds.

Let $\widehat{B}$ be the solution to (6). Writing $F(B)$ as the objective function, the convexity of the optimization problem implies that

$$\begin{aligned} 0 &\leq \langle \nabla F(\widehat{B}), B^* - \widehat{B}\rangle_{\text{HS}} \\ &= \frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, \widehat{B}\rangle_{\text{HS}})\langle X_i, \widehat{B} - B^*\rangle_{\text{HS}} \\ &\quad + \lambda\langle W_{\widehat{B}}, B^* - \widehat{B}\rangle_{\text{HS}}, \end{aligned}$$

for any $W_{\widehat{B}} \in \partial\|\widehat{B}\|_{*,s}$. By the monotonicity of (sub)gradients of convex functions, we have that, for any $W \in \partial\|B^*\|_{*,s}$, $\langle W - W_{\widehat{B}}, B^* - \widehat{B}\rangle_{\text{HS}} \geq 0$, and therefore

$$0 \leq \frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, \widehat{B}\rangle_{\text{HS}})\langle X_i, \widehat{B} - B^*\rangle_{\text{HS}} + \lambda\langle W, B^* - \widehat{B}\rangle_{\text{HS}}.$$

Let $H = \widehat{B} - B^*$. Using the fact that $(y_i - \langle X_i, \widehat{B}\rangle_{\text{HS}})\langle X_i, \widehat{B} - B^*\rangle_{\text{HS}} = \xi_i\langle X_i, H\rangle_{\text{HS}} - \langle X_i, H\rangle_{\text{HS}}^2$, we have

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \langle X_i, H\rangle_{\text{HS}}^2 &\leq \frac{1}{n}\sum_{i=1}^n \xi_i\langle X, H\rangle_{\text{HS}} - \lambda\langle W, H\rangle_{\text{HS}} \\ &\leq \frac{\lambda}{20}\|H\|_{*,s} - \lambda\langle W, H\rangle_{\text{HS}}. \end{aligned}$$

By Lemma 1, there exists $W \in \partial\|B^*\|_{*,s}$ such that

$$\langle W, H\rangle_{\text{HS}} \geq \frac{1}{10}\|H\|_{*,s} - 5\|H\|_F.$$

Therefore, we have

$$\frac{1}{n}\sum_{i=1}^n \langle X_i, H\rangle_{\text{HS}}^2 \leq 5\lambda\|H\|_F - \frac{\lambda}{20}\|H\|_{*,s}.$$

Because the left side of this inequality is nonnegative, we have $\|H\|_{*,s} \leq 100\|H\|_F$. Then, by restricted lower isometry, we have

$$\|H\|_F^2 \lesssim \lambda\|H\|_F.$$

The result immediately follows. □

B. Sparse PCA Proof Sketch

The proof of Theorem 2 is somewhat messier than the proof of Theorem 1 above, so we do not go into all of the details here. We refer the reader to Section C for the full proof.

If $\widehat{P}$ is an optimal solution of (7), one can obtain, similarly to the proof of Theorem 1, that

$$\langle \widehat{\sum}, H\rangle_{\text{HS}} \geq \lambda\langle W, H\rangle_{\text{HS}}$$

for any $W \in \partial\|P_1\|_{*,s}$, where $H = \widehat{P} - P_1$. Choosing W according to Lemma 1, we obtain

$$\langle \widehat{\sum}, H\rangle_{\mathrm{HS}} \geq \lambda\left(\frac{1}{10}\|H\|_{*,s} - 5\|H\|_F\right).$$

By analysis similar to Lemma 2, one can show that

$$|\langle \widehat{\sum} - \Sigma, H\rangle_{\mathrm{HS}}| \lesssim \sqrt{\sigma_1\sigma_2\frac{s\log(ep/s)}{n}}\|H\|_{*,s} + \sigma_1\sqrt{\frac{t}{n}}|\langle H, P_1\rangle_{\mathrm{HS}}|$$

with probability at least $1 - e^{-t} - 3\ e^{-s}(s/p)^s$ when $n \gtrsim s\log(ep/s)$. For λ chosen so that the coefficient of $\|H\|_{*,s}$ above is $\leq \lambda/10$, we get, on this event,

$$\langle \Sigma, H\rangle_{\mathrm{HS}} \gtrsim -\lambda\|H\|_F - \sigma_1\sqrt{\frac{t}{n}}|\langle H, P_1\rangle_{\mathrm{HS}}|.$$

Now, note that because $\|\widehat{P}\|_* \leq 1$, we have the following:

- $|\langle H, P_1\rangle_{\mathrm{HS}}| = 1 - \langle \widehat{P}, P_1\rangle_{\mathrm{HS}}$, and
- $\langle \Sigma, H\rangle_{\mathrm{HS}} = \sigma_1(\langle \widehat{P}, P_1\rangle_{\mathrm{HS}} - 1) + \langle \Sigma_2, \widehat{P}\rangle_{\mathrm{HS}} \leq \sigma_1(\langle \widehat{P}, P_1\rangle_{\mathrm{HS}} - 1) + \sigma_2(1 - \langle \widehat{P}, P_1\rangle_{\mathrm{HS}})$.

Then, using the assumption that $n \gtrsim \frac{\sigma_1^2}{(\sigma_1-\sigma_2)^2}t$, we get

$$\begin{aligned}(\sigma_1 - \sigma_2)(1 - \langle \widehat{P}, P_1\rangle_{\mathrm{HS}}) &\lesssim \left(\sigma_1 - \sigma_2 - \sigma_1\sqrt{\frac{t}{n}}\right)(1 - \langle \widehat{P}, P_1\rangle_{\mathrm{HS}}) \\ &\lesssim \lambda\|H\|_F.\end{aligned}$$

Finally, one can show that $\|\widehat{P}\|_F \leq \|\widehat{P}\|_* \leq 1$ implies $\|H\|_F^2 \lesssim 1 - \langle \widehat{P}, P_1\rangle_{\mathrm{HS}}$, which immediately gives the result.

V. Computational Limitations and a Practical Algorithm for Phase Retrieval

Although the mixed atomic norm $\|\cdot\|_{*,s}$ is a powerful theoretical tool, it is not clear how to calculate (let alone optimize) it for a general matrix in practice, since it is defined as an infimum over infinite sets of possible factorizations.

A warning that computations with these atomic regularizers may be difficult in general is that they can be used to get $O_{\log}(s)$ sample complexity for sparse PCA, which, as discussed in Section I-C, is widely believed to be impossible with efficient algorithms.

In this section, we will analyze the convex programs more carefully, with a particular focus on phase retrieval.[5] We will analyze the optimality conditions via a dual problem and thereby develop a heuristic algorithm.

This problem was studied in greater generality in [36]. Their Corollary 1 is similar to our Corollary 1. However, our analysis of the dual problem is quite different from their perturbation argument, and we can much more easily apply our method to the sparse PCA optimization problem (7) with its inequality constraint. Furthermore, we think the reader will benefit from our deriving the optimality conditions from more elementary principles for the particular problem we are trying to solve.

[5]While our algorithmic approach led to strong empirical performance for sparse phase retrieval, the approach was less effective for sparse PCA. We leave a more thorough investigation of this phenomenon for future work.

A. Factorization, Duality, and Optimality Conditions

To move toward a practical algorithm, we consider optimizing (6) in factored form; rather than optimizing over B directly, we optimize over the factors $\{u_k, v_k\}$ of a factorization $B = \sum_k u_k \otimes v_k$. Then (6) is equivalent to

$$\min_{\{u_k,v_k\}\subset\mathbf{R}^p} \frac{1}{2n}\sum_{i=1}^n\left(y_i - \left\langle X_i, \sum_k u_k\otimes v_k\right\rangle_{\mathrm{HS}}\right)^2 + \lambda\sum_k \theta_s(u_k, v_k). \quad (8)$$

The obvious drawback to this form is that the optimization problem is no longer convex; therefore, it is not clear whether finding a global minimum is computationally feasible.

To determine how well a factored algorithm works (e.g., to certify optimality), we examine a dual problem to (6). We formulate the dual via a trick found in [39]: note that $b^2/2 = \max_a\ ab - a^2/2$ (achieved if and only if $a = b$), and therefore

$$\begin{aligned}&\min_{B\in\mathbf{R}^{p\times p}} \frac{1}{2n}\sum_{i=1}^n(y_i - \langle X_i, B\rangle_{\mathrm{HS}})^2 + \lambda\|B\|_{*,s} \\ &\quad= \min_{B\in\mathbf{R}^{p\times p}} \frac{1}{2n}\sum_{i=1}^n\max_{\alpha_i}\ \left(2\alpha_i(y_i - \langle X_i, B\rangle_{\mathrm{HS}}) - \alpha_i^2\right) + \lambda\|B\|_{*,s} \\ &\quad\geq \max_{\alpha\in\mathbf{R}^n}\ \left[\frac{1}{n}\sum_{i=1}^n\left(\alpha_i y_i - \frac{\alpha_i^2}{2}\right)\right. \\ &\qquad\left. + \min_{B\in\mathbf{R}^{p\times p}}\left(\lambda\|B\|_{*,s} - \frac{1}{n}\sum_{i=1}^n\alpha_i\langle X_i, B\rangle_{\mathrm{HS}}\right)\right],\end{aligned}$$

where the inequality comes from swapping the maximum over $\alpha = (\alpha_1, \ldots, \alpha_n)$ and the minimum over B.

Define the dual norm $\|\cdot\|_{*,s}^*$ by

$$\|Z\|_{*,s}^* = \max_{\substack{B\in\mathbf{R}^{p\times p}\\ \|B\|_{*,s}\leq 1}}\ \langle Z, B\rangle_{\mathrm{HS}}.$$

Because $\|\cdot\|_{*,s}^*$ is nonnegatively homogeneous,

$$\min_{B\in\mathbf{R}^{p\times p}}\left(\lambda\|B\|_{*,s} - \left\langle\frac{1}{n}\sum_{i=1}^n\alpha_i X_i, B\right\rangle_{\mathrm{HS}}\right) = \begin{cases}0 & \text{if } \left\|\frac{1}{n}\sum_{i=1}^n\alpha_i X_i\right\|_{*,s}^* \leq \lambda \\ -\infty & \text{otherwise.}\end{cases}$$

Therefore, a dual formulation of (6) is the convex problem

$$\max_{\alpha\in\mathbf{R}^n}\left(\frac{1}{n}\sum_{i=1}^n\alpha_i y_i - \frac{\alpha_i^2}{2}\right)\ \text{s.t.}\ \left\|\frac{1}{n}\sum_{i=1}^n\alpha_i X_i\right\|_{*,s}^* \leq \lambda. \quad (9)$$

Before we go further, note that,

$$\|Z\|_{*,s}^* = \max_{\substack{u,v\in\mathbf{R}^p\\ \theta_s(u,v)\leq 1}}\ \langle Zu, v\rangle.$$

To see this, note that

$$\|Z\|_{*,s}^* = \sup \left\{ \langle Z, B\rangle_{\mathrm{HS}} : B, \{u_k\}, \{v_k\}, B = \sum_k u_k \otimes v_k, \sum_k \theta_s(u_k, v_k) \le 1 \right\}$$
$$= \sup \left\{ \sum_k \langle Z u_k, v_k\rangle : \sum_k \theta_s(u_k, v_k) \le 1 \right\}$$
$$= \sup \left\{ \sum_{k=1}^K \langle Z u_k, v_k\rangle : K \ge 1, \sum_{k=1}^K \theta_s(u_k, v_k) \le 1 \right\}.$$

For any finite sequence $\{u_k, v_k\}_{k=1}^K$ with $\sum_{k=1}^K \theta_s(u_k, v_k) \le 1$, if we let $k^* = \arg\max_{1\le k\le K} \frac{\langle Z u_k, v_k\rangle}{\theta_s(u_k, v_k)}$ and set $\tilde{u} = \frac{u_{k^*}}{\sqrt{\theta_s(u_{k^*}, v_{k^*})}}$ and $\tilde{v} = \frac{v_{k^*}}{\sqrt{\theta_s(u_{k^*}, v_{k^*})}}$, we will always have $\langle Z\tilde{u}, \tilde{v}\rangle \ge \sum_{k=1}^K \langle Z u_k, u_k\rangle$. Therefore,

$$\|Z\|_{*,s}^* = \sup \{\langle Z u, v\rangle : \theta_s(u, v) \le 1\}.$$

We can replace the supremum by a maximum because the objective function is continuous and the constraint set is compact.

Returning to the optimization problem, note that a feasible point α for the dual problem gives us a *lower* bound on the primal optimal value. If there exist $B \in \mathbf{R}^{p\times p}$, $\alpha \in \mathbf{R}^n$ such that α is feasible and the two objective functions are *equal*, then we know B is optimal for the primal problem. More precisely, (B, α) is an optimal primal-dual pair if and only if

(a) the primal objective function at B equals the dual objective functions at α, and
(b) α is feasible, i.e., $\left\|\frac{1}{n}\sum_{i=1}^n \alpha_i X_i\right\|_{*,s}^* \le \lambda$.

From the derivation of the dual problem above, (a) requires $\alpha_i = y_i - \langle X_i, B\rangle_{\mathrm{HS}}$. Making this substitution, setting the objective functions equal, and simplifying gives one direction of the following result:

Lemma 4: B solves (6) if and only if both of the following hold:

(a) $\frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, B\rangle_{\mathrm{HS}})\langle X_i, B\rangle_{\mathrm{HS}} = \lambda\|B\|_{*,s}$.
(b) $\left\|\frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, B\rangle_{\mathrm{HS}}) X_i\right\|_{*,s}^* \le \lambda$.

Proof: We have already shown that these conditions are *sufficient* for optimality. To see the other direction (that these conditions are *necessary* for optimality), note that $Z := \frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, B\rangle_{\mathrm{HS}}) X_i$ is the negative gradient of the empirical loss at B. Because condition (b) is equivalent to

$$\langle Z u, v\rangle \le \lambda\theta_s(u, v)\ \forall u \in \mathbf{R}^p,$$

if (b) does not hold, there exists some $\bar{u}, \bar{v} \in \mathbf{R}^p$ such that $\langle Z\bar{u}, \bar{v}\rangle > \lambda\theta_s(\bar{u}, \bar{v})$, and then we can decrease the objective function by moving to $B + \epsilon\bar{u}\otimes\bar{v}$ for some sufficiently small $\epsilon > 0$. Thus (b) is a necessary condition for the optimality of B.

Now suppose (b) holds, but (a) does not. Condition (b) implies that $\langle Z, B\rangle_{\mathrm{HS}} \le \lambda\|B\|_{*,s}$, so we must have $\langle Z, B\rangle_{\mathrm{HS}} < \lambda\|B\|_{*,s}$.

Let $B = \sum_k u_k \otimes v_k$ be an optimal factorization with respect to the definition of $\|B\|_{*,s}$, that is, such that $\|B\|_{*,s} = \sum_k \theta_s(u_k, v_k)$ (we assume, for clarity, that an optimal factorization exists—if not, we could use an approximation argument). There must be some u_k, v_k such that $\langle Z u_k, v_k\rangle < \lambda\theta_s(u_k, v_k)$. Then, modifying B by replacing (u_k, v_k) with $((1-\epsilon)u_k, (1-\epsilon)v_k)$ for some sufficiently small $\epsilon > 0$ will decrease the objective function.

□

Note that the proof of Lemma 4 gives us an explicit way to improve the objective function whenever one of the optimality conditions is not satisfied.

Applying our derivation to the factored optimization problem, we get the following result:

Corollary 1: B solves (6) and $B = \sum_k u_k \otimes v_k$ is an optimal factorization with respect to $\|\cdot\|_{*,s}$ (equivalently, $\{u_k, v_k\}$ solve (8)) if and only if the following hold:

1) For all k, $\frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, B\rangle_{\mathrm{HS}})\langle X_i u_k, v_k\rangle = \lambda\theta_s(u_k, v_k)$.
2) $\left\|\frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, B\rangle_{\mathrm{HS}}) X_i\right\|_{*,s}^* \le \lambda$; equivalently, for all $u, v \in \mathbf{R}^p$,

$$\frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, B\rangle_{\mathrm{HS}})\langle X_i u, v\rangle \le \lambda\theta_s(u, v).$$

Note that we have broken out condition (a) into individual equalities (rather than equating the sums of each side); condition (b) allows us to do this. It is even easier to find a descent direction when one of these conditions fails to hold, since the objective function of (8) already depends explicitly on the vectors u_k, v_k.

Note that condition (a) is much easier to verify than condition (b). We refer to $\{u_k, v_k\}$ as a *first-order stationary point* if it satisfies condition (a), since this is equivalent to a zero subgradient on the (nonzero) u_k's and v_k's (cf. Proposition 2 in [36]).

Although we are not focusing on sparse PCA here, it may be interesting to compare Corollary 1 to what we get for sparse PCA, particularly as PCA may be a fundamentally more difficult problem. A dual problem to (7) is

$$\arg\max_{Z\in\mathbf{R}^{p\times p}} -\|Z\| \text{ s.t } \|\widehat{\sum} - Z\|_{*,s}^* \le \lambda.$$

The following lemma gives (redundant) optimality conditions:

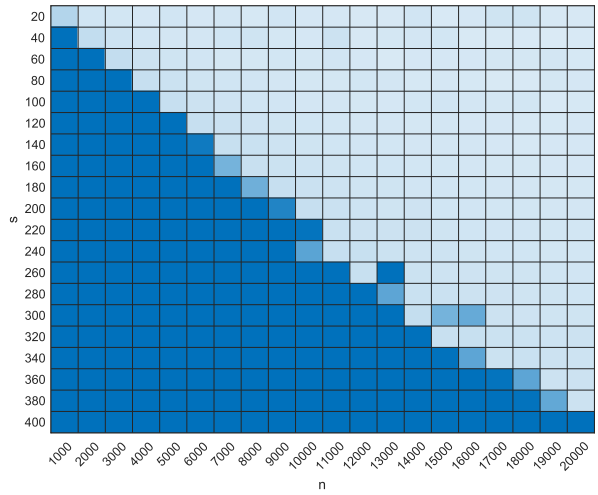
Lemma 5: P solves (7) if and only if $\|P\|_* = 1$ and there exists $Z \in \mathbf{R}^{p\times p}$ such that

1) $\|\widehat{\sum} - Z\|_{*,s}^* \le \lambda$,
2) $\langle \widehat{\sum} - Z, P\rangle_{\mathrm{HS}} = \lambda\|P\|_{*,s}$,
3) $\langle Z, P\rangle_{\mathrm{HS}} = \|Z\| = \|Z\|\|P\|_*$, and
4) $\|Z\| = \langle \widehat{\sum}, P\rangle_{\mathrm{HS}} - \lambda\|P\|_{*,s}$.

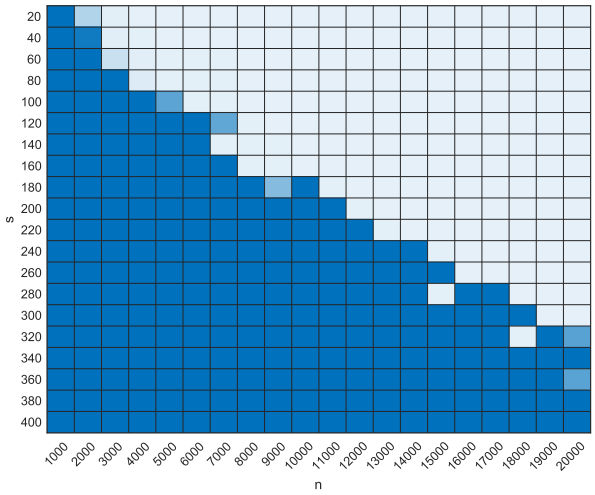
In the PCA case, the semidefinite version of the problem is somewhat simpler due to the fact that the nuclear norm becomes a trace. If we solve

$$\widehat{P} = \arg\min_{P\succeq 0} -\langle \widehat{\sum}, P\rangle_{\mathrm{HS}} + \lambda\Theta_s(P) \text{ s.t. } \mathrm{tr}(P) \le 1,$$

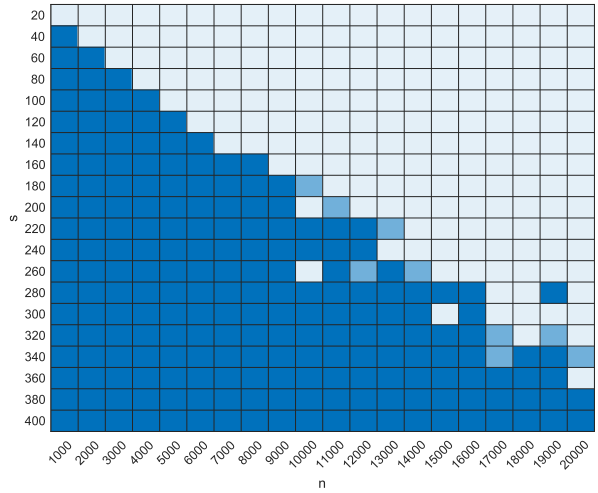
we get similar theoretical error guarantees as Theorem 2. Furthermore, $P = \sum_k u_k \otimes u_k$ solves this optimization program and $\{u_k\}$ is an optimal factorization with respect to



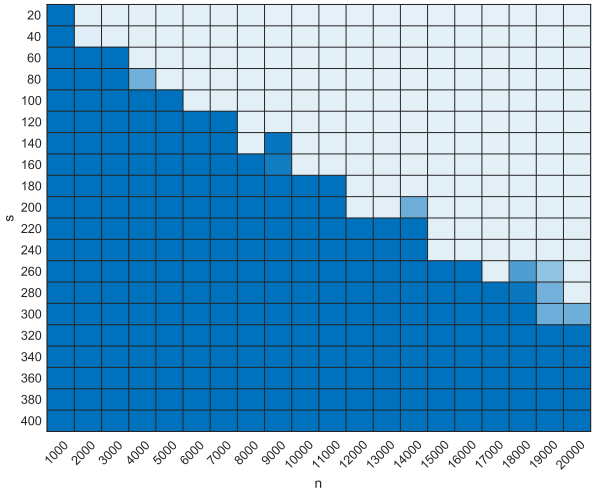
(a) Our algorithm



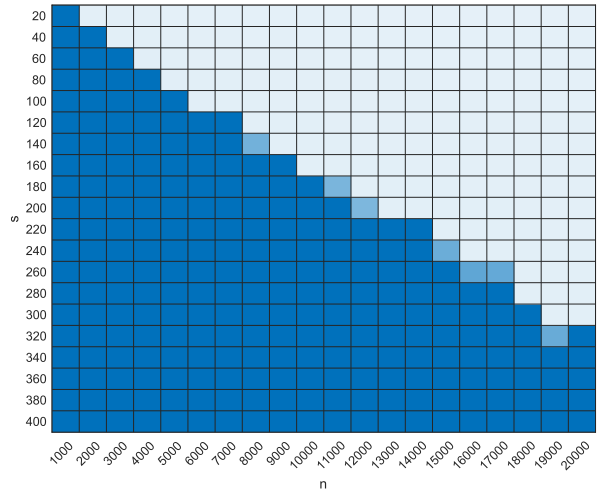
(b) SWF [9]



(c) GAMP [13]



(d) SPARTA [8]



(e) CoPRAM [11]

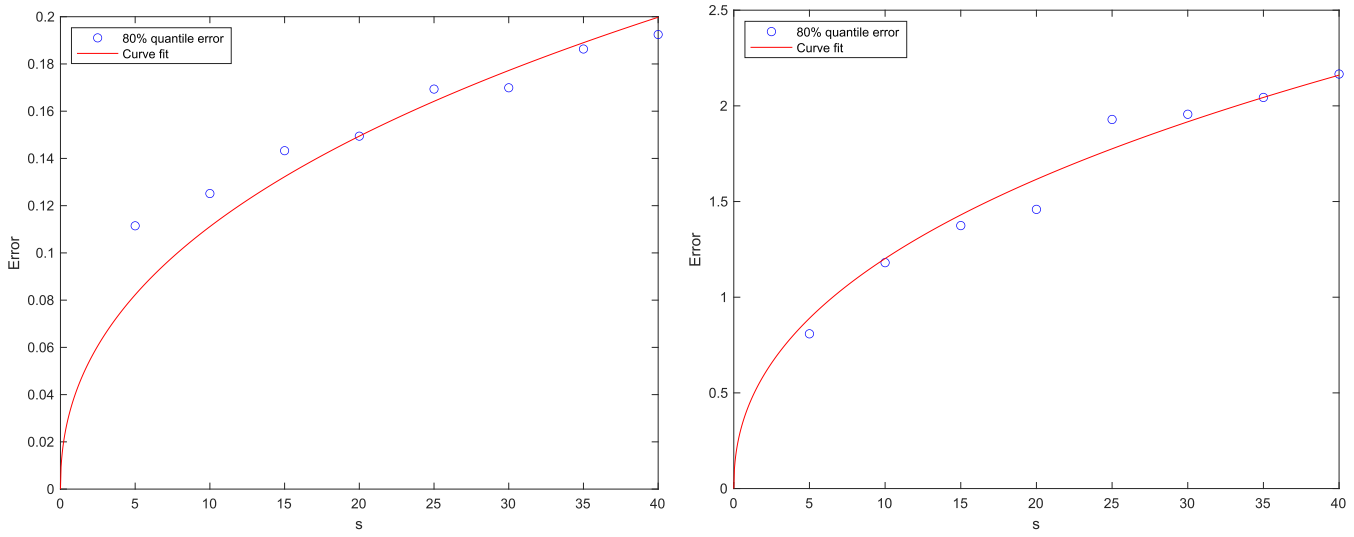
Fig. 1. Phase transition plots. Colors represent 80% quantile error over 20 trials (darker colors correspond to higher error). We used $p = 20{,}000$, $\|\beta^*\|_2 = 1$, and $\sigma = 0.05$. All algorithms were run on the same data.

Θ_s if and only if P is feasible and, for all $u \in \mathbf{R}^p$.

$$\langle \widehat{\sum} u, u\rangle + \left(\lambda \sum_k \theta_s(u_k, u_k) - \langle \widehat{\sum}, P\rangle_{\mathrm{HS}}\right)\|u\|_2^2 \le \theta_s(u,u).$$

B. A First Factored Algorithm, a Computational Snag, and a Heuristic

The results of the previous section give a simple abstract recipe for finding a global optimum of (6):



(a) Gaussian noise ($\|\beta^*\|_2 = 1$, $\sigma = 0.2$) (b) Poisson noise ($\|\beta^*\|_2 = 10$)

Fig. 2. Plot of $\|\hat{\beta} - \beta^*\|_2$ vs. s (80% quantile over 10 trials). All simulations use $p = 8{,}000$ and $n = 4{,}000$. Blue circles are actual data; the red curves are of the form $c\sqrt{s \log \frac{ep}{s}}$, where the scaling factor c is chosen to give minimum mean absolute deviation.

1) We optimize (8) over a fixed number r of rank-1 factors (i.e., vectors $u_1, \ldots, u_r, v_1, \ldots, v_r$) until we reach a first-order stationary point (by satisfying condition (a) in Corollary 1). Note that whenever condition (a) is not satisfied, it is easy to find a descent direction, since we can simply rescale the vectors u_k, v_k in a similar manner to the second part of the proof of Lemma 4.
2) At a first-order stationary point, if condition (b) in Corollary 1 holds, we have reached the global minimum. Otherwise, as in the first part of the proof of Lemma 4, there exists $\tilde{u}, \tilde{v} \in \mathbf{R}^p$ such that $\frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, B\rangle_{\mathrm{HS}})\langle X_i \tilde{u}, \tilde{v}\rangle > \lambda \theta_s(\tilde{u}, \tilde{v})$. We set $(u_{r+1}, v_{r+1}) = (\epsilon\tilde{u}, \epsilon\tilde{v})$ for $\epsilon > 0$ small enough to decrease the objective function and go back to step 1.

The algorithm is guaranteed to terminate with a finite r by [36, Theorem 2].

The most difficult part to implement is step 2. Checking condition (b) requires maximizing a bilinear form on vectors u, v under a bound on $\theta_s(u, v)$. If we could maximize this for general bilinear forms, we could also solve sparse PCA (see Section V-A), so we suspect it is not possible. However, this does not preclude positive results that exploit the particular structure of the phase retrieval problem.

To implement a practical algorithm, we take a very simple shortcut: instead of checking condition (b) over *all* vectors $u, v \in \mathbf{R}^p$, we check it over *1-sparse* vectors. We simply calculate whether any element of $\frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, B\rangle_{\mathrm{HS}}) X_i$ is greater than $(1 + 1/\sqrt{s})^2 \lambda$. Although we have not yet found a robust theoretical justification, we will see in the next section that this heuristic works reasonably well in practice. We summarize our high-level practical algorithm in Algorithm 1.

Algorithm 1 High-Level Sparse Phase Retrieval Algorithm

1: $r \leftarrow 1$
2: Initialize u_1, v_1 (e.g., some spectral algorithm)
3: **while** not Converged **do**
4: Optimize (8) over $\{u_1, \ldots u_r\}, \{v_1, \ldots, v_r\}$ with first-order method until condition (a) in Corollary 1 is satisfied
5: $Z \leftarrow \frac{1}{n}\sum_{i=1}^n (y_i - \langle X_i, B\rangle_{\mathrm{HS}}) X_i$, where $B = \sum_{k=1}^r u_k \otimes v_k$
6: **if** $Z_{ij} > (1 + 1/\sqrt{s})^2 \lambda$ for any $i, j \in \{1, \ldots, p\}$ **then**
7: $r \leftarrow r + 1$
8: $u_{r+1} \leftarrow \epsilon e_j, v_{r+1} \leftarrow \epsilon e_i$, where $\epsilon > 0$ is sufficiently small to decrease objective function.
9: **else**
10: Converged $\leftarrow$ true
11: **end if**
12: **end while**
13: **return** $\{u_1, \ldots, u_r\}, \{v_1, \ldots, v_r\}$

C. Simulation Results

We implemented Algorithm 1 in MATLAB and ran a variety of simulations to illustrate its performance with respect to both sample complexity and noise performance. The interested reader can view our code[6] to see more details, but some of the more salient features are the following:

- Line 5 of Algorithm 1 is implemented with alternating minimization over $U = [u_1 \cdots u_r] \in \mathbf{R}^{p \times r}$ and $V = [v_1 \cdots v_r] \in \mathbf{R}^{p \times r}$.
- After each alternating minimization step, we "rebalance" U and V (i.e., rescale each u_k, v_k to force $\theta_s(u_k, u_k) = \theta_s(u_k, v_k) = \theta_s(v_k, v_k)$).

[6]https://github.com/admcrae/spr2021

- Each minimization problem over U or V is convex, and we solve it with an accelerated proximal gradient descent algorithm.
- The proximal step requires solving a convex problem of the form

$$\arg\min_{y\in\mathbf{R}^p} \langle x, y\rangle + \frac{1}{2}\|y\|_2^2 + a\|y\|_2 + b\|y\|_1$$

for arbitrary $x \in \mathbf{R}^p$ and $a, b > 0$. This can be solved in closed form by soft-thresholding x with threshold b and then rescaling.

All of our simulations used i.i.d. Gaussian measurement vectors $x \sim \mathcal{N}(0, I_p)$.

1) Figure 1 shows phase transition diagrams of performance versus sample size n and sparsity s for our algorithm and a variety of alternatives. Note that qualitatively, all these algorithms have similar performance in terms of sample complexity. Interestingly, all of them appear only to require (within a log factor) a number of samples *linear* in the sparsity s. This demonstrates a gap between the empirical performance of all these algorithms and the best theoretical guarantees that have been proved so far.
2) Figure 2 shows plots of the error versus sparsity s for both Gaussian noise and Poisson noise. Note that in both cases, the error roughly follows the predicted $\sqrt{s\log(p/s)}$ scaling.

VI. Conclusion

We have shown that estimators for sparse phase retrieval and sparse PCA obtained by solving a convex program ((6) for sparse phase retrieval and (7) for sparse PCA) with the abstract mixed atomic norm (5) as a regularizer satisfy optimal statistical guarantees in terms of sample complexity and error. For sparse phase retrieval, we have derived a practical heuristic algorithm whose performance matches that of existing state-of-the-art algorithms.

Our work suggests new methods for analyzing these problems (and others with similar sparse factored structure, such as sparse blind deconvolution). It also suggests interesting new research directions in sparse recovery and in optimization. For example, it would be very useful to study *why* our heuristic approach appears to work well for sparse phase retrieval as well as whether it is possible to do even better. A related problem is to prove that sparse phase retrieval has linear sample complexity with practical algorithms (or that it doesn't, along with why current empirical results seem to suggest otherwise). Similarly, the atomic matrix norm (along with other similar norms) invites further analysis, particularly in how well we can optimize it (where this may depend on the structure of the problem in which it is used). The interplay between statistical guarantees and computational complexity theory (e.g., in sparse PCA) may be very interesting here.

Appendix A
Detailed Analysis of Mixed Norm

In this section, we explore several important properties of the mixed norm $\|\cdot\|_{*,s}$.

First, we show that matrices with small mixed norm can be written as a convex combination of sparse rank-1 matrices.

Lemma 6: For any matrix A, we can write $A = \sum a_i u_i \otimes v_i$, where each u_i and v_i has unit ℓ_2 norm and is s-sparse, and $\sum|a_i| \le \|A\|_{*,s}$.

Consequently, for any matrix Z,

$$\sup_{\|A\|_{*,s}\le 1} \langle Z, A\rangle_{\mathrm{HS}} \le \sup_{\substack{\|u\|_2,\|v\|_2\le 1\\ \|u\|_0,\|v\|_0\le s}} \langle Zu, v\rangle.$$

Proof: The consequence follows from the first statement immediately by the fact that any unit-atomic-norm A is in the convex hull of rank-1 s-sparse atoms. We now prove the first statement of the Lemma.

Because $\|\cdot\|_{*,s}$ is defined as an atomic norm over rank-1 atoms, it suffices to prove the result for rank-1 A. Therefore, we will show that any rank-1 matrix $x \otimes y$ can be written as $x \otimes y = \sum u_i \otimes v_i$, where each u_i and v_i is s-sparse, and $\sum\|u_i\|_2\|v_i\|_2 \le \theta_s(x, y)$.

Indeed, a standard result from sparsity theory (see, e.g., Exercise 10.3.7 in [1]) says that any vector z can be written as $z = \sum z_i$, where each z_i is s-sparse, and $\sum\|z_i\|_2 \le \|z\|_2 + \frac{1}{\sqrt{s}}\|z\|_1$. Applying this to both x and y, we have

$$x \otimes y = \left(\sum_i x_i\right)\left(\sum_j y_j\right) = \sum_{i,j} x_i \otimes y_j,$$

where each x_i and y_j is s-sparse, and

$$\begin{aligned}\sum_{i,j}\|x_i\|_2\|y_j\|_2 &= \left(\sum_i\|x_i\|_2\right)\left(\sum_j\|y_j\|_2\right)\\ &\le \left(\|x\|_2 + \frac{\|x\|_1}{\sqrt{s}}\right)\left(\|y\|_2 + \frac{\|y\|_1}{\sqrt{s}}\right)\\ &= \theta_s(x, y).\end{aligned}$$

□

To prove Lemma 1, we need to find a suitable subgradient of $\|\cdot\|_{*,s}$ at the point $B = \beta \otimes \beta$. Let $I \subset \{1, \dots, p\}$ denote the indices for which the entries of β are nonzero. With some abuse of notation, we also write I as the subspace of $\mathbf{R}^{p\times p}$ whose matrices are zero except at entries $(i, j) \in I \times I$. We also denote $T = \{x \otimes \beta + \beta \otimes y : x, y \in \mathbf{R}^p\}$. We will denote the orthogonal projections onto these subspaces and various orthogonal complements and intersections by $\mathcal{P}_I$, $\mathcal{P}_T$, $\mathcal{P}_{T\cap I^\perp}$, etc. We will also on occasion denote the orthogonal projection onto $\mathrm{span}\{\beta\} \subset \mathbf{R}^p$ or its orthogonal complement (in I) by $\mathcal{P}_\beta$, $\mathcal{P}_{\beta^\perp}$, $\mathcal{P}_{\beta^\perp\cap I}$, etc.

According to [36, Proposition 1], a matrix $W \in \partial\|B\|_{*,s}$ if the following two properties hold:

1) $\langle W\beta, \beta\rangle = \theta_s(\beta, \beta)$, and
2) $\langle Wu, v\rangle \le \theta_s(u, v)$ for all $u, v \in \mathbf{R}^p$.

It is easy to check that the matrix $W_\beta := w_\beta \otimes w_\beta$, where $w_\beta := \frac{\beta}{\|\beta\|_2} + \frac{1}{\sqrt{s}}\operatorname{sign}\beta$, is a subgradient. However, as with the subgradients of the ordinary nuclear norm, a much broader set of matrices satisfies these properties:

Lemma 7: Suppose β is s-sparse, and let $B = \beta \otimes \beta$. Any matrix of the form $W = W_\beta + W^\perp \in \partial\|B\|_{*,s}$ where $W^\perp$

can be any matrix in one of the following three families (or any convex combination thereof):

1) $W^\perp = \frac{1}{\sqrt{s}}(w_\beta \otimes \tilde{u} + \tilde{v} \otimes w_\beta)$, where $\tilde{u}, \tilde{v} \in I^\perp$ and $\|\tilde{u}\|_\infty, \|\tilde{v}\|_\infty \le 1$.
2) $W^\perp \in T^\perp$ and $\|W\| \le 1$.
3) $W^\perp = \mathcal{P}_{T^\perp \cap I^\perp}(\widetilde{W})$ for $\widetilde{W}$ satisfying $\langle \widetilde{W}u, v\rangle \le \frac{1}{5}\theta_s(u,v)$ for all $u, v \in \mathbf{R}^p$.

Proof: For each case, note that $\langle W^\perp \beta, \beta\rangle = 0$, so we only need to show that $\langle Wu, v\rangle \le \theta_s(u,v)$ for all $u, v \in \mathbf{R}^p$.

We will use the following simple fact many times: for any vector $u \in \mathbf{R}^p$,

$$|\langle w_\beta, u\rangle| \le \|\mathcal{P}_\beta(u)\|_2 + \frac{\|\mathcal{P}_I(u)\|_1}{\sqrt{s}}.$$

We prove each case separately.

a) Case 1: Let $\tilde{u}, \tilde{v} \in I^\perp$ with $\|\tilde{u}\|_\infty, \|\tilde{v}\|_\infty \le 1$. Let

$$W = w_\beta \otimes w_\beta + \frac{1}{\sqrt{s}}(w_\beta \otimes \tilde{u} + \tilde{v} \otimes w_\beta).$$

Then, for any $u, v \in \mathbf{R}^p$,

$$\begin{aligned}
\langle Wu, v\rangle &= \langle w_\beta, u\rangle\langle w_\beta, v\rangle + \frac{1}{\sqrt{s}}(\langle w_\beta, v\rangle\langle \tilde{u}, u\rangle + \langle \tilde{v}, v\rangle\langle w_\beta, u\rangle) \\
&\le \left(\|\mathcal{P}_\beta(u)\|_2 + \frac{\|\mathcal{P}_I(u)\|_1}{\sqrt{s}}\right)\left(\|\mathcal{P}_\beta(v)\|_2 + \frac{\|\mathcal{P}_I(v)\|_1}{\sqrt{s}}\right) \\
&\quad + \left(\|\mathcal{P}_\beta(v)\|_2 + \frac{\|\mathcal{P}_I(v)\|_1}{\sqrt{s}}\right)\frac{\|\mathcal{P}_{I^\perp}(u)\|_1}{\sqrt{s}} \\
&\quad + \left(\|\mathcal{P}_\beta(u)\|_2 + \frac{\|\mathcal{P}_I(u)\|_1}{\sqrt{s}}\right)\frac{\|\mathcal{P}_{I^\perp}(v)\|_1}{\sqrt{s}} \\
&\le \left(\|\mathcal{P}_\beta(u)\|_2 + \frac{\|u\|_1}{\sqrt{s}}\right)\left(\|\mathcal{P}_\beta(v)\|_2 + \frac{\|v\|_1}{\sqrt{s}}\right) \\
&\le \theta_s(u,v),
\end{aligned}$$

where the penultimate inequality uses the fact that $\|z\|_1 = \|\mathcal{P}_I(z)\|_1 + \|\mathcal{P}_{I^\perp}(z)\|_1$ for any vector z.

b) Case 2: Let $W^\perp \in T^\perp$ such that $\|W\| \le 1$. Let $u, v \in \mathbf{R}^p$. Note that $\langle W^\perp u, v\rangle \le \|\mathcal{P}_{\beta^\perp}(u)\|_2\|\mathcal{P}_{\beta^\perp}(v)\|_2$. Then

$$\begin{aligned}
\langle Wu, v\rangle &= \langle w_\beta, u\rangle\langle w_\beta, v\rangle + \langle W^\perp u, v\rangle \\
&\le \left(\|\mathcal{P}_\beta(u)\|_2 + \frac{\|\mathcal{P}_I(u)\|_1}{\sqrt{s}}\right)\left(\|\mathcal{P}_\beta(v)\|_2 + \frac{\|\mathcal{P}_I(v)\|_1}{\sqrt{s}}\right) \\
&\quad + \|\mathcal{P}_{\beta^\perp}(u)\|_2\|\mathcal{P}_{\beta^\perp}(v)\|_2 \\
&\le \theta_s(u,v),
\end{aligned}$$

where the last inequality uses that fact that

$$\|\mathcal{P}_\beta(u)\|_2\|\mathcal{P}_\beta(v)\|_2 + \|\mathcal{P}_{\beta^\perp}(u)\|_2\|\mathcal{P}_{\beta^\perp}(v)\|_2 \le \|u\|_2\|v\|_2.$$

c) Case 3: Let $\widetilde{W} \in \mathbf{R}^{p\times p}$ satisfy $\langle \widetilde{W}u, v\rangle \le \frac{1}{5}\theta_s(u,v)$. Let $W = W_\beta + \mathcal{P}_{T^\perp \cap I^\perp}(\widetilde{W})$. Then, for $u, v \in \mathbf{R}^p$,

$$\begin{aligned}
\langle Wu, v\rangle &= \langle w_\beta, u\rangle\langle w_\beta, v\rangle + \langle \mathcal{P}_{T^\perp\cap I^\perp}(\widetilde{W}), v \otimes u\rangle_{\mathrm{HS}} \\
&= \langle w_\beta, u\rangle\langle w_\beta, v\rangle + \langle \widetilde{W}, \mathcal{P}_{T^\perp\cap I^\perp}(v \otimes u)\rangle_{\mathrm{HS}} \\
&= \langle w_\beta, u\rangle\langle w_\beta, v\rangle + \langle \widetilde{W}\mathcal{P}_{I^\perp}(u), \mathcal{P}_{\beta^\perp\cap I}(v)\rangle \\
&\quad + \langle \widetilde{W}\mathcal{P}_{\beta^\perp\cap I}(u), \mathcal{P}_{I^\perp}(v)\rangle \\
&\quad + \langle \widetilde{W}\mathcal{P}_{I^\perp}(u), \mathcal{P}_{I^\perp}(v)\rangle.
\end{aligned}$$

As before,

$$\begin{aligned}
\langle w_\beta, u\rangle\langle w_\beta, v\rangle &\le \left(\|\mathcal{P}_\beta(u)\|_2 + \frac{\|\mathcal{P}_I(u)\|_1}{\sqrt{s}}\right) \\
&\quad \times \left(\|\mathcal{P}_\beta(v)\|_2 + \frac{\|\mathcal{P}_I(v)\|_1}{\sqrt{s}}\right).
\end{aligned}$$

Next,

$$\begin{aligned}
&\langle \widetilde{W}\mathcal{P}_{I^\perp}(u), \mathcal{P}_{\beta^\perp\cap I}(v)\rangle \\
&\quad \le \frac{1}{5}\left(\|\mathcal{P}_{I^\perp}(u)\|_2 + \frac{\|\mathcal{P}_{I^\perp}(u)\|_1}{\sqrt{s}}\right) \\
&\qquad \times \left(\|\mathcal{P}_{\beta^\perp\cap I}(v)\|_2 + \frac{\|\mathcal{P}_{\beta^\perp\cap I}(v)\|_1}{\sqrt{s}}\right) \\
&\quad \le \frac{2}{5}\left(\|\mathcal{P}_{\beta^\perp}(u)\|_2 + \frac{\|\mathcal{P}_{I^\perp}(u)\|_1}{\sqrt{s}}\right)\|\mathcal{P}_{\beta^\perp}(v)\|_2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\langle \widetilde{W}\mathcal{P}_{\beta^\perp\cap I}(u), \mathcal{P}_{I^\perp}(v)\rangle \\
&\qquad \le \frac{2}{5}\|\mathcal{P}_{\beta^\perp}(u)\|_2\left(\|\mathcal{P}_{\beta^\perp}(v)\|_2 + \frac{\|\mathcal{P}_{I^\perp}(v)\|_1}{\sqrt{s}}\right).
\end{aligned}$$

Combining these with the obvious bound on $\langle \widetilde{W}\mathcal{P}_{I^\perp}(u), \mathcal{P}_{I^\perp}(v)\rangle$, we get

$$\begin{aligned}
\langle Wu, v\rangle &\le \left(\|\mathcal{P}_\beta(u)\|_2 + \frac{\|\mathcal{P}_I(u)\|_1}{\sqrt{s}}\right) \\
&\quad \times \left(\|\mathcal{P}_\beta(v)\|_2 + \frac{\|\mathcal{P}_I(v)\|_1}{\sqrt{s}}\right) \\
&\quad + \left(\|\mathcal{P}_{\beta^\perp}(u)\|_2 + \frac{\|\mathcal{P}_{I^\perp}(u)\|_1}{\sqrt{s}}\right) \\
&\quad \times \left(\|\mathcal{P}_{\beta^\perp}(v)\|_2 + \frac{\|\mathcal{P}_{I^\perp}(v)\|_1}{\sqrt{s}}\right)
\end{aligned}$$

To bound this last expression, we consider the terms that we get from multiplying everything out. Note again that

$$\|\mathcal{P}_\beta(u)\|_2\|\mathcal{P}_\beta(v)\|_2 + \|\mathcal{P}_{\beta^\perp}(u)\|_2\|\mathcal{P}_{\beta^\perp}(v)\|_2 \le \|u\|_2\|v\|_2,$$

and also

$$\|\mathcal{P}_I(u))\|_1\|\mathcal{P}_I(v)\|_1 + \|\mathcal{P}_{I^\perp}(u)\|_1\|\mathcal{P}_{I^\perp}(v)\|_1 \le \|u\|_1\|v\|_1.$$

For the cross-terms, note that

$$\begin{aligned}
&\|\mathcal{P}_\beta(u)\|_2\|\mathcal{P}_I(v)\|_1 + \|\mathcal{P}_{\beta^\perp}(u)\|_2\|\mathcal{P}_{I^\perp}(v)\|_1 \\
&\quad \le \min_{c>0}\left(c\frac{\|\mathcal{P}_\beta(u)\|_2^2 + \|\mathcal{P}_{\beta^\perp}(u)\|_2^2}{2}\right. \\
&\qquad \left. + \frac{1}{c}\frac{\|\mathcal{P}_I(v)\|_1^2 + \|\mathcal{P}_{I^\perp}(v)\|_1^2}{2s}\right) \\
&\quad \le \min_{c>0}\left(c\frac{\|u\|_2^2}{2} + \frac{1}{c}\frac{\|v\|_1^2}{2s}\right) \\
&\quad = \frac{1}{\sqrt{s}}\|u\|_2\|v\|_1.
\end{aligned}$$

The similar inequality holds for u and v reversed. Therefore,

$$\begin{aligned}
\langle Wu, v\rangle &\le \left(\|u\|_2 + \frac{\|u\|_1}{\sqrt{s}}\right)\left(\|v\|_2 + \frac{\|v\|_1}{\sqrt{s}}\right) \\
&= \theta_s(u,v).
\end{aligned}$$

□

With this, we can prove Lemma 1.

Proof of Lemma 1: Let $A \in \mathbf{R}^{p\times p}$. We choose a subgradient $W \in \partial\|B\|_{*,s}$ as follows: Let

$$W = W_\beta + \frac{1}{10}\big(W_1^\perp + 4\ W_2^\perp + 5\ W_3^\perp\big),$$

where we choose $W_i^\perp$, $i = 1, 2, 3$, as follows:

1) If $\mathcal{P}_{T\cap I^\perp}(A) = \beta \otimes u + v \otimes \beta$ where $u, v \in I^\perp$, choose

$$W_1^\perp = \frac{1}{\sqrt{s}}(w_\beta \otimes \tilde{u} + \tilde{v} \otimes w_\beta),$$

where $\tilde{u}, \tilde{v} \in I^\perp$, $\|\tilde{u}\|_\infty, \|\tilde{v}\|_\infty \le 1$ and $\langle \tilde{u}, u\rangle = \|u\|_1$, $\langle \tilde{v}, v\rangle = \|v\|_1$. Then

$$\begin{aligned}\langle W_1^\perp u, v\rangle &= \left(\|\beta\|_2 + \frac{\|\beta\|_1}{\sqrt{s}}\right)\frac{\|u\|_1 + \|v\|_1}{\sqrt{s}}\\ &\ge \theta_s(\beta, u) + \theta_s(\beta, v) - 2\|\beta\|_2(\|u\|_2 + \|v\|_2)\\ &\ge \|\mathcal{P}_{T\cap I^\perp}(A)\|_{*,s} - 2\sqrt{2}\|\mathcal{P}_{T\cap I^\perp}(A)\|_F\\ &\ge \|\mathcal{P}_{T\cap I^\perp}(A)\|_{*,s} - 2\sqrt{2}\|A\|_F.\end{aligned}$$

2) Choose $W_2^\perp \in T^\perp \cap I$ with $\|W_2^\perp\| \le 1$ such that $\langle W_2^\perp, A\rangle_{\mathrm{HS}} = \|\mathcal{P}_{T^\perp\cap I}(A)\|_* \ge \frac{1}{4}\|\mathcal{P}_{T^\perp\cap I}(A)\|_{*,s}$. This last norm inequality holds because every vector in I is s-sparse.
3) Choose $W_3^\perp$ according to Lemma 7 such that $\langle W_3^\perp, A\rangle_{\mathrm{HS}} = \frac{1}{5}\|\mathcal{P}_{T^\perp\cap I^\perp}(A)\|_{*,s}$.

Then, using the fact that $\|W_\beta\|_F = \|w_\beta\|_2^2 \le 4$, we have

$$\begin{aligned}\langle W, A\rangle_{\mathrm{HS}} &= \langle W_\beta, A\rangle_{\mathrm{HS}} + \frac{1}{10}\langle W_1^\perp, A\rangle_{\mathrm{HS}}\\ &\quad + \frac{4}{10}\langle W_2^\perp, A\rangle_{\mathrm{HS}} + \frac{5}{10}\langle W_3^\perp, A\rangle_{\mathrm{HS}}\\ &\ge -4\|A\|_F - \frac{1}{10}\|\mathcal{P}_{T\cap I}(A)\|_{*,s} + \frac{1}{10}\|\mathcal{P}_{T\cap I}(A)\|_{*,s}\\ &\quad + \frac{1}{10}\Big(\|\mathcal{P}_{T\cap I^\perp}(A)\|_{*,s} - 2\sqrt{2}\|A\|_F\Big)\\ &\quad + \frac{4}{10}\cdot\frac{1}{4}\|\mathcal{P}_{T^\perp\cap I}(A)\|_{*,s} + \frac{5}{10}\cdot\frac{1}{5}\|\mathcal{P}_{T^\perp\cap I^\perp}(A)\|_{*,s}\\ &\ge \frac{1}{10}\|A\|_{*,s} - \left(4 + \frac{\sqrt{2}}{5}\right)\|A\|_F - \frac{1}{10}\|\mathcal{P}_{T\cap I}(A)\|_{*,s}\\ &\ge \frac{1}{10}\|A\|_{*,s} - 5\|A\|_F,\end{aligned}$$

where the last inequality uses the fact that $\|\mathcal{P}_{T\cap I}(A)\|_{*,s} \le 4\|\mathcal{P}_{T\cap I}(A)\|_* \le 4\sqrt{2}\|A\|_F$. □

Appendix B
Empirical Process and Restricted Lower Isometry Bounds

Proof of Lemma 2: By Lemma 6, it suffices to show

$$\sup_{\substack{\|u\|_2=\|v\|_2=1\\ \|u\|_0,\|v\|_0\le s}} \langle Zu, v\rangle \lesssim \sigma\sqrt{\frac{s\log(ep/s)}{n}} + \frac{M}{n^{1-c}}\left(s\log\frac{ep}{s}\right)^{\eta+1}$$

where, again, $Z = \frac{1}{n}\sum_i G_i$.

We first consider the random variable $\langle Zu, v\rangle$ for fixed unit-norm u and v. We have

$$\langle Zu, v\rangle = \frac{1}{n}\sum_{i=1}^n \langle G_i u, v\rangle.$$

This is the sum of independent copies of the zero-mean random variable $\langle Gu, v\rangle$. By assumption,

$$\mathbf{E}\langle Gu, v\rangle^2 \le \sigma^2$$

and, for $\alpha \ge 3$,

$$\|\langle Gu, v\rangle\|_\alpha \le M\alpha^{\eta+1}.$$

Then, by [40, Theorem 3.1], for any $\delta > 0$, with probability at least $1 - \delta$,

$$\langle Zu, v\rangle \lesssim \sigma\sqrt{\frac{\log\delta^{-1}}{n}} + \frac{M\alpha^{\eta+1}}{n^{1-1/\alpha}}\delta^{-1/\alpha}.$$

We then use a covering argument similar to that in [41]. Let J_1 and J_2 be any two subspaces of s-sparse vectors in $\mathbf{R}^p$. The unit sphere S_{J_i} in J_i can be covered within a resolution of $1/4$ by at most 9^s points ([1, Corollary 4.2.13], for example). Let $\mathcal{N}_{J_1}, \mathcal{N}_{J_2}$ be optimal $1/4$-covering sets. For each $x \in S_{J_i}$, let $n_i(x)$ be the closest point in $\mathcal{N}_{J_i}$. Then

$$\begin{aligned}\sup_{\substack{u\in S_{J_1}\\ v\in S_{J_2}}} \langle Zu, v\rangle &= \sup_{\substack{u\in S_{J_1}\\ v\in S_{J_2}}} \langle Zn_1(u), n_2(v)\rangle + \langle Z(u - n_1(u)), v\rangle\\ &\quad + \langle Zn_1(u), v - n_2(v)\rangle\\ &\le \max_{\substack{u\in\mathcal{N}_{J_1}\\ v\in\mathcal{N}_{J_2}}} \langle Zu, v\rangle + \frac{1}{2}\sup_{\substack{u\in S_{J_1}\\ v\in S_{J_2}}} \langle Zu, v\rangle,\end{aligned}$$

so

$$\sup_{\substack{u\in S_{J_1}\\ v\in S_{J_2}}} \langle Zu, v\rangle \le 2\max_{\substack{u\in\mathcal{N}_{J_1}\\ v\in\mathcal{N}_{J_2}}} \langle Zu, v\rangle.$$

Let

$$\mathcal{N} = \bigcup_{s\text{-sparse } J_1, J_2} \mathcal{N}_{J_1} \times \mathcal{N}_{J_2}.$$

Clearly,

$$\begin{aligned}\sup_{\substack{\|u\|_2=\|v\|_2=1\\ \|u\|_0,\|v\|_0\le s}} \langle Zu, v\rangle &= \sup_{s\text{-sparse } J_1, J_2}\ \sup_{\substack{u\in S_{J_1}\\ v\in S_{J_2}}} \langle Zu, v\rangle\\ &\le 2\max_{(u,v)\in\mathcal{N}} \langle Zu, v\rangle.\end{aligned}$$

There are $\binom{p}{s} \le \left(\frac{ep}{s}\right)^s$ s-sparse subspaces of $\mathbf{R}^p$, so $|\mathcal{N}| \le \left(9^s\left(\frac{ep}{s}\right)^s\right)^2$.

By a union bound and substituting δ above with $\delta/|\mathcal{N}|$, we then have, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\begin{aligned}\sup_{\substack{\|u\|_2=\|v\|_2=1\\ \|u\|_0,\|v\|_0\le s}} \langle Zu, v\rangle &\lesssim \sigma\sqrt{\frac{s\log(ep/s)}{n} + \frac{\log\delta^{-1}}{n}}\\ &\quad + \frac{M\alpha^{\eta+1}}{n^{1-1/\alpha}}\left(\frac{ep}{s}\right)^{2s/\alpha}\delta^{-1/\alpha}.\end{aligned}$$

Taking $\delta = e^{-s}(s/p)^s$ and $\alpha \approx s\log\frac{Cp}{s}$, we get, with probability at least $1 - e^{-s}(s/p)^s$,

$$\sup_{\substack{\|u\|_2=\|v\|_2=1\\ \|u\|_0,\|v\|_0\le s}} \langle Zu, v\rangle \lesssim \sigma\sqrt{\frac{s\log(ep/s)}{n}} + \frac{M}{n^{1-c}}\left(s\log\frac{ep}{s}\right)^{\eta+1}.$$

□

We will need the following variant of Lemma 2 for both the sparse PCA results and our restricted lower isometry lemma:

Lemma 8: Let $G_1, \dots, G_n$ be i.i.d. copies of a random matrix $G \in \mathbf{R}^{p\times p}$, where, for all $u, v \in \mathbf{R}^p$, $\langle Gu, v\rangle$ has zero mean,

$$\mathbf{E}\langle Gu, v\rangle^2 \lesssim \|u\|_2^2\|v\|_2^2$$

and $\langle Gu, v\rangle$ is sub-exponential in the sense that $\|\langle Gu, v\rangle\|_\alpha \lesssim \alpha\|u\|_2\|v\|_2$ for all $\alpha \geq 2$.

Let

$$Z = \frac{1}{n}\sum_{i=1}^n G_i$$

For any integer $s \geq 1$, with probability at least $1 - e^{-s}(s/p)^s$,

$$\sup_{\|A\|_{*,s}\leq 1} \langle Z, A\rangle_{\mathrm{HS}} \lesssim \sqrt{\frac{s\log(ep/s)}{n}} + \frac{s\log(ep/s)}{n}.$$

Furthermore, for $n \gtrsim s\log\frac{ep}{s}$,

$$\mathbf{E}\sup_{\|A\|_{*,s}\leq 1} \langle Z, A\rangle_{\mathrm{HS}} \lesssim \sqrt{\frac{s\log(ep/s)}{n}}.$$

We omit the proof, as it is nearly identical to the proof of Lemma 2. We simply replace the Fuk-Nagaev inequality with a Bernstein inequality. With this, we can prove our restricted lower isometry lemma:

Proof of Lemma 3: If $X = x \otimes x$, by a straightforward calculation, for any $p \times p$ matrix A,

$$\begin{aligned}\mathbf{E}\langle X, A\rangle_{\mathrm{HS}}^2 = &\sum_{i\neq j} A_{ii}A_{jj}\,\mathbf{E}(x^{(i)})^2(x^{(j)})^2 \\ &+ 2\sum_{i\neq j} A_{ij}^2\,\mathbf{E}(x^{(i)})^2(x^{(j)})^2 \\ &+ \sum_i A_{ii}^2\,\mathbf{E}(x^{(i)})^4.\end{aligned}$$

Using the facts that $\mathbf{E}(x^{(i)})^2 = 1$ for each i and $x^{(i)}$ and $x^{(j)}$ are independent when $i \neq j$, we have

$$\begin{aligned}\mathbf{E}\langle X, A\rangle_{\mathrm{HS}}^2 &= \sum_{i,j} A_{ii}A_{jj} + 2\sum_{i\neq j} A_{ij}^2 + \sum_i A_{ii}^2(\mathbf{E}(x^{(i)})^4 - 1) \\ &\geq (\operatorname{tr} A)^2 + \min\{2, \mathbf{E}(x^1)^4 - 1\}\|A\|_F^2 \\ &\gtrsim \|A\|_F^2.\end{aligned}$$

The last inequality uses the assumption that $\mathbf{E}(x^{(1)})^4 > 1$.

By the Hanson-Wright inequality for sub-Gaussian vectors [42], we have

$$\mathbf{E}(\langle X, A\rangle_{\mathrm{HS}}^2 - \mathbf{E}\langle X, A\rangle_{\mathrm{HS}}^2)^2 \lesssim \|A\|_F^4,$$

so $\mathbf{E}\langle X, A\rangle_{\mathrm{HS}}^4 \lesssim (\mathbf{E}\langle X, A\rangle_{\mathrm{HS}}^2)^2$. By the Paley-Zygmund inequality, we then have, for some $c_1, c_2 > 0$,

$$\inf_{A\in\mathbf{R}^{p\times p}} \mathrm{P}(|\langle X, A\rangle_{\mathrm{HS}}| \geq c_1\|A\|_F) \geq c_2.$$

The remainder of the proof is a small-ball argument ([43]; see also [44] for an excellent introduction).

Let

$$S = \{A \in \mathbf{R}^{p\times p} : \|A\|_F = 1;\ \|A\|_{*,s} \leq C\}.$$

We will prove that

$$\inf_{A\in S} \frac{1}{n}\sum_{i=1}^n \langle X_i, A\rangle_{\mathrm{HS}}^2 \geq c$$

with high probability for some constant $c > 0$.

By [44, Proposition 5.1], for any $t > 0$, we have, with probability at least $1 - e^{-t^2/2}$,

$$\begin{aligned}&\inf_{A\in S}\sqrt{\frac{1}{n}\sum_{i=1}^n \langle X_i, A\rangle_{\mathrm{HS}}^2} \\ &\qquad\gtrsim c_1c_2 - 2\,\mathbf{E}\sup_{A\in S}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle X_i, A\rangle_{\mathrm{HS}}\right) - \frac{1}{\sqrt{n}}c_1\ t,\end{aligned}$$

where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. Rademacher random variables independent of everything else.

Set $Z = \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i$, and note that $G_i = \varepsilon_i X_i$, $i = 1, \dots, n$, satisfy the requirements of Lemma 8. Then

$$\mathbf{E}\sup_{A\in S}\langle Z, A\rangle_{\mathrm{HS}} \lesssim C\sqrt{\frac{s\log(ep/s)}{n}}.$$

Choosing n large enough and $t = \sqrt{2bn}$ for small enough $b > 0$ completes the proof.

□

Appendix C
Proof of Sparse PCA Error Bound

Proof of Theorem 2: By a similar argument to that in the proof of Theorem 1 in Section IV-A, the solution $\widehat{P}$ to (7) satisfies

$$\langle\widehat{\sum}, -H\rangle_{\mathrm{HS}} \leq \lambda\langle W, -H\rangle_{\mathrm{HS}}$$

for $H = \widehat{P} - P_1$ and any $W \in \partial\|P_1\|_{*,s}$. Choosing W according to Lemma 1 (as in the proof of Theorem 1), we obtain

$$\langle\widehat{\sum}, H\rangle_{\mathrm{HS}} \geq \lambda\left(\frac{1}{10}\|H\|_{*,s} - 5\|H\|_F\right).$$

We first consider the difference between $\langle\widehat{\sum}, H\rangle_{\mathrm{HS}}$ and $\langle\Sigma, H\rangle_{\mathrm{HS}}$. Since the distribution of $\widehat{\sum}$ is independent of μ, we assume, without loss of generality, that $\mu = 0$. We write $x_i = \Sigma^{1/2}z_i$, where $z_i \sim \mathcal{N}(0, I_p)$, and $\Sigma^{1/2} = \sqrt{\sigma_1}P_1 + \Sigma_2^{1/2}$. We therefore want to bound

$$\langle\widehat{\sum} - \Sigma, H\rangle_{\mathrm{HS}} = \langle\Sigma^{1/2}(Z - I_p - \bar{z}\otimes\bar{z})\Sigma^{1/2}, H\rangle_{\mathrm{HS}},$$

where $Z = \frac{1}{n}\sum_{i=1}^n z_i \otimes z_i$ and $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$.

Let $H^\perp$ denote the component of H orthogonal (in Hilbert-Schmidt inner product) to P_1. We have

$$H = \langle H, P_1\rangle_{\mathrm{HS}}P_1 + H^\perp.$$

First, for all $t \le n$, with probability at least $1 - e^{-t}$,

$$\begin{aligned}\left|\langle\widehat{\sum} - \Sigma, P_1\rangle_{\mathrm{HS}}\right| &= \sigma_1\left|\frac{1}{n}\sum_{i=1}^n(\langle z_i, v_1\rangle^2 - 1) - \langle\bar{z}, v_1\rangle^2\right| \\ &\le \sigma_1\left|\frac{1}{n}\sum_{i=1}^n(\langle z_i, v_1\rangle^2 - 1)\right| + \langle\bar{z}, v_1\rangle^2 \\ &\lesssim \sigma_1\left(\sqrt{\frac{t}{n}} + \frac{t}{n}\right) \\ &\lesssim \sigma_1\sqrt{\frac{t}{n}},\end{aligned}$$

where the second-to-last inequality follows from applying a Bernstein inequality to the sum and an ordinary Gaussian tail bound to the $\mathcal{N}(0, 1/n)$ random variable $\langle\bar{z}, v_1\rangle$.

To analyze the remainder, denote the portion of $\widehat{\sum}$ orthogonal to P_1 as

$$\begin{aligned}\widehat{\sum}^{\perp} &= \widehat{\sum} - \langle\widehat{\sum}, P_1\rangle_{\mathrm{HS}} P_1 \\ &= \frac{1}{n}\sum_{i=1}^n\Bigg(\sqrt{\sigma_1}\langle z_i, v_1\rangle(v_1 \otimes (\Sigma_2^{1/2}z_i) \\ &\qquad + (\Sigma_2^{1/2}z_i) \otimes v_1) + (\Sigma_2^{1/2}z_i)^{\otimes 2}\Bigg) \\ &\qquad - (\Sigma_2^{1/2}\bar{z})^{\otimes 2}.\end{aligned}$$

Note that for each i, $\langle v_1, z_i\rangle$ is independent of $\Sigma_2^{1/2}z_i$. By Lemma 8, with probability at least $1 - 2\ e^{-s}(s/p)^s$,

$$\begin{aligned}&\sup_{\|A\|_{*,s}\le 1}\langle\widehat{\sum}^{\perp} + (\Sigma_2^{1/2}\bar{z})^{\otimes 2} - \Sigma_2, A\rangle_{\mathrm{HS}} \\ &\quad\le 2\sup_{\|A\|_{*,s}\le 1}\left\langle\frac{1}{n}\sum_{i=1}^n\sqrt{\sigma_1}\langle z_i, v_1\rangle v_1 \otimes (\Sigma_2^{1/2}z_i), A\right\rangle_{\mathrm{HS}} \\ &\qquad + \sup_{\|A\|_{*,s}\le 1}\left\langle\frac{1}{n}\sum_{i=1}^n(\Sigma_2^{1/2}z_i)^{\otimes 2} - \Sigma_2, A\right\rangle_{\mathrm{HS}} \\ &\quad\lesssim (\sqrt{\sigma_1\sigma_2} + \sigma_2)\left(\sqrt{\frac{s\log(ep/s)}{n}} + \frac{s\log(ep/s)}{n}\right) \\ &\quad\lesssim \sqrt{\sigma_1\sigma_2}\sqrt{\frac{s\log(ep/s)}{n}}.\end{aligned}$$

Lemma 8 also gives, with probability at least $1 - e^{-s}(s/p)^s$,

$$\begin{aligned}&\sup_{\|A\|_{*,s}\le 1}\langle(\Sigma_2^{1/2}\bar{z})^{\otimes 2}, A\rangle_{\mathrm{HS}} \\ &\quad\le \sup_{\|A\|_{*,s}\le 1}\langle(\Sigma_2^{1/2}\bar{z})^{\otimes 2} - \mathbf{E}(\Sigma_2^{1/2}\bar{z})^{\otimes 2}, A\rangle_{\mathrm{HS}} \\ &\qquad + \sup_{\|A\|_{*,s}\le 1}\langle\mathbf{E}(\Sigma_2^{1/2}\bar{z})^{\otimes 2}, A\rangle_{\mathrm{HS}} \\ &\quad\lesssim \sigma_2\sqrt{\frac{s\log(ep/s)}{n}}.\end{aligned}$$

Therefore,

$$\sup_{\|A\|_{*,s}\le 1}\langle\widehat{\sum}^{\perp} - \Sigma_2, A\rangle_{\mathrm{HS}} \lesssim \sqrt{\sigma_1\sigma_2}\sqrt{\frac{s\log(ep/s)}{n}}$$

with probability at least $1 - 3\ e^{-s}(s/p)^s$.

Let λ be chosen with a large enough constant to ensure that on this event,

$$\sup_{\|A\|_{*,s}\le 1}\langle\widehat{\sum}^{\perp} - \Sigma_2, A\rangle_{\mathrm{HS}} \le \frac{\lambda}{10}.$$

Then

$$|\langle\widehat{\sum}^{\perp} - \Sigma_2, H\rangle_{\mathrm{HS}}| \le \frac{\lambda}{10}\|H\|_{*,s}.$$

We then have

$$\begin{aligned}&\sigma_1\langle P_1, H\rangle_{\mathrm{HS}} + \langle\Sigma_2, H\rangle_{\mathrm{HS}} \\ &\quad= \langle\Sigma, H\rangle_{\mathrm{HS}} \\ &\quad= \langle\widehat{\sum}, H\rangle_{\mathrm{HS}} + \langle\Sigma - \widehat{\sum}, H\rangle_{\mathrm{HS}} \\ &\quad\ge \lambda\left(\frac{1}{10}\|H\|_{*,s} - 5\|H\|_F\right) - \sigma_1\sqrt{\frac{t}{n}}|\langle H, P_1\rangle_{\mathrm{HS}}| \\ &\qquad - \frac{\lambda}{10}\|H\|_{*,s} \\ &\quad= -5\lambda\|H\|_F - \sigma_1\sqrt{\frac{t}{n}}|\langle H, P_1\rangle_{\mathrm{HS}}|.\end{aligned}$$

Note that

$$\langle H, P_1\rangle_{\mathrm{HS}} = \langle\widehat{P}, P_1\rangle_{\mathrm{HS}} - 1 \le 0,$$

and $\langle\Sigma_2, H\rangle_{\mathrm{HS}} = \langle\Sigma_2, \widehat{P}\rangle_{\mathrm{HS}}$, so

$$\sigma_1\left(1 - \sqrt{\frac{t}{n}}\right)(\langle\widehat{P}, P_1\rangle_{\mathrm{HS}} - 1) + \langle\widehat{P}, \Sigma_2\rangle_{\mathrm{HS}} \gtrsim -\lambda\|H\|_F.$$

Note that $\langle\widehat{P}, \Sigma_2\rangle_{\mathrm{HS}} \le \sigma_2\|\mathcal{P}_{T^\perp}(\widehat{P})\|_*$, where $T^\perp$ is (similarly to before) the matrix subspace with rows and columns orthogonal to v_1. Note that $1 \ge \|\widehat{P}\|_* \ge \langle\widehat{P}, P_1\rangle_{\mathrm{HS}} + \|\mathcal{P}_{T^\perp}(\widehat{P})\|_*$, so $\langle\widehat{P}, \Sigma_2\rangle_{\mathrm{HS}} \le \sigma_2(1 - \langle\widehat{P}, P_1\rangle_{\mathrm{HS}})$.

Combining this with the previous inequality and requiring $n \gtrsim \left(\frac{\sigma_1}{\sigma_1 - \sigma_2}\right)^2 t$, we have

$$\begin{aligned}&(\sigma_1 - \sigma_2)(1 - \langle\widehat{P}, P_1\rangle_{\mathrm{HS}}) \\ &\quad\lesssim \left(\sigma_1\left(1 - \sqrt{\frac{t}{n}}\right) - \sigma_2\right)(1 - \langle\widehat{P}, P_1\rangle_{\mathrm{HS}}) \\ &\quad\lesssim \lambda\|H\|_F.\end{aligned}$$

To bound $\|H\|_F$, note that we can write

$$\widehat{P} = av_1 \otimes v_1 + v_1 \otimes u + w \otimes v_1 + \mathcal{P}_{T^\perp}(\widehat{P}),$$

where $a = \langle\widehat{P}, P_1\rangle_{\mathrm{HS}}$ and $u, w \perp v_1$. Then

$$1 \ge \|\widehat{P}\|_*^2 \ge \|\widehat{P}\|_F^2 = a^2 + \|u\|_2^2 + \|w\|_2^2 + \|\mathcal{P}_{T^\perp}(\widehat{P})\|_F^2,$$

and therefore

$$\begin{aligned}\|H\|_F^2 &= (1 - a)^2 + \|u\|_2^2 + \|w\|_2^2 + \|\mathcal{P}_{T^\perp}(\widehat{P})\|_F^2 \\ &\le (1 - a)^2 + 1 - a^2 \\ &= 2(1 - a) \\ &= 2(1 - \langle\widehat{P}, P_1\rangle_{\mathrm{HS}}).\end{aligned}$$

From this, we have $(\sigma_1 - \sigma_2)\|H\|_F^2 \lesssim \lambda\|H\|_F$, from which the result immediately follows. □

APPENDIX D
PROOF OF POISSON VARIANCE/MOMENT BOUNDS

If x satisfies Assumption 1 and, conditioned on x, $y \sim \text{Poisson}(\langle x, \beta^* \rangle^2)$, then, for unit-norm $u \in \mathbf{R}^p$,

$$\begin{aligned}\mathbf{E}\,\xi^2\langle x,u\rangle^4 &= \mathbf{E}\big[\mathbf{E}[\xi^2 \mid x]\langle x,u\rangle^4\big] \\ &= \mathbf{E}\langle x,\beta^*\rangle^2\langle x,u\rangle^4 \\ &\lesssim \|\beta^*\|_2^2.\end{aligned}$$

Also,

$$\begin{aligned}\|\xi\langle x,u\rangle^2\|_\alpha &= \Big(\mathbf{E}\big|\xi\langle x,u\rangle^2\big|^\alpha\Big)^{1/\alpha} \\ &= \big(\mathbf{E}\big[\mathbf{E}[|\xi|^\alpha \mid x]|\langle x,u\rangle|^{2\alpha}\big]\big)^{1/\alpha} \\ &\lesssim \sqrt{\alpha}\big(\mathbf{E}|\langle x,\beta^*\rangle|^\alpha|\langle x,u\rangle|^{2\alpha}\big)^{1/\alpha} + \alpha\|\langle x,u\rangle^2\|_\alpha \\ &\lesssim \alpha^2(\|\beta^*\|_2+1),\end{aligned}$$

where the first inequality uses the standard Poisson centered moment bound

$$\|Z - \mathbf{E}\,Z\|_\alpha \lesssim \sqrt{\alpha\lambda} + \alpha$$

if $Z \sim \text{Poisson}(\lambda)$.

REFERENCES

[1] R. Vershynin, *High-Dimensional Probability, An Introduction With Applications in Data Science*. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[2] E. J. Candès, T. Strohmer, and V. Voroninski, "PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming," *Commun. Pure Appl. Math.*, vol. 66, no. 8, pp. 1241–1274, Aug. 2013.

[3] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.

[4] E. J. Candès and X. Li, "Solving quadratic equations via phaselift when there are about as many equations as unknowns," *Found. Comput. Math.*, vol. 14, no. 5, pp. 1017–1026, Oct. 2014.

[5] C. Thrampoulidis and A. S. Rawat, "Lifting high-dimensional non-linear models with Gaussian regressors," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2019, pp. 3206–3215.

[6] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2013, pp. 2796–2804.

[7] T. Cai, X. Li, and Z. Ma, "Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow," *Ann. Statist.*, vol. 44, no. 5, pp. 2221–2251, 2016.

[8] G. Wang, L. Zhang, G. B. Giannakis, M. Akçakaya, and J. Chen, "Sparse phase retrieval via truncated amplitude flow," *IEEE Trans. Signal Process.*, vol. 66, no. 2, pp. 479–491, Jan. 2018.

[9] Z. Yuan, H. Wang, and Q. Wang, "Phase retrieval via sparse Wirtinger flow," *J. Comput. Appl. Math.*, vol. 355, pp. 162–173, Aug. 2019.

[10] Z. Yang, L. F. Yang, E. X. Fang, T. Zhao, Z. Wang, and M. Neykov, "Misspecified nonconvex statistical optimization for sparse phase retrieval," *Math. Program.*, vol. 176, nos. 1–2, pp. 545–571, Jul. 2019.

[11] G. Jagatap and C. Hegde, "Sample-efficient algorithms for recovering structured signals from magnitude-only measurements," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4434–4456, Jul. 2019.

[12] M. Soltanolkotabi, "Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization," *IEEE Trans. Inf. Theory*, vol. 65, no. 4, pp. 2374–2400, Apr. 2019.

[13] P. Schniter and S. Rangan, "Compressive phase retrieval via generalized approximate message passing," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 1043–1055, Feb. 2015.

[14] M. Bakhshizadeh, A. Maleki, and S. Jalali, "Using black-box compression algorithms for phase retrieval," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7978–8001, Dec. 2020.

[15] P. Hand and V. Voroninski, "Compressed sensing from phaseless Gaussian measurements via linear programming in the natural parameter space," 2016, *arXiv:1611.05985*.

[16] F. Salehi, E. Abbasi, and B. Hassibi, "Learning without the phase: Regularized PhaseMax achieves optimal sample complexity," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 8641–8652. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/b91f4f4d36fa98a94ac5584af95594a0-Abstract.html

[17] S. Bahmani and J. Romberg, "A flexible convex relaxation for phase retrieval," *Electron. J. Statist.*, vol. 11, no. 2, pp. 5254–5281, 2017.

[18] T. Goldstein and C. Studer, "PhaseMax: Convex phase retrieval via basis pursuit," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2675–2689, Apr. 2018.

[19] X. Li and V. Voroninski, "Sparse signal recovery from quadratic measurements via convex programming," *SIAM J. Math. Anal.*, vol. 45, no. 5, pp. 3019–3033, 2013.

[20] H. Ohlsson, A. Y. Yang, R. Dong, and S. S. Sastry, "Compressive phase retrieval from squared output measurements via semidefinite programming," in *Proc. IFAC Symp. Syst. Identificat.*, vol. 16, 2012, pp. 89–94.

[21] H. Ohlsson, A. Yang, R. Dong, and S. Sastry, "CPRL—An extension of compressive sensing to the phase retrieval problem," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2012, pp. 1367–1375. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/3a066bda8c96b9478bb0512f0a43028c-Abstract.html

[22] S. Bahmani and J. Romberg, "Efficient compressive phase retrieval with constrained sensing vectors," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 523–531.

[23] M. Iwen, A. Viswanathan, and Y. Wang, "Robust sparse phase retrieval made easy," *Appl. Comput. Harmon. Anal.*, vol. 42, no. 1, pp. 135–142, 2017.

[24] V. Koltchinskii and K. Lounici, "Concentration inequalities and moment bounds for sample covariance operators," *Bernoulli*, vol. 23, no. 1, pp. 110–133, 2017.

[25] V. Q. Vu and J. Lei, "Minimax rates of estimation for sparse PCA in high dimensions," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2012, pp. 1278–1286.

[26] H. Zou and L. Xue, "A selective overview of sparse principal component analysis," *Proc. IEEE*, vol. 106, no. 8, pp. 1311–1320, Aug. 2018.

[27] T. T. Cai, Z. Ma, and Y. Wu, "Sparse PCA: Optimal rates and adaptive estimation," *Ann. Statist.*, vol. 41, no. 6, pp. 3074–3110, 2013.

[28] A. Birnbaum, I. M. Johnstone, B. Nadler, and D. Paul, "Minimax bounds for sparse PCA with noisy high-dimensional data," *Ann. Statist.*, vol. 41, no. 3, pp. 1055–1084, 2013.

[29] Q. Berthet and P. Rigollet, "Complexity theoretic lower bounds for sparse principal component detection," in *Proc. Conf. Learn. Theory (COLT)*, 2013, pp. 1046–1066.

[30] T. Wang, Q. Berthet, and R. J. Samworth, "Statistical and computational trade-offs in estimation of sparse principal components," *Ann. Statist.*, vol. 44, no. 5, pp. 1896–1930, Oct. 2016.

[31] C. Gao, Z. Ma, and H. H. Zhou, "Sparse CCA: Adaptive estimation and computational barriers," *Ann. Statist.*, vol. 45, no. 5, pp. 2074–2101, Oct. 2017.

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

[33] M. Kliesch, S. J. Szarek, and P. Jung, "Simultaneous structures in convex signal recovery—Revisiting the convex combination of norms," *Frontiers Appl. Math. Statist.*, vol. 5, p. 23, May 2019.

[34] F. Diestel and J. Swart, "The metric theory of tensor products (Grothendieck's résumé revisited) Part 1: Tensor norms," *Quaestiones Mathematicae*, vol. 25, no. 1, pp. 37–72, Mar. 2002.

[35] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, pp. 805–849, Dec. 2012.

[36] B. D. Haeffele and R. Vidal, "Structured low-rank matrix factorization: Global optimality, algorithms, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 6, pp. 1468–1482, Jun. 2020.

[37] E. Richard, G. R. Obozinski, and J.-P. Vert, "Tight convex relaxations for sparse matrix factorization," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 27, 2014, pp. 3284–3292. [Online]. Available: https://papers.nips.cc/paper/5408-tight-convex-relaxations-for-sparse-matrix-factorization

[38] Y. Chen and E. J. Candès, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 739–747.
[39] T. Zhang, "On the dual formulation of regularized linear systems with convex risks," *Mach. Learn.*, vol. 46, no. 1, pp. 91–129, Jan. 2002.
[40] E. Rio, "About the constants in the Fuk–Nagaev inequalities," *Electron. Commun. Probab.*, vol. 22, pp. 1–12, Jan. 2017.
[41] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approx.*, vol. 28, no. 3, pp. 253–263, Dec. 2008.
[42] M. Rudelson and R. Vershynin, "Hanson-Wright inequality and sub-Gaussian concentration," *Electron. Commun. Probab.*, vol. 18, pp. 1–9, Jan. 2013.
[43] S. Mendelson, "Learning without concentration," *J. ACM*, vol. 62, no. 3, pp. 1–25, Jun. 2015.
[44] J. A. Tropp, "Convex recovery of a structured signal from independent random linear measurements," in *Sampling Theory, a Renaissance*, G. E. Pfander, Ed. Cham, Switzerland: Springer, 2015, pp. 67–101.

Andrew D. McRae (Member, IEEE) received the B.S. degree in electrical engineering, the B.S. degree in mathematics, the M.S. degree in electrical engineering, the M.S. degree in mathematics, and the Ph.D. degree in electrical engineering from the Georgia Institute of Technology in 2015, 2016, 2016, 2021, and 2022, respectively. He is currently a Post-Doctoral Researcher with the Institute of Mathematics, École Polytechnique Fédérale de Lausanne (EPFL). His research interests include learning theory and high-dimensional signal processing and statistics, with a particular focus on the role of optimization.

Justin Romberg (Fellow, IEEE) received the B.S.E.E., M.S., and Ph.D. degrees from Rice University, Houston, TX, USA, in 1997, 1999, and 2004, respectively. From 2004 to 2006, he was a Post-Doctoral Scholar in applied and computational mathematics at the California Institute of Technology. He is currently the Schlumberger Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, where he has been on the Faculty since 2006. His research interests include intersection of signal processing, machine learning, optimization, and applied probability. In 2009, he received the PECASE Award and a Packard Fellowship. In 2010, he was named a Rice University Outstanding Young Engineering Alumnus. In 2021, he received the IEEE Jack S. Kilby Signal Processing Medal.

Mark A. Davenport (Senior Member, IEEE) received the B.S.E.E., M.S., and Ph.D. degrees in electrical and computer engineering from Rice University in 2004, 2007, and 2010, respectively. He is currently a Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology. Prior to this, he spent 2010–2012 as an NSF Mathematical Sciences Post-Doctoral Research Fellow with the Department of Statistics, Stanford University, and as a Visitor with the Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie. His research interests include the fundamental role that low-dimensional models and optimization play in signal processing, statistical inference, and machine learning. He was a recipient of the National Science Foundation CAREER Award, the Air Force Office of Scientific Research Young Investigator Award, a Sloan Research Fellowship, and a Presidential Early Career Award for Scientists and Engineers (PECASE). He also serves as an Associate Editor for IEEE TRANSACTIONS ON INFORMATION THEORY and a Senior Area Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING.