FOUNDATIONS OF
COMPUTATIONAL
MATHEMATICS
The Journal of the Society for the Foundations of **Computational Mathematics**

CrossMark

Learning Semidefinite Regularizers

Yong Sheng Soh[1] · **Venkat Chandrasekaran**[1,2]

Abstract Regularization techniques are widely employed in optimization-based approaches for solving ill-posed inverse problems in data analysis and scientific computing. These methods are based on augmenting the objective with a penalty function, which is specified based on prior domain-specific expertise to induce a desired structure in the solution. We consider the problem of learning suitable regularization functions from data in settings in which precise domain knowledge is not directly available. Previous work under the title of 'dictionary learning' or 'sparse coding' may be viewed as learning a regularization function that can be computed via linear programming. We describe generalizations of these methods to learn regularizers that can be computed and optimized via semidefinite programming. Our framework for learning such semidefinite regularizers is based on obtaining structured factorizations of data matrices, and our algorithmic approach for computing these factorizations combines recent techniques for rank minimization problems along with an operator analog of Sinkhorn scaling. Under suitable conditions on the input data, our algorithm provides a locally linearly convergent method for identifying the correct regularizer that promotes the type of structure contained in the data. Our analysis is based on the stability properties of Operator Sinkhorn scaling and their relation to geometric aspects of determinantal

✉ Yong Sheng Soh
ysoh@caltech.edu

Venkat Chandrasekaran
venkatc@caltech.edu

[1] Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125, USA

[2] Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125, USA

🙋 Springer

varieties (in particular tangent spaces with respect to these varieties). The regularizers obtained using our framework can be employed effectively in semidefinite programming relaxations for solving inverse problems.

Keywords Atomic norm · Convex optimization · Low-rank matrices · Nuclear norm · Operator scaling · Representation learning

Mathematics Subject Classification Primary 90C25 · Secondary 90C22 · 15A24 · 41A45 · 52A41

1 Introduction

Regularization techniques are widely employed in the solution of inverse problems in data analysis and scientific computing due to their effectiveness in addressing difficulties due to ill-posedness. In their most common manifestation, these methods take the form of penalty functions added to the objective in optimization-based approaches for solving inverse problems. The purpose of the penalty function is to induce a desired structure in the solution, and these functions are specified based on prior domain-specific expertise. For example, regularization is useful for promoting smoothness, sparsity, low energy, and large entropy in solutions to inverse problems in image analysis, statistical model selection, and the geosciences [10,12,13,16,17,22,43,50,61]. In this paper, we study the question of *learning* suitable regularization functions from data in settings in which precise domain knowledge is not directly available. The regularizers obtained using our framework are specified as convex functions that can be computed efficiently via semidefinite programming, and therefore, they can be employed in tractable convex optimization approaches for solving inverse problems.

We begin our discussion by highlighting the geometric aspects of regularizers that make them effective in promoting a desired structure. In particular, we focus on a family of convex regularizers that are useful for inducing a general form of sparsity in solutions to inverse problems. Sparse data descriptions provide a powerful formalism for specifying low-dimensional structure in high-dimensional data, and they feature prominently in a range of problem domains. For example, natural images are often well approximated by a small number of wavelet coefficients, financial time series may be characterized by low-complexity factor models, and a small number of genetic markers may constitute a signature for disease. Concretely, suppose $\mathcal{A} \subset \mathbb{R}^d$ is a (possibly infinite) collection of elementary building blocks or atoms. Then, $\mathbf{y} \in \mathbb{R}^d$ is said to have a sparse representation using the atomic set $\mathcal{A}$ if $\mathbf{y}$ can be expressed as follows:

$$\mathbf{y} = \sum_{i=1}^{k} c_i \mathbf{a}_i, \quad \mathbf{a}_i \in \mathcal{A}, c_i \geq 0,$$

for a relatively small number k. As an illustration, if $\mathcal{A} = \{\pm \mathbf{e}^{(j)}\}_{j=1}^{d} \subset \mathbb{R}^d$ is the collection of signed standard basis vectors in $\mathbb{R}^d$, then concisely described objects with these atoms are those vectors in $\mathbb{R}^d$ consisting of a small number of nonzero coordinates. Similarly, if $\mathcal{A}$ is the set of rank-one matrices, then the corresponding sparsely

represented entities are low-rank matrices; see [16] for a more exhaustive collection of examples. An important virtue of sparse descriptions based on an atomic set $\mathcal{A}$ is that employing the *atomic norm* induced by $\mathcal{A}$—the gauge function of the atomic set $\mathcal{A}$—as a regularizer in inverse problems offers a natural convex optimization approach for obtaining solutions that have a sparse representation using $\mathcal{A}$ [16]. Continuing with the examples of vectors with few nonzero coordinates and of low-rank matrices, regularization with the ℓ_1 norm (the gauge function of the signed standard basis vectors) and with the matrix nuclear norm (the gauge function of the unit Euclidean norm rank-one matrices) are prominent techniques for promoting the corresponding sparse descriptions in solutions to inverse problems [12,13,17,22,26,43,50,61]. The reason for the effectiveness of atomic norm regularization is the favorable facial structure of the convex hull of $\mathcal{A}$, which has the feature that all its low-dimensional faces contain points that have a sparse description using $\mathcal{A}$. Indeed, in many contemporary data analysis applications the solutions of regularized optimization problems with generic input data tend to lie on low-dimensional faces of sublevel sets of the regularizer [14,22,50]. Based on this insight, atomic norm regularization has been shown to be effective in a range of tasks such as statistical denoising, model selection, and system identification [8,47,54].

The difficulty with employing an atomic norm regularizer in practice is that one requires prior domain knowledge of the atomic set $\mathcal{A}$—the extreme points of the atomic norm ball—that underlies a sparse description of the desired solution in an inverse problem. While such information may be available based on domain expertise in some problems (e.g., certain classes of signals having a sparse representation in a Fourier basis), identifying a suitable atomic set is challenging for many contemporary data sets that are high-dimensional and are typically presented to an analyst in an unstructured fashion. In this paper, we study the question of learning a suitable regularizer directly from observations $\{\mathbf{y}^{(j)}\}_{j=1}^n \subset \mathbb{R}^d$ of a collection of structured signals or models of interest. Specifically, as motivated by the preceding discussion, our objective is to identify a norm $\|\cdot\|$ in $\mathbb{R}^d$ such that each $\mathbf{y}^{(j)}/\|\mathbf{y}^{(j)}\|$ lies on a low-dimensional face of the unit ball of $\|\cdot\|$. An equivalent formulation of this question in terms of extreme points is that we want to obtain an atomic set $\mathcal{A}$ such that each $\mathbf{y}^{(j)}$ has a sparse representation using $\mathcal{A}$; the corresponding regularizer is simply the atomic norm induced by $\mathcal{A}$. A norm with these characteristics is adapted to the structure contained in the data $\{\mathbf{y}^{(j)}\}_{j=1}^n$, and it can be used subsequently as a regularizer in inverse problems to promote solutions with the same type of structure as in the collection $\{\mathbf{y}^{(j)}\}_{j=1}^n$.

When considered in full generality, our question is somewhat ill-posed for several reasons. First, if $\|\cdot\|$ is a norm that satisfies the properties described above with respect to the data $\{\mathbf{y}^{(j)}\}_{j=1}^n$, then so does $\alpha\|\cdot\|$ for any positive scalar α. This issue is addressed by learning a norm from a suitably scaled class of regularizers. A second source of difficulty is that the Euclidean norm $\|\cdot\|_{\ell_2}$ trivially satisfies our requirements for a regularizer as each $\mathbf{y}^{(j)}/\|\mathbf{y}^{(j)}\|_{\ell_2}$ is an extreme point of the Euclidean norm ball in $\mathbb{R}^d$; indeed, this is the regularizer employed in ridge regression. The atomic set in this case is the collection of all points with Euclidean norm equal to one, i.e., the dimension of this set is $d - 1$. However, data sets in many applications throughout science and engineering are well approximated as sparse combinations of elements of atomic sets of

much smaller dimension [7,10,16,21,37,46,49]. Identifying such lower-dimensional atomic sets is critical in inverse problems arising in high-dimensional data analysis in order to address the curse of dimensionality; in particular, as discussed in some of these preceding references, the benefits of atomic norm regularization in problems with large ambient dimension d are a consequence of measure concentration phenomena that crucially rely on the small dimensionality of the associated atomic set in comparison with d. We circumvent this second difficulty in learning a regularizer by considering atomic sets with appropriately bounded dimension. A third challenge with our question as it is stated is that the gauge function of the set $\{\pm \mathbf{y}^{(j)}/\|\mathbf{y}^{(j)}\|_{\ell_2}\}_{j=1}^{n}$ also satisfies the requirements for a suitable atomic norm as each $\mathbf{y}^{(j)}/\|\mathbf{y}^{(j)}\|_{\ell_2}$ is an extreme point of the unit ball of this regularizer. However, such a regularizer suffers from overfitting and does not generalize well as it is excessively tuned to the data set $\{\mathbf{y}^{(j)}\}_{j=1}^{n}$. Further, for large n this gauge function becomes intractable to characterize and it does not offer a computationally efficient approach for regularization. We overcome this complication by considering regularizers that have effectively parametrized sets of extreme points and consequently are tractable to compute.

The problem of learning a suitable polyhedral regularizer—an atomic norm with a unit ball that is a polytope—from data points $\{\mathbf{y}^{(j)}\}_{j=1}^{n}$ corresponds to identifying an appropriate *finite* atomic set to concisely describe each $\mathbf{y}^{(j)}$. This problem is equivalent to the question of 'dictionary learning' (also called 'sparse coding') on which there is a substantial amount of prior work [1–6,33,46,52,53,56,59,60,65] (see also the survey articles in [25,41]). To see this connection, suppose without loss of generality that we parametrize a finite atomic set via a matrix $L \in \mathbb{R}^{d \times p}$ so that the columns of L and their negations specify the atoms. The associated atomic norm ball is the image under L of the ℓ_1 ball in $\mathbb{R}^p$. The columns of L are typically scaled to have unit Euclidean norm to address the scaling issues mentioned previously (see Sect. 2.4). The number of columns p may be larger than d (i.e., the 'overcomplete' regime), and it controls the complexity of the atomic set as well as the computational tractability of describing the atomic norm. With this parametrization, learning a polyhedral regularizer to promote the type of structure contained in $\{\mathbf{y}^{(j)}\}_{j=1}^{n}$ may be viewed as obtaining a matrix L (given a target number of columns p) such that each $\mathbf{y}^{(j)}$ is well approximated as $L\mathbf{x}^{(j)}$ for a vector $\mathbf{x}^{(j)} \in \mathbb{R}^p$ with few nonzero coordinates. Computing such a representation of the data is precisely the objective in dictionary learning, although this problem is typically not phrased as a quest for a polyhedral regularizer in the literature. We remark further on some recent algorithmic developments in dictionary learning in Sects. 1.3.1 and 2.4, and we contrast these with the methods proposed in the present paper.

1.1 From Polyhedral to Semidefinite Regularizers

The objective of this paper is to investigate the problem of learning more general non-polyhedral atomic norm regularizers; in other words, the associated set of extreme points may be *infinite*. On the approximation-theoretic front, infinite atomic sets offer the possibility of concise descriptions of data sets with much richer types of structure than those with a sparse representation using finite atomic sets; in turn, the associated

regularizers could promote a broader class of structured solutions to inverse problems than polyhedral regularizers. On the computational front, many families of convex optimization problems beyond linear programs can be solved tractably and reliably [45]. However, building on the challenges outlined previously, there are two important factors in identifying non-polyhedral regularizers from data. First, it is crucial that any infinite atomic set $\mathcal{A}$ we consider has an effective parametrization so that it is tractable to characterize data that have a sparse representation using the elements of $\mathcal{A}$. Second, we require that the convex hull of the atomic set $\mathcal{A}$ has an efficient description so that the associated atomic norm provides a computationally tractable regularizer. As described next, we address these concerns by considering atomic sets that are efficiently parametrized as algebraic varieties (of a particular form) and that have convex hulls with tractable semidefinite descriptions. Thus, previous efforts in the dictionary learning literature on identifying finite atomic sets may be viewed as learning zero-dimensional ideals, whereas our approach corresponds to learning atomic sets that are larger-dimensional varieties. From a computational viewpoint, dictionary learning provides atomic norm regularizers that are computed via linear programming, while our framework leads to semidefinite programming regularizers. Consequently, although our framework is based on a much richer family of atomic sets in comparison with the finite sets considered in dictionary learning, we still retain efficiency of parametrization and computational tractability based on semidefinite representability.

Formally, we consider atomic sets in $\mathbb{R}^d$ that are images of rank-one matrices:

$$\mathcal{A}_q(\mathcal{L}) = \left\{ \mathcal{L}\left(\mathbf{u}\mathbf{v}'\right) \mid \mathbf{u}, \mathbf{v} \in \mathbb{R}^q, \ \|\mathbf{u}\|_{\ell_2} = 1, \ \|\mathbf{v}\|_{\ell_2} = 1 \right\}, \tag{1}$$

where $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ specifies a linear map. We focus on settings in which the dimension q is such that $q^2 > d$, so the atomic sets $\mathcal{A}_q(\mathcal{L})$ that we study in this paper are projections of rank-one matrices from a larger-dimensional space (in analogy to the overcomplete regime in dictionary learning). By construction, elements of $\mathbb{R}^d$ that have a sparse representation using the atomic set $\mathcal{A}_q(\mathcal{L})$ are those that can be specified as the image under $\mathcal{L}$ of *low-rank matrices* in $\mathbb{R}^{q \times q}$. As the convex hull of unit Euclidean norm rank-one matrices in $\mathbb{R}^{q \times q}$ is the nuclear norm ball in $\mathbb{R}^{q \times q}$, the corresponding atomic norm ball is given by:

$$\mathrm{conv}\left(\mathcal{A}_q(\mathcal{L})\right) = \left\{ \mathcal{L}(X) \mid X \in \mathbb{R}^{q \times q}, \ \|X\|_\star \leq 1 \right\}, \tag{2}$$

where $\|X\|_\star := \sum_i \sigma_i(X)$. As the nuclear norm ball has a tractable semidefinite description [26,50], the atomic norm induced by $\mathcal{A}_q(\mathcal{L})$ can be computed efficiently using semidefinite programming.

Given a collection of data points $\{\mathbf{y}^{(j)}\}_{j=1}^n \subset \mathbb{R}^d$ and a target dimension q, our goal is to find a linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ such that each $\mathbf{y}^{(j)}$, upon normalization by the gauge function of $\mathcal{A}_q(\mathcal{L})$, lies on a low-dimensional face of $\mathrm{conv}(\mathcal{A}_q(\mathcal{L}))$. For each $\mathbf{y}^{(j)}$ to have this property, it must have a sparse representation using the atomic set $\mathcal{A}_q(\mathcal{L})$; that is, there must exist a low-rank matrix $X^{(j)} \in \mathbb{R}^{q \times q}$ with $\mathbf{y}^{(j)} = \mathcal{L}(X^{(j)})$. The matrix $X^{(j)}$ provides a concise description of $\mathbf{y}^{(j)} \in \mathbb{R}^d$ in the

	Dictionary learning	Our work
Atomic set	$\{\pm L\mathbf{e}^{(i)} \mid \mathbf{e}^{(i)} \in \mathbb{R}^p$ is the i'th standard basis vector$\}$ $L : \mathbb{R}^p \to \mathbb{R}^d$ (linear map)	$\{\mathcal{L}(\mathbf{uv}') \mid \mathbf{u}, \mathbf{v} \in \mathbb{R}^q,$ $\|\mathbf{u}\|_{\ell_2} = \|\mathbf{v}\|_{\ell_2} = 1\}$ $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ (linear map)
Algebraic/geometric structure of atoms	Zero-dimensional ideal	Image of determinantal variety
Concisely specified data using atomic set	Image under L of sparse vectors	Image under $\mathcal{L}$ of low-rank matrices
Atomic norm ball	$\{L\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^p, \|\mathbf{x}\|_{\ell_1} \leq 1\}$	$\{\mathcal{L}(X) \mid X \in \mathbb{R}^{q \times q}, \|X\|_\star \leq 1\}$
Computing atomic norm regularizer	Linear programming	Semidefinite programming
Learning regularizer from data $\{\mathbf{y}^{(j)}\}_{j=1}^n$	Identify L and sparse $\mathbf{x}^{(j)} \in \mathbb{R}^p$ such that $\mathbf{y}^{(j)} \approx L\mathbf{x}^{(j)}$ for each j	Identify $\mathcal{L}$ and low-rank $X^{(j)} \in \mathbb{R}^{q \times q}$ such that $\mathbf{y}^{(j)} \approx \mathcal{L}(X^{(j)})$ for each j

Fig. 1 A comparison between prior work on dictionary learning and the present paper

higher-dimensional space $\mathbb{R}^{q \times q}$. Consequently, the problem of learning a semidefinite-representable regularizer with a unit ball that is a linear image of the nuclear norm ball may be phrased as one of *matrix factorization*. In particular, let $Y = [\mathbf{y}^{(1)} | \cdots | \mathbf{y}^{(n)}] \in \mathbb{R}^{d \times n}$ denote the data matrix, and let $\mathcal{L}_i \in \mathbb{R}^{q \times q}$, $i = 1, \ldots, d$ be the matrix that specifies the linear functional corresponding to the i'th component of a linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$. Then, our objective can be viewed as one of finding a collection of matrices $\{\mathcal{L}_i\}_{i=1}^d \subset \mathbb{R}^{q \times q}$ specifying linear functionals and a set of low-rank matrices $\{X^{(j)}\}_{j=1}^n \subset \mathbb{R}^{q \times q}$ specifying concise descriptions such that:

$$Y_{i,j} = \left\langle \mathcal{L}_i, X^{(j)} \right\rangle \quad i = 1, \ldots, d, \ j = 1, \ldots, n. \tag{3}$$

Here $\langle A, B \rangle = \text{trace}(A'B)$ denotes the trace inner product between matrices. Note the distinction with dictionary learning in which one seeks a factorization of the data matrix Y such that the $X^{(j)}$'s are sparse vectors as opposed to low-rank matrices as in our approach. Figure 1 summarizes the key differences between dictionary learning and the present paper.

1.2 An Alternating Update Algorithm for Matrix Factorization

A challenge with identifying a semidefinite regularizer by factoring a given data matrix as in (3) is that such a factorization is not unique. Specifically, consider any linear map $M : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$ that is a rank-preserver, i.e., $\text{rank}(M(X)) = \text{rank}(X)$ for all $X \in \mathbb{R}^{q \times q}$; examples of rank-preservers include operators that act via conjugation by non-singular matrices and the transpose operation. If each $\mathbf{y}^{(j)} = \mathcal{L}(X^{(j)})$ for a linear map $\mathcal{L}$ and low-rank matrices $\{X^{(j)}\}_{j=1}^n$, then we also have that each $\mathbf{y}^{(j)} = \mathcal{L} \circ M^{-1}(M(X^{(j)}))$, where by construction each $X^{(j)}$ has the same rank as the corresponding $M(X^{(j)})$. This non-uniqueness presents a difficulty as the image of the nuclear norm ball under a linear map $\mathcal{L}$ is, in general, different than it is under $\mathcal{L} \circ M^{-1}$ for an arbitrary rank-preserver M. Consequently, due to its invariances the

factorization (3) does not uniquely specify a regularizer. We investigate this point in Sect. 2.2 by analyzing the structure of rank-preserving linear maps, and we describe an approach to associate a unique regularizer to a family of linear maps obtained from equivalent factorizations. Our method entails putting linear maps in an appropriate 'canonical' form using the Operator Sinkhorn iterative procedure, which was developed by Gurvits to solve certain quantum matching problems [34]; this algorithm is an operator analog of the diagonal congruence scaling technique for nonnegative matrices developed by Sinkhorn [55].

In Sect. 2, we describe an alternating update algorithm to compute a factorization of form (3). With the $\mathcal{L}_i$'s fixed, updating the $X^{(j)}$'s entails the solution of affine rank minimization problems. Although this problem is intractable in general [44], in recent years several tractable heuristics have been developed and proven to succeed under suitable conditions [30,36,50]. With the $X^{(j)}$'s fixed, the $\mathcal{L}_i$'s are updated by solving a least-squares problem followed by an application of the Operator Sinkhorn iterative procedure to put the map $\mathcal{L}$ in a canonical form as described above. Our alternating update approach is a generalization of methods that are widely employed in dictionary learning for identifying finite atomic sets (see Sect. 2.4).

Section 3 contains the main theorem of this paper on the local linear convergence of our alternating update algorithm. Specifically, suppose a collection of data points $\{\mathbf{y}^{(j)}\}_{j=1}^n \subset \mathbb{R}^d$ is generated as $\mathbf{y}^{(j)} = \mathcal{L}^\star(X^{(j)^\star})$, $j = 1, \ldots, n$ for a linear map $\mathcal{L}^\star : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ that is nearly isometric restricted to low-rank matrices (formally, $\mathcal{L}^\star$ satisfies a *restricted isometry property* [50]) and a collection $\{X^{(j)^\star}\}_{j=1}^n \subset \mathbb{R}^{q \times q}$ of low-rank matrices that is isotropic in a well-defined sense. Given the data $\{\mathbf{y}^{(j)}\}_{j=1}^n$ as input, our alternating update approach is locally linearly convergent to a linear map $\hat{\mathcal{L}} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ with the property that the image of the nuclear norm ball in $\mathbb{R}^{q \times q}$ under $\hat{\mathcal{L}}$ is equal to its image under $\mathcal{L}^\star$, i.e., our procedure identifies the appropriate regularizer that promotes the type of structure contained in the data $\{\mathbf{y}^{(j)}\}_{j=1}^n$; see Theorem 3. Our analysis relies on geometric aspects of determinantal varieties (in particular tangent spaces with respect to these varieties) and their relation to stability properties of Operator Sinkhorn scaling.

We demonstrate the utility of our framework with a series of experimental results on synthetic as well as real data in Sect. 4.

1.3 Related Work

1.3.1 Dictionary Learning

As outlined above, our approach for learning a regularizer from data may be viewed as a semidefinite programming generalization of dictionary learning. The alternating update algorithm we propose in Sect. 2.3 for computing a factorization (3) generalizes similar methods previously developed for dictionary learning [1,3,4,46] (see Sect. 2.4), and the local convergence analysis of our algorithm in Sect. 3 also builds on previous analyses for dictionary learning [1,4]. In contrast to these previous results, the development and the analysis of our method in the present paper are more challenging due to the invariances and associated identifiability issues underlying the factorization

(3), which necessitate the incorporation of the Operator Sinkhorn scaling procedure in our algorithm.

An unresolved matter in our paper—one that has been investigated previously in the context of dictionary learning—is the question of a suitable initialization for our algorithm. In particular, our theory states that our algorithm exhibits linear convergence to the desired solution provided the initial guess is sufficiently close to a linear map that specifies the correct regularizer (in an appropriate metric). We employ random initializations in our experiments with real data in Sect. 4.2, and these are useful in identifying effective semidefinite regularizers that outperform polyhedral regularizers obtained via dictionary learning. Random initialization is the most common technique utilized in practice in dictionary learning as well as in many other structured matrix factorization problems arising in data analysis. To build support for this idea, several researchers have proven that random initialization succeeds with high probability in recovering a desired factorization under suitable conditions in a number of problems [29,58], including in a restricted form of dictionary learning [59,60] in which the polyhedral regularizer is specified as the image of the ℓ_1 ball under an invertible linear map (as described previously, dictionary learning in full generality allows for polyhedral regularizers that may be specified as an image of the ℓ_1 ball under a many-to-one linear map). In a different direction, some recent papers also describe data-driven initialization strategies for dictionary learning based on variants of clustering [2,5]. It would be of interest to develop both these sets of ideas in our context, and we comment on this point in Sect. 5.

1.3.2 Lifts of Convex Sets

A second body of work with which our paper is conceptually related is the literature on lift-and-project representations (or extended formulations) of convex sets. A tractable lift-and-project representation refers to a description of a 'complicated' convex set in $\mathbb{R}^d$ as the projection of a more concisely specified convex set in $\mathbb{R}^{d'}$, with the lifted dimension d' not being too much larger than the original dimension d. As discussed in [32,66], obtaining a suitably structured factorization—of a different nature than that considered in the present paper—of the *slack matrix* of a polytope (and more generally, of the slack operator of a convex set) corresponds to identifying an efficient lift-and-project description of the polytope. On the other hand, we seek a structured factorization of a *data matrix* to identify a convex set (i.e., the unit ball of a regularizer) with an efficient extended formulation and with the additional requirement that the data points (upon suitable scaling) lie on low-dimensional faces of the set. This latter stipulation arises in our context from data analysis considerations, and it is a distinction between our setup and the optimization literature on extended formulations.

1.3.3 Sinkhorn Scaling

A third topic with which our paper has synergies—and to which we make contributions in the course of our analysis—is the literature on Sinkhorn scaling. This algorithm is an iterative procedure for transforming an entrywise nonnegative matrix to a doubly stochastic matrix by diagonal congruence scaling [55]. There is a substantial body of

work on the properties of this algorithm (see [35] and the references therein) as well as on its applications in domains such as combinatorial optimization (approximating the permanent of a matrix [40]) and data analysis (efficiently computing distances between probability distributions [19]). The operator analog of Sinkhorn scaling was developed by Gurvits, and this work was motivated by certain operator analogs of the bipartite matching problem that arise in matroid theory [34]. To the best of our knowledge, our work represents the first application of Operator Sinkhorn scaling in a problem in data analysis. Further, in our investigation of the properties of Algorithm 1, we describe results on the stability of Operator Sinkhorn scaling; these may be of independent interest beyond the specific context of our paper (see "Appendix C").

1.4 Paper Outline

In Sect. 2, we discuss our alternating update algorithm for computing the factorization (3) based on an analysis of the invariances arising in (3). Section 3 gives the main theoretical result concerning the local linear convergence of the algorithm described in Sects. 2, and 4 describes numerical results obtained using our algorithm. We conclude with a discussion of further research directions in Sect. 5.

Notation We denote the Euclidean norm by $\|\cdot\|_{\ell_2}$. We denote the operator or spectral norm by $\|\cdot\|_2$. The kth largest singular value of a linear map is denoted by $\sigma_k(\cdot)$, and the largest and smallest eigenvalues of a self-adjoint linear map are denoted by $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$, respectively. The space of $q \times q$ symmetric matrices is denoted $\mathbb{S}^q$, and the set of $q \times q$ symmetric positive-definite matrices is denoted $\mathbb{S}^q_{++}$. The projection map onto a subspace $\mathcal{V}$ is denoted $\mathcal{P}_\mathcal{V}$. The restriction of a linear map M to a subspace $\mathcal{V}$ is denoted by $M_\mathcal{V}$. Given a self-adjoint linear map $M : \mathcal{V} \to \mathcal{V}$ with $\mathcal{V}$ being a subspace of a vector space $\bar{\mathcal{V}}$, we denote the extension of M to $\bar{\mathcal{V}}$ by $[M]_{\bar{\mathcal{V}}} : \bar{\mathcal{V}} \to \bar{\mathcal{V}}$; the component in $\mathcal{V}$ of the image of any $\mathbf{x} \in \bar{\mathcal{V}}$ under this map is $M\mathcal{P}_\mathcal{V}(\mathbf{x})$, while the component in $\mathcal{V}^\perp$ is the origin. Given a vector space $\mathcal{V}$, we denote the set of linear operators from $\mathcal{V}$ to $\mathcal{V}$ by $\mathrm{End}(\mathcal{V})$. Given matrices $A, B \in \mathbb{R}^{q \times q}$, the linear map $A \boxtimes B \in \mathrm{End}(\mathbb{R}^{q \times q})$ is specified as $A \boxtimes B : X \to \langle B, X \rangle A$. The Kronecker product between two linear maps is specified using the standard $\otimes$ notation. For a collection of matrices $\mathfrak{X} := \{X^{(j)}\}_{j=1}^n \subset \mathbb{R}^{q \times q}$, the covariance is specified as $\Sigma(\mathfrak{X}) = \frac{1}{n} \sum_{j=1}^n X^{(j)} \boxtimes X^{(j)}$. Two quantities associated with this covariance that play a role in our analysis are $\Lambda(\mathfrak{X}) = \frac{1}{2}(\lambda_{\max}(\Sigma(\mathfrak{X})) + \lambda_{\min}(\Sigma(\mathfrak{X})))$ and $\Delta(\mathfrak{X}) = \frac{1}{2}(\lambda_{\max}(\Sigma(\mathfrak{X})) - \lambda_{\min}(\Sigma(\mathfrak{X})))$. Given a matrix $X \in \mathbb{R}^{q \times q}$ of rank r, the tangent space at X with respect to the algebraic variety of $q \times q$ matrices of rank at most r is specified as[1]:

$$\mathcal{T}(X) = \{XA + BX \mid A, B \in \mathbb{R}^{q \times q}\}.$$

---

[1] A rank-r matrix $X \in \mathbb{R}^{q \times q}$ is a smooth point with respect to the variety of $q \times q$ matrices of rank at most r.

2 An Alternating Update Algorithm for Learning Semidefinite Regularizers

In this section, we describe an alternating update algorithm to factor a given data matrix $Y = [\mathbf{y}^{(1)} | \cdots | \mathbf{y}^{(n)}] \in \mathbb{R}^{d \times n}$ as in (3). As discussed previously, the difficulty with obtaining a semidefinite regularizer using a factorization (3) is the existence of infinitely many equivalent factorizations due to the invariances underlying (3). We begin by investigating and addressing this issue in Sects. 2.1 and 2.2, and then, we discuss our algorithm to obtain a regularizer in Sect. 2.3. We contrast our method with techniques that have previously been developed in the context of dictionary learning in Sect. 2.4.

2.1 Identifiability Issues

Building on the discussion in the introduction, for a linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ obtained from the factorization (3) and for any linear rank-preserver $M : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$, there exists an equivalent factorization in which the linear map is $\mathcal{L} \circ M$ (note that M^{-1} is also a rank-preserver if M is a rank-preserver). As the image of the nuclear norm ball in $\mathbb{R}^{q \times q}$ is not invariant under an arbitrary rank-preserver, a regularizer cannot be obtained uniquely from a factorization due to the existence of equivalent factorizations that lead to non-equivalent regularizers. To address this difficulty, we describe an approach to associate a *unique* regularizer to a family of linear maps obtained from equivalent factorizations. We begin by analyzing the structure of rank-preserving linear maps based on the following result [42]:

Theorem 1 [42, Theorem 1], [64, Theorem 9.6.2] *An invertible linear operator $M :$ $\mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$ is a rank-preserver if and only if M is of one of the following two forms for non-singular matrices $W_1, W_2 \in \mathbb{R}^{q \times q}$: $M(X) = W_1 X W_2$ or $M(X) = W_1 X' W_2$.*

This theorem brings the preceding discussion into sharper focus, namely, that the lack of identifiability boils down to the fact that the nuclear norm is not invariant under conjugation of its argument by arbitrary non-singular matrices. However, we note that the nuclear norm ball is invariant under the transpose operation and under conjugation by orthogonal matrices. This observation leads naturally to the idea of employing the *polar decomposition* to describe a rank-preserver:

Corollary 1 *Every rank-preserver $M : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$ can be uniquely decomposed as $M = M^{\mathrm{or}} \circ M^{\mathrm{pd}}$ for rank-preservers $M^{\mathrm{pd}} : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$ and $M^{\mathrm{or}} : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$ with the following properties:*

- *The operator M^{pd} is specified as $M^{\mathrm{pd}}(X) = P_1 X P_2$ for some positive-definite matrices $P_1, P_2 \in \mathbb{S}_{++}^q$.*
- *The operator M^{or} is of one of the following two forms for orthogonal matrices $U_1, U_2 \in \mathbb{R}^{q \times q}$: $M^{\mathrm{or}}(X) = U_1 X U_2$ or $M^{\mathrm{or}}(X) = U_1 X' U_2$.*

Proof The result follows by combining Theorem 1 with the polar decomposition. □

We refer to rank-preservers of the type M^{pd} in this corollary as *positive-definite rank-preservers* and to those of the type M^{or} as *orthogonal rank-preservers*. This corollary highlights the point that the key source of difficulty in identifying a regularizer uniquely from a factorization is due to positive-definite rank-preservers. A natural approach to address this challenge is to put a given linear map $\mathcal{L}$ into a 'canonical' form that removes the ambiguity due to positive-definite rank-preservers. In other words, we seek a distinguished subset of *normalized* linear maps with the following properties: (*a*) for a linear map $\mathcal{L}$, the set $\{\mathcal{L} \circ M^{\mathrm{pd}} \mid M^{\mathrm{pd}}$ is a positive-definite rank-preserver$\}$ intersects the collection of normalized maps at precisely one point, and (*b*) for any normalized linear map $\mathcal{L}$, every element of the set $\{\mathcal{L} \circ M^{\mathrm{or}} \mid M^{\mathrm{or}}$ is an orthogonal rank-preserver$\}$ is also normalized. The following definition possesses both of these attributes:

Definition 1 Let $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ be a linear map, and let $\mathcal{L}_i \in \mathbb{R}^{q \times q}$, $i = 1, \ldots, d$ be the component linear functionals of $\mathcal{L}$. Then, $\mathcal{L}$ is said to be *normalized* if $\sum_{i=1}^{d} \mathcal{L}_i {\mathcal{L}_i}' = qI$ and $\sum_{i=1}^{d} {\mathcal{L}_i}' \mathcal{L}_i = qI$.

The utility of this definition in resolving our identifiability issue is based on a paper by Gurvits [34]. Specifically, for a generic linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$, the results in [34] imply that there exists a *unique* positive-definite rank-preserver $N_{\mathcal{L}} : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$ so that $\mathcal{L} \circ N_{\mathcal{L}}$ is normalized (see Corollary 2 in the sequel); this feature address our first requirement above. One can also check that the second requirement above is satisfied by this definition—any normalized linear map composed with any orthogonal rank-preserver is also normalized. Further, the collection of normalized maps defined above may be viewed as an affine algebraic variety specified by polynomials of degree two. One can check that any notion of normalization (specified as a real variety) that satisfies the two attributes described previously cannot be an affine space and therefore must be specified by polynomials of degree at least two. Consequently, our definition of normalization is in some sense also as 'simple' as possible from an algebraic perspective.[2]

In addition to satisfying these appealing properties, our notion of normalization also possesses an important computational attribute—given a (generic) linear map, a normalizing positive-definite rank-preserver for the map can be computed using the Operator Sinkhorn iterative procedure developed in [34]. Thus, the following method offers a natural approach for uniquely associating a regularizer to an equivalence class of factorizations.

Obtaining a regularizer from a linear map: Given a linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ obtained from a factorization (3), the unit ball of the regularizer we associate with this factorization is the image of the nuclear norm ball in $\mathbb{R}^{q \times q}$ under the linear map $\mathcal{L} \circ N_{\mathcal{L}}$; here $N_{\mathcal{L}}$ is the unique positive-definite rank-preserver that normalizes $\mathcal{L}$ (as discussed in the sequel in Corollary 2, such unique normalizing rank-preservers exist for generic maps $\mathcal{L}$).

---

[2] Note that any affine variety over the reals may be defined by polynomials of degree at most two by suitably adding extra variables; in our discussion here on normalization, we consider varieties defined without additional variables.

The soundness of this approach follows from the fact that linear maps from equivalent factorizations produce the same regularizer. We prove a result on this point in the next section (see Proposition 1), and we also discuss algorithmic consequences of the Operator Sinkhorn scaling procedure of [34].

2.2 Normalizing Maps via Operator Sinkhorn Scaling

From the discussion in the preceding section, a key step in associating a unique regularizer to a collection of equivalent factorizations is to normalize a given linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$. In this section, we describe how this may be accomplished by appealing to the work of Gurvits [34].

Given a linear operator $\mathsf{T} : \mathbb{S}^q \to \mathbb{S}^q$ that leaves the positive-semidefinite cone invariant, Gurvits consider the question of the existence (and computation) of positive-definite matrices $P_1, P_2 \in \mathbb{S}^q_{++}$ such that the rescaled operator $\tilde{\mathsf{T}} = (P_1 \otimes P_1) \circ \mathsf{T} \circ (P_2 \otimes P_2)$ has the property that $\tilde{\mathsf{T}}(I) = \tilde{\mathsf{T}}'(I) = I$, i.e., the identity matrix is an eigenmatrix of the rescaled operator $\tilde{\mathsf{T}}$ and its adjoint [34]. This problem is an operator analog of the classical problem of transforming entrywise square nonnegative matrices to doubly stochastic matrices by diagonal congruence scaling. This *matrix scaling* problem was originally studied by Sinkhorn [55], and he developed an iterative solution technique that is known as Sinkhorn scaling. Gurvits developed an operator analog of classical Sinkhorn scaling that proceeds by alternately performing the updates $\mathsf{T} \leftarrow (\mathsf{T}(I)^{-1/2} \otimes \mathsf{T}(I)^{-1/2}) \circ \mathsf{T}$ and $\mathsf{T} \leftarrow \mathsf{T} \circ (\mathsf{T}'(I)^{-1/2} \otimes \mathsf{T}'(I)^{-1/2})$, this sequence of operations is known as the *Operator Sinkhorn Iteration*. The next theorem concerning the convergence of this iterative method is proved in [34]. Following the terminology in [34], a linear operator $\mathsf{T} : \mathbb{S}^q \to \mathbb{S}^q$ is *rank indecomposable* if it satisfies the inequality rank $(\mathsf{T}(Z)) >$ rank(Z) for all $Z \succeq 0$ with $1 \leq$ rank$(Z) < q$; this condition is an operator analog of a matrix being irreducible.

Theorem 2 [34, Theorems 4.6 and 4.7] *Let* $\mathsf{T} : \mathbb{S}^q \to \mathbb{S}^q$ *be a rank indecomposable linear operator. There exist* unique *positive-definite matrices* $P_1, P_2 \in \mathbb{S}^q_{++}$ *with* $\det(P_1) = 1$ *such that* $\tilde{\mathsf{T}} = (P_1 \otimes P_1) \circ \mathsf{T} \circ (P_2 \otimes P_2)$ *satisfies the conditions* $\tilde{\mathsf{T}}(I) = \tilde{\mathsf{T}}'(I) = I$. *Moreover, the Operator Sinkhorn Iteration initialized with* T *converges to* $\tilde{\mathsf{T}}$.

Remark The condition $\det(P_1) = 1$ is imposed purely to avoid the ambiguity that arises from setting $P_1 \leftarrow \alpha P_1$ and $P_2 \leftarrow \frac{1}{\alpha} P_2$ for positive scalars α. Other than this degree of freedom, there are no other positive-definite matrices that satisfy the property that the rescaled operator $\tilde{\mathsf{T}}$ in this theorem as well as its adjoint both have the identity as a eigenmatrix.

These ideas and results are directly relevant in our context as follows. For any linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$, we may associate an operator $\mathsf{T}_{\mathcal{L}} : \mathbb{S}^q \to \mathbb{S}^q$ defined as $\mathsf{T}_{\mathcal{L}}(Z) = \frac{1}{q} \sum_{i=1}^d \mathcal{L}_i Z \mathcal{L}_i'$, which has the property that it leaves the positive-semidefinite cone invariant. Rescaling the operator $\mathsf{T}_{\mathcal{L}}$ via positive-definite matrices $P_1, P_2 \in \mathbb{S}^q_{++}$ to obtain $\tilde{\mathsf{T}}_{\mathcal{L}} = (P_1 \otimes P_1) \circ \mathsf{T}_{\mathcal{L}} \circ (P_2 \otimes P_2)$ corresponds to conjugating the component linear functionals $\{\mathcal{L}_i\}_{i=1}^d$ of $\mathcal{L}$ by P_1 and P_2. Consequently, rescaling

---

Algorithm 1 Normalizing a linear map via the Operator Sinkhorn Iteration

---

Input: A linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ with component functionals $\mathcal{L}_i$, $i = 1, \ldots, d$
Require: A normalized map $\mathcal{L} \circ M$ where $M : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ is a rank-preserver that acts via conjugation by positive-definite matrices
Algorithm: Repeat until convergence
1. $R = \sum_{i=1}^d \mathcal{L}_i \mathcal{L}_i{}'$
2. $\mathcal{L}_i \leftarrow \sqrt{q} R^{-\frac{1}{2}} \mathcal{L}_i$, $i = 1, \ldots, d$
3. $C = \sum_{i=1}^d \mathcal{L}_i{}' \mathcal{L}_i$
4. $\mathcal{L}_i \leftarrow \sqrt{q} \mathcal{L}_i C^{-\frac{1}{2}}$, $i = 1, \ldots, d$

$\mathsf{T}_{\mathcal{L}}$ so that $\tilde{\mathsf{T}}_{\mathcal{L}} = (P_1 \otimes P_1) \circ \mathsf{T}_{\mathcal{L}} \circ (P_2 \otimes P_2)$ and its adjoint both have the identity as an eigenmatrix is equivalent to composing $\mathcal{L}$ by a positive-definite rank-preserver $N = P_1 \otimes P_2$ so that $\mathcal{L} \circ N$ is normalized. Based on this correspondence, Algorithm 1 gives a specialization of the general Operator Sinkhorn Iteration to our setting for normalizing a linear map $\mathcal{L}$.[3] We also have the following corollary to Theorem 2:

Corollary 2 *Let* $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ *be a linear map, and suppose* $\mathrm{rank}(\sum_{i=1}^d \mathcal{L}_i Z \mathcal{L}_i{}')$ $> \mathrm{rank}(Z)$ *for all* $Z \succeq 0$ *with* $1 \leq \mathrm{rank}(Z) < q$ *(i.e., the operator* $\mathsf{T}_{\mathcal{L}}(Z) = \frac{1}{q} \sum_{i=1}^d \mathcal{L}_i Z \mathcal{L}_i{}'$ *is rank indecomposable). There exists a* unique *positive-definite rank-preserver* $N_{\mathcal{L}} : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$ *such that* $\mathcal{L} \circ N_{\mathcal{L}}$ *is normalized. Moreover, Algorithm 1 initialized with* $\mathcal{L}$ *converges to* $\mathcal{L} \circ N_{\mathcal{L}}$.

Proof The existence of a positive-definite rank-preserver $N_{\mathcal{L}}$ and the convergence of Algorithm 1 follow directly from Theorem 2. We need to prove that $N_{\mathcal{L}}$ is unique. Let $\tilde{N}_{\mathcal{L}} : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$ be any positive-definite rank-preserver such that $\mathcal{L} \circ \tilde{N}_{\mathcal{L}}$ is normalized. By Theorem 1, there exists positive-definite matrices $P_1, P_2, \tilde{P}_1, \tilde{P}_2$ such that $N_{\mathcal{L}} = P_1 \otimes P_2$ and $\tilde{N}_{\mathcal{L}} = \tilde{P}_1 \otimes \tilde{P}_2$. Without loss of generality, we may assume that $\det(P_1) = \det(\tilde{P}_1) = 1$. By Theorem 2, we have $P_1 = \tilde{P}_1$ and $P_2 = \tilde{P}_2$ and consequently that $N_{\mathcal{L}} = \tilde{N}_{\mathcal{L}}$. □

Generic linear maps $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ (for $d \geq 2$) satisfy the condition $\mathrm{rank}(\sum_{i=1}^d \mathcal{L}_i Z \mathcal{L}_i{}') > \mathrm{rank}(Z)$ for all $Z \succeq 0$ with $1 \leq \mathrm{rank}(Z) < q$. Therefore, this assumption in Corollary 2 is not particularly restrictive. A consequence of the uniqueness of the positive-definite rank-preserver $N_{\mathcal{L}}$ in Corollary 2 is that our normalization scheme associates a unique regularizer to every collection of equivalent factorizations:

Proposition 1 *Let* $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ *be a linear map, and suppose* $\mathrm{rank}(\sum_{i=1}^d \mathcal{L}_i Z \mathcal{L}_i{}')$ $> \mathrm{rank}(Z)$ *for all* $Z \succeq 0$ *with* $1 \leq \mathrm{rank}(Z) < q$. *Let* $M : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$ *be any rank-preserver. Suppose* $N_{\mathcal{L}}$ *and* $N_{\mathcal{L} \circ M}$ *are positive-definite rank-preservers such that* $\mathcal{L} \circ N_{\mathcal{L}}$ *and* $\mathcal{L} \circ M \circ N_{\mathcal{L} \circ M}$ *are normalized. Then, the image of the nuclear norm ball under* $\mathcal{L} \circ N_{\mathcal{L}}$ *is the same as it is under* $\mathcal{L} \circ M \circ N_{\mathcal{L} \circ M}$.

---

[3] Algorithm 1 requires the computation of a matrix square root at every iteration. By virtue of the fact that the operator $\mathsf{T}_{\mathcal{L}}$ which we wish to rescale is completely positive, it is possible to normalize $\mathcal{L}$ using only rational matrix operations via a modified scheme known as the Rational Operator Sinkhorn Iteration [34].

Remark Note that if the linear map $\mathcal{L}$ satisfies the property that $\mathrm{rank}(\sum_{i=1}^{d} \mathcal{L}_i Z \mathcal{L}_i') > \mathrm{rank}(Z)$ for all $Z \succeq 0$ with $1 \leq \mathrm{rank}(Z) < q$, then so does the linear map $\mathcal{L} \circ M$ for any rank-preserver M.

Proof As $M^{-1} \circ N_{\mathcal{L}}$ is a rank-preserver, we can apply Corollary 1 to obtain the decomposition $M^{-1} \circ N_{\mathcal{L}} = \bar{M}^{\mathrm{or}} \circ \bar{M}^{\mathrm{pd}}$, where $\bar{M}^{\mathrm{or}}$ is an orthogonal rank-preserver and $\bar{M}^{\mathrm{pd}}$ is a positive-definite rank-preserver.

We claim that $N_{\mathcal{L} \circ M} = M^{-1} \circ N_{\mathcal{L}} \circ \bar{M}^{\mathrm{or}'}$. First, we have $M^{-1} \circ N_{\mathcal{L}} \circ \bar{M}^{\mathrm{or}'} = \bar{M}^{\mathrm{or}} \circ \bar{M}^{\mathrm{pd}} \circ \bar{M}^{\mathrm{or}'}$, which implies that this operator is positive-definite. Next, we note that a linear map that is obtained by right multiplication of a normalized linear map with an orthogonal rank-preserver is also normalized, and hence, the linear map $\mathcal{L} \circ M \circ M^{-1} \circ N_{\mathcal{L}} \circ \bar{M}^{\mathrm{or}'} = \mathcal{L} \circ N_{\mathcal{L}} \circ \bar{M}^{\mathrm{or}'}$ is normalized. By applying Corollary 2, we conclude that $N_{\mathcal{L} \circ M} = M^{-1} \circ N_{\mathcal{L}} \circ \bar{M}^{\mathrm{or}'}$.

Consequently, we have $\mathcal{L} \circ M \circ N_{\mathcal{L} \circ M} = \mathcal{L} \circ N_{\mathcal{L}} \circ \bar{M}^{\mathrm{or}'}$. As the nuclear norm ball is invariant under the action of the orthogonal rank-preserver $\bar{M}^{\mathrm{or}'}$, it follows that the image of the nuclear norm ball under the map $\mathcal{L} \circ N_{\mathcal{L}}$ is the same as it is under the map $\mathcal{L} \circ M \circ N_{\mathcal{L} \circ M}$. □

The polynomial-time complexity of the (general) Operator Sinkhorn iterative procedure—in terms of the number of iterations required to obtain a desired accuracy to the fixed point—has recently been established in [28]. In summary, this approach provides a computationally tractable method to normalize linear maps, and consequently to associate a unique regularizer to a collection of equivalent factorizations.

2.3 An Alternating Update Algorithm for Matrix Factorization

Given the resolution of the identifiability issues in the preceding two sections, we are now in a position to describe an algorithmic approach for computing a factorization (3) of a data matrix $Y = [\mathbf{y}^{(1)} | \cdots | \mathbf{y}^{(n)}] \in \mathbb{R}^{d \times n}$ to obtain a semidefinite regularizer that promotes the type of structure contained in Y. Specifically, given a target dimension q, our objective is to obtain a normalized linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ and a collection $\{X^{(j)}\}_{j=1}^{n}$ of low-rank matrices such that $\sum_{i=1}^{n} \|\mathbf{y}^{(j)} - \mathcal{L}(X^{(j)})\|_{\ell_2}^2$ is minimized. Our procedure is an alternating update technique that sequentially updates the low-rank $X^{(j)}$'s followed by an update of $\mathcal{L}$. We assume that our algorithm is provided with a data matrix $Y \in \mathbb{R}^{d \times n}$, a target dimension q, and an initial guess for the normalized map $\mathcal{L}$. Our method is summarized in Algorithm 3.

2.3.1 Updating the Low-Rank Matrices $\{X^{(j)}\}_{j=1}^{n}$

In this stage, a normalized linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ is fixed, and the objective is to find low-rank matrices $\{X^{(j)}\}_{j=1}^{n}$ such that $\mathbf{y}^{(j)} \approx \mathcal{L}(X^{(j)})$ for each $j = 1, \ldots, n$. Without the requirement that the $X^{(j)}$'s be low rank, such linear inverse problems are ill-posed in our context as q^2 is typically taken to be larger than d. With the low-rank restriction, this problem is well-posed and it is known as the *affine rank minimization problem*. This problem is NP-hard in general [44]. However, due to its prevalence in

---

Algorithm 2 Obtaining a low-rank matrix near an affine space via Singular Value Projection

---

Input: A linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$, a point $\mathbf{y} \in \mathbb{R}^d$, a target rank r, an initial guess $X \in \mathbb{R}^{q \times q}$, and a damping parameter $\nu \in (0, 1]$

Require: A matrix $\hat{X}$ of rank at most r such that $\|\mathbf{y} - \mathcal{L}(\hat{X})\|_{\ell_2}$ is minimized, i.e., solve (5)

Initialization $X = 0$

Algorithm: Repeat until convergence

1. $X \leftarrow X + \nu \mathcal{L}'(\mathbf{y} - \mathcal{L}(X))$ (i.e., take a gradient step with respect to the objective of (5))
2. Compute top-r singular vectors and singular values of X: $U_r, V_r \in \mathbb{R}^{q \times r}$, $\Sigma_r \in \mathbb{R}^{r \times r}$
3. $X \leftarrow U_r \Sigma_r V_r'$

---

a range of application domains [26,50], significant efforts have been devoted toward the development of tractable heuristics that are useful in practice and that succeed on certain families of problem instances. We describe next two popular heuristics for this problem.

The first approach—originally proposed by Fazel in her thesis [26] and subsequently analyzed in [12,50]—is based on a convex relaxation in which the rank constraint is replaced by the nuclear norm penalty, which leads to the following convex program:

$$\hat{X} = \arg\min_{X \in \mathbb{R}^{q \times q}} \quad \frac{1}{2} \|\mathbf{y} - \mathcal{L}(X)\|_{\ell_2}^2 + \lambda \|X\|_\star. \tag{4}$$

Here $\mathbf{y} \in \mathbb{R}^d$ and $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ are the problem data specifying the affine space near which we seek a low-rank solution, and the parameter $\lambda > 0$ provides a trade-off between fidelity to the data (i.e., fit to the specified affine space) and rank of the solution $\hat{X}$. This problem is a semidefinite program, and it can solved to a desired precision in polynomial time using standard software [45,62].

Another popular method for the affine rank minimization problem is based on directly attempting to solve the following non-convex optimization problem via alternating projection for a specified rank $r < q$:

$$\hat{X} = \arg\min_{X \in \mathbb{R}^{q \times q}} \quad \|\mathbf{y} - \mathcal{L}(X)\|_{\ell_2}^2$$
$$\text{s.t.} \quad \text{rank}(X) \leq r. \tag{5}$$

This problem is intractable to solve globally in general, but the heuristic described in Algorithm 2 provides an approach that provably succeeds under certain conditions [30,36]. The utility of this method in comparison with convex program (4) is that applying the procedure described in Algorithm 2 is much more tractable in large-scale settings in comparison with solving (4).

The analyses in [27,30,36,50] rely on the map $\mathcal{L}$ satisfying the following type of restricted isometry condition introduced in [50]:

Definition 2 Consider a linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$. For each $k = 1, \ldots, q$ the *restricted isometry constant of order k* is defined as the smallest $\delta_k(\mathcal{L})$ such that:

$$1 - \delta_k(\mathcal{L}) \leq \frac{\|\mathcal{L}(X)\|_{\ell_2}^2}{\|X\|_{\ell_2}^2} \leq 1 + \delta_k(\mathcal{L})$$

for all matrices $X \in \mathbb{R}^{q \times q}$ with rank less than or equal to k.

If a linear map $\mathcal{L}$ has a small restricted isometry constant for some order k, then the affine rank minimization problem is, in some sense, well-posed when restricted to matrices of rank less than or equal to k. The results in [27,30,36,50] go much further by demonstrating that if $\mathbf{y} = \mathcal{L}(X^\star) + \boldsymbol{\epsilon}$ for $\boldsymbol{\epsilon} \in \mathbb{R}^d$ and with $\mathrm{rank}(X^\star) \leq r$, and if the map $\mathcal{L}$ satisfies a bound on the restricted isometry constant $\delta_{4r}(\mathcal{L})$, then both convex program (4) and the procedure in Algorithm 2 applied to solve (5) provide solutions $\hat{X}$ such that $\|\hat{X} - X^\star\|_{\ell_2} \lesssim C\|\boldsymbol{\epsilon}\|_{\ell_2}$. Due to the qualitative similarity in the performance guarantees for these approaches, either of them is appropriate as a subroutine for updating the $X^{(j)}$'s in our alternating update method for computing a factorization of a given data matrix $Y \in \mathbb{R}^{d \times n}$. Algorithm 3 is therefore stated in a general manner to retain this flexibility. In our main theoretical result in Sect. 3.3, we assume that the $X^{(j)}$'s are updated by solving (5) using the heuristic outlined in Algorithm 2; our analysis could equivalently be carried out by assuming that the $X^{(j)}$'s are updated by solving (4).

2.3.2 Updating the Linear Map $\mathcal{L}$

In this stage, the low-rank matrices $\{X^{(j)}\}_{j=1}^n$ are fixed and the goal is to obtain a normalized linear map $\mathcal{L}$ such that $\sum_{i=1}^n \|\mathbf{y}^{(j)} - \mathcal{L}(X^{(j)})\|_{\ell_2}^2$ is minimized. Our procedure for this update consists of two steps. First, we solve the following least-squares problem:

$$\tilde{\mathcal{L}} = \underset{\substack{\bar{\mathcal{L}}:\mathbb{R}^{q \times q} \to \mathbb{R}^d \\ \bar{\mathcal{L}} \text{ is a linear map}}}{\arg\min} \sum_{i=1}^n \|\mathbf{y}^{(j)} - \bar{\mathcal{L}}(X^{(j)})\|_{\ell_2}^2 \tag{6}$$

This problem can be solved, for example, via a pseudoinverse computation. Next, we apply the procedure described in Algorithm 1 to the updated $\tilde{\mathcal{L}}$ obtained from (6) in order to normalize it.

2.4 Comparison with Dictionary Learning

As described in Sect. 1.1, the dictionary learning literature considers the following factorization problem: given a collection of data points $\{\mathbf{y}^{(j)}\}_{j=1}^n \subset \mathbb{R}^d$ and a target dimension p, find a linear map $L : \mathbb{R}^p \to \mathbb{R}^d$ and a collection of sparse vectors $\{\mathbf{x}^{(j)}\}_{j=1}^n \subset \mathbb{R}^p$ such that $\mathbf{y}^{(j)} = L\mathbf{x}^{(j)}$ for each j. As with (3), the linear map L does not lead to a unique polyhedral regularizer. Specifically, for any linear sparsity-preserver $M : \mathbb{R}^p \to \mathbb{R}^p$, there is an equivalent factorization in which the linear map is LM. In parallel to Corollary 1, one can check that M is a sparsity-preserver if and only if M is a composition of a positive-definite diagonal matrix and a signed permutation

---

Algorithm 3 Computing a factorization via alternating updates

---

Input: A data matrix $Y = [\mathbf{y}^{(1)} | \cdots | \mathbf{y}^{(n)}] \in \mathbb{R}^{d \times n}$, a target dimension q, an initial guess for a normalized linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$, a target rank $r < q$

Require: A normalized linear map $\hat{\mathcal{L}} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ and a collection of matrices $\{\hat{X}^{(j)}\}_{j=1}^n$ with rank at most r such that $\sum_{i=1}^n \|\mathbf{y}^{(j)} - \hat{\mathcal{L}}(\hat{X}^{(j)})\|_{\ell_2}^2$ is minimized

Algorithm: Repeat until convergence

1.[Update $X^{(j)}$'s; $\mathcal{L}$ fixed] Obtain matrices $\{X^{(j)}\}_{j=1}^n$ of rank at most r such that $\sum_{i=1}^n \|\mathbf{y}^{(j)} - \mathcal{L}(X^{(j)})\|_{\ell_2}^2$ is minimized. This can be accomplished either via Algorithm 2 or by solving (4) for a suitable choice of λ.

2.[Update $\mathcal{L}$; $X^{(j)}$'s fixed] $\tilde{\mathcal{L}} \leftarrow \underset{\substack{\tilde{\mathcal{L}}:\mathbb{R}^{q \times q} \to \mathbb{R}^d \\ \tilde{\mathcal{L}} \text{ is a linear map}}}{\arg\min} \sum_{i=1}^n \|\mathbf{y}^{(j)} - \tilde{\mathcal{L}}(X^{(j)})\|_{\ell_2}^2$

3.[Normalize $\mathcal{L}$] Normalize updated linear map from previous step using Algorithm 1.

---

matrix. Since the ℓ_1 ball is invariant under the action of a signed permutation, the main source of difficulty in obtaining a unique regularizer from a factorization is due to sparsity-preservers that are positive-definite diagonal matrices. A common convention in dictionary learning that addresses this identifiability issue is to require that each of the columns of L has unit Euclidean norm; for a generic linear map L, there is a unique positive-definite diagonal matrix D such that LD consists of unit-norm columns. Adopting a similar reasoning as in Sect. 2.2, one can check that this normalization resolves the issue of associating a unique regularizer to an equivalence of factorizations.

The most popular approach for computing a factorization in dictionary learning is based on alternately updating the map L and the sparse vectors $\{\mathbf{x}^{(j)}\}_{j=1}^n$. For a fixed linear map L, updating the $\mathbf{x}^{(j)}$'s entails the solution of a sparse linear inverse problem for each j. That is, for each j we seek a sparse vector $\mathbf{x}^{(j)}$ in the affine space $\mathbf{y}^{(j)} = L\mathbf{x}$. Although this problem in NP-hard in general, there is a significant literature on tractable heuristics that succeed under suitable conditions [13, 14, 17, 22–24]; indeed, this work predates and served as a foundation for the literature on the affine rank minimization problem. Prominent examples include the lasso [61], which is a convex relaxation approach akin to (4), and iterative hard thresholding [9], which is analogous to Algorithm 2. For a fixed collection $\{\mathbf{x}^{(j)}\}_{j=1}^n$, the linear map L is then updated by solving a least-squares problem followed by a rescaling of the columns so that they have unit Euclidean norm.

We note that each step in this procedure has a direct parallel to a corresponding step of Algorithm 3. In summary, our proposed approach for obtaining a semidefinite regularizer via matrix factorization is a generalization of previous methods in the dictionary learning literature for obtaining a polyhedral regularizer.

3 Convergence Analysis of Our Algorithm

This section describes the main theoretical result on the local convergence of our algorithm. We begin by discussing the setup and an outline of our analysis in Sects. 3.1 and 3.2, respectively. The statement of our main theorem with deterministic conditions is given in Sect. 3.3, and we describe natural random ensembles that satisfy these

deterministic conditions with high probability in Sect. 3.4. The proof of our theorem is discussed in Sect. 3.5.

3.1 Theoretical Setup

The setup underlying our main theorem is as follows. We assume that we are given a collection of data points $\{\mathbf{y}^{(j)\star}\}_{j=1}^{n} \subset \mathbb{R}^d$ with each $\mathbf{y}^{(j)\star} = \mathcal{L}^\star\left(X^{(j)\star}\right)$, where $\mathcal{L}^\star : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ is a linear map and $\mathfrak{X}^\star := \{X^{(j)\star}\}_{j=1}^{n} \subset \mathbb{R}^{q \times q}$ is a collection of low-rank matrices. Without loss of generality, we may take $\mathcal{L}^\star$ to be normalized and surjective. Our objective is to obtain a linear map $\hat{\mathcal{L}} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ with the property that the image of the nuclear norm ball in $\mathbb{R}^{q \times q}$ under $\mathcal{L}^\star$ is the same as it is under $\hat{\mathcal{L}}$. To this end, we seek a linear map $\hat{\mathcal{L}}$ that can be expressed as the composition of $\mathcal{L}^\star$ with an orthogonal rank-preserver (recall that the nuclear norm ball is invariant under the action of an orthogonal rank-preserver).

As this goal is distinct from the more restrictive requirement that $\hat{\mathcal{L}}$ must equal $\mathcal{L}^\star$, we need an appropriate measure of the 'distance' of a linear map to $\mathcal{L}^\star$. A convenient approach to addressing this issue is to express a linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ in terms of $\mathcal{L}^\star$ as follows, given any linear rank-preserver $M : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$:

$$\mathcal{L} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}) \circ M, \tag{7}$$

Here $\mathsf{I} \in \mathrm{End}(\mathbb{R}^{q \times q})$ is the identity map and the error term $\mathsf{E} = \mathcal{L}^{\star+} \circ (\mathcal{L} \circ M^{-1} - \mathcal{L}^\star) \in \mathrm{End}(\mathbb{R}^{q \times q})$; the assumption that $\mathcal{L}^\star$ is surjective is key as $\mathcal{L}^{\star+}$ is the right inverse of $\mathcal{L}^\star$. By varying the rank-preserver M in (7), the error term E changes. If there exists an *orthogonal* rank-preserver M such that the corresponding error E is small, then in some sense the image of the nuclear norm ball under $\mathcal{L}$ is close to the image under $\mathcal{L}^\star$. This observation suggests that the closeness between $\mathcal{L}$ and $\mathcal{L}^\star$ may be measured as the smallest error E that one can obtain by varying M over the set of orthogonal rank-preservers. The following result suggests that one can in fact vary M over *all* rank-preservers, provided we have the additional condition that $\mathcal{L}$ is also normalized. The additional flexibility provided by varying M over all rank-preservers is well suited to characterizing the effects of normalization via Operator Sinkhorn scaling in our analysis, as described in the next section.

Proposition 2 *Suppose $\mathcal{L}, \mathcal{L}^\star : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ are normalized linear maps such that (i) $\mathcal{L}^\star$ satisfies the restricted isometry condition $\delta_1(\mathcal{L}^\star) \le 1/10$, and (ii) $\mathcal{L} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}) \circ M$ for a linear rank-preserver M with $\|\mathsf{E}\|_{\ell_2} \le 1/(150\sqrt{q}\|\mathcal{L}^\star\|_2)$. Then, there exists an orthogonal rank-preserver M^{or} such that $\|M^{\mathrm{or}} - M\|_2 \le 300\sqrt{q}\|\mathcal{L}^\star\|_2\|\mathsf{E}\|_{\ell_2}$.*

In words, if both $\mathcal{L}$ and $\mathcal{L}^\star$ are normalized and if there exists a rank-preserver M such that $\|\mathsf{E}\|_{\ell_2}$ is small in (7), then M is close to an orthogonal rank-preserver;[4] in turn, this implies that the image of the nuclear norm ball under $\mathcal{L}^\star$ is close to the image

---

[4] The restricted isometry condition in Proposition 2 is a mild one; we require a stronger restricted isometry condition on $\mathcal{L}^\star$ in Theorem 3.

of the nuclear norm ball under $\mathcal{L}$. These observations motivate the following definition as a measure of the distance between normalized linear maps $\mathcal{L}^\star, \mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ for surjective $\mathcal{L}^\star$:

$$\xi_{\mathcal{L}^\star}(\mathcal{L}) := \inf\{\|\mathsf{E}\|_{\ell_2} \mid \exists \mathsf{E} \in \mathrm{End}(\mathbb{R}^{q \times q}) \text{ and a rank-preserver } M \in \mathrm{End}(\mathbb{R}^{q \times q})$$
$$\text{s.t. } \mathcal{L} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}) \circ M\}. \tag{8}$$

In Sect. 3.3, our main result gives conditions under which the sequence of normalized linear maps obtained from Algorithm 3 converges to $\mathcal{L}^\star$ in terms of the distance measure ξ.

3.2 An Approach for Proving a Local Convergence Result

We describe a high-level approach for proving a local convergence result, which motivates the definition of the key parameters that govern the performance of our algorithm. Our proof strategy is to demonstrate that under appropriate conditions the sequence of normalized iterates $\mathcal{L}^{(t)}$ obtained from Algorithm 3 satisfies $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)}) \leq \gamma \xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})$ for a suitable $\gamma < 1$. To bound $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)})$ with respect to $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})$, we consider each of the three steps in Algorithm 3. Fixing notation before, we proceed, let $\mathcal{L}^{(t)} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t)}) \circ M^{(t)}$ for some linear rank-preserver $M^{(t)}$ and for a corresponding error term $\mathsf{E}^{(t)}$. Our objective is to show that there exists a linear rank-preserver $M^{(t+1)}$ and corresponding error term $\mathsf{E}^{(t+1)}$ with $\mathcal{L}^{(t+1)} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t+1)}) \circ M^{(t+1)}$, so that $\|\mathsf{E}^{(t+1)}\|_{\ell_2}$ is suitably bounded above in terms of $\|\mathsf{E}^{(t)}\|_{\ell_2}$. By taking limits, we obtain the desired result in terms of $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})$ and $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)})$.

The first step of Algorithm 3 involves the solution of the following optimization problem for each $j = 1, \ldots, n$:

$$\hat{X}^{(j)} = \underset{X \in \mathbb{R}^{q \times q}}{\arg\min} \left\| \mathbf{y}^{(j)\star} - \mathcal{L}^{(t)}(X) \right\|_{\ell_2}^2 \text{ s.t. } \mathrm{rank}(X) \leq r.$$

As $\mathcal{L}^{(t)} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t)}) \circ M^{(t)}$ and as $\mathbf{y}^{(j)\star} = \mathcal{L}^\star \left(X^{(j)\star} \right)$, the preceding problem can be reformulated in the following manner:

$$M^{(t)}(\hat{X}^{(j)}) = \underset{\tilde{X} \in \mathbb{R}^{q \times q}}{\arg\min} \left\| \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t)}) \left(X^{(j)\star} \right) - \mathcal{L}^\star \circ \mathsf{E}^{(t)} \left(X^{(j)\star} \right) \right.$$
$$\left. - \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t)})(\tilde{X}) \right\|_{\ell_2}^2$$
$$\text{s.t. } \mathrm{rank}(\tilde{X}) \leq r.$$

If $\mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t)})$ satisfies a suitable restricted isometry condition and if $\|\mathcal{L}^\star \circ \mathsf{E}^{(t)} \left(X^{(j)\star} \right)\|_{\ell_2}$ is small, then the results in [30,36] (as described in Sect. 2.3.1) imply that $M^{(t)}(\hat{X}^{(j)}) \approx X^{(j)\star}$. In other words, if $\|\mathsf{E}^{(t)}\|_{\ell_2}$ is small and if $\mathcal{L}^\star$ satisfies

a restricted isometry condition, then $M^{(t)}(\hat{X}^{(j)}) \approx X^{(j)\star}$; the following result states matters formally:

Proposition 3 *Let $\mathcal{L}^\star : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ be a linear map such that (i) $\mathcal{L}^\star$ is normalized, and (ii) $\mathcal{L}^\star$ satisfies the restricted isometry condition $\delta_{4r}(\mathcal{L}^\star) \leq \frac{1}{20}$. Suppose $\mathcal{L} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}) \circ \mathsf{M}$ such that (i) M is a linear rank-preserver and (ii) $\|\mathsf{E}\|_{\ell_2} \leq \min\{1/(50\sqrt{q}), 1/(120r^2\|\mathcal{L}^\star\|_2)\}$. Finally, suppose $\mathbf{y} = \mathcal{L}^\star(X^\star)$, where $X^\star \in \mathbb{R}^{q \times q}$ is a rank-r matrix such that $\sigma_r(X^\star) \geq \sigma_1(X^\star)/2$ and that $\hat{X}$ is the optimal solution to*

$$\hat{X} = \underset{X \in \mathbb{R}^{q \times q}}{\arg\min} \ \|\mathbf{y} - \mathcal{L}(X)\|_{\ell_2}^2 \ \text{s.t.} \ \text{rank}(X) \leq r. \tag{9}$$

Then,

$$M(\hat{X}) = X^\star - \left[\left(\mathcal{L}_{\mathcal{T}(X^\star)}^{\star\prime}\mathcal{L}_{\mathcal{T}(X^\star)}^\star\right)^{-1}\right]_{\mathbb{R}^{q \times q}} \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}\left(X^\star\right) + G,$$

where $\|G\|_{\ell_2} \leq 800r^{5/2}\|\mathcal{L}^\star\|_2^2\|X^\star\|_2\|\mathsf{E}\|_{\ell_2}^2$.

In this proposition, the conclusion is well defined as the linear map $\mathcal{L}_{\mathcal{T}(X^\star)}^{\star\prime}\mathcal{L}_{\mathcal{T}(X^\star)}^\star :$ $\mathcal{T}(X^\star) \to \mathcal{T}(X^\star)$ is invertible due to the restricted isometry condition on $\mathcal{L}^\star$ (see Lemma 1). The proof appears in "Appendix F", and it relies primarily on the first-order optimality conditions of the problem (5). To ensure that the conditions required by this proposition hold, we assume in our main theorem in Sect. 3.3 that $\mathcal{L}^\star$ satisfies the restricted isometry property for rank-r matrices and that the initial guess $\mathcal{L}^{(0)}$ that is supplied to Algorithm 3 is such that $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(0)})$ is small (with a sufficiently good initial guess and by an inductive hypothesis, we have that there exists an error term $\mathsf{E}^{(t)}$ at iteration t such that $\|\mathsf{E}^{(t)}\|_{\ell_2}$ is small).

The second step of Algorithm 3 entails the solution of a least-squares problem. To describe the implications of this step in detail, we consider the linear maps $\mathbf{X}^\star : \mathbf{z} \mapsto \sum_{j=1}^n X^{(j)\star}\mathbf{z}_j$ and $\hat{\mathbf{X}} : \mathbf{z} \mapsto \sum_{j=1}^n \hat{X}^{(j)}\mathbf{z}_j$ from $\mathbb{R}^n$ to $\mathbb{R}^{q \times q}$. With this notation, the second step of Algorithm 3 results in the linear map $\mathcal{L}^{(t)}$ being updated as follows:

$$\tilde{\mathcal{L}}^{(t+1)} = \mathcal{L}^\star \circ \mathbf{X}^\star \circ \hat{\mathbf{X}}^+. \tag{10}$$

In order for the normalized version of $\tilde{\mathcal{L}}^{(t+1)}$ to be close to $\mathcal{L}^\star$ (in terms of the distance measure ξ), we require a deeper understanding of the structure of $\mathbf{X}^\star \circ \hat{\mathbf{X}}^+$, which is the focus of the next proposition. This result relies on the set $\mathcal{X}^\star$ being suitably isotropic, as characterized by the quantities $\Delta(\mathcal{X}^\star)$ and $\Lambda(\mathcal{X}^\star)$.

Proposition 4 *Let $\{A^{(j)}\}_{j=1}^n \subset \mathbb{R}^{q \times q}$ and $\{B^{(j)}\}_{j=1}^n \subset \mathbb{R}^{q \times q}$ be two collections of matrices, and let $\mathbf{A} : \mathbf{z} \mapsto \sum_{j=1}^n A^{(j)}\mathbf{z}_j$ and $\mathbf{B} : \mathbf{z} \mapsto \sum_{j=1}^n B^{(j)}\mathbf{z}_j$ be linear maps from $\mathbb{R}^n$ to $\mathbb{R}^{q \times q}$ associated with these ensembles. Let $\mathsf{Q} : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$ be any invertible linear operator and denote $\omega = \max_j \left\|\mathsf{Q}(B^{(j)}) - A^{(j)}\right\|_{\ell_2}$. If $\omega \leq$*

$$\frac{\sqrt{\Lambda\left(\{A^{(j)}\}_{j=1}^n\right)}}{20} \text{ and if } \frac{\Delta\left(\{A^{(j)}\}_{j=1}^n\right)}{\Lambda\left(\{A^{(j)}\}_{j=1}^n\right)} \le \frac{1}{6}, \text{ then}$$

$$\mathbf{A} \circ \mathbf{B}^+ = \left(\mathsf{I} - \frac{1}{n\Lambda\left(\{A^{(j)}\}_{j=1}^n\right)} \sum_{j=1}^n \left(\mathsf{Q}(B^{(j)}) - A^{(j)}\right) \boxtimes A^{(j)} + \mathsf{F}\right) \circ \mathsf{Q}, \quad (11)$$

where $\|\mathsf{F}\|_{\ell_2} \le 20q \dfrac{\omega^2}{\Lambda\left(\{A^{(j)}\}_{j=1}^n\right)} + 2q \dfrac{\omega\Delta\left(\{A^{(j)}\}_{j=1}^n\right)}{\Lambda\left(\{A^{(j)}\}_{j=1}^n\right)^{3/2}}.$

The proof of this proposition appears in "Appendix G", and it consists of two key elements. First, as ω is bounded, the operator $\mathbf{A} \circ \mathbf{B}^+$ may be approximated as $\mathbf{A} \circ \mathbf{A}^+ \circ \mathsf{Q}$. Second, as the set $\{A^{(j)}\}_{j=1}^n$ is near isotropic based on the assumptions involving $\Delta(\{A^{(j)}\}_{j=1}^n)$ and $\Lambda(\{A^{(j)}\}_{j=1}^n)$, one can show that $\mathbf{A} \circ \mathbf{A}^+$ can be expanded suitably around the identity map I. In the context of our analysis, we apply the conclusions of Proposition 4 with the choice of $A^{(j)} = X^{(j)\star}$, $B^{(j)} = \hat{X}^{(j)}$, and $\mathsf{Q} = M^{(t)}$.

The final step of our analysis is to consider the effect of normalization on the map $\tilde{\mathcal{L}}^{(t)}$ in (10). Denoting the positive-definite rank-preserver that normalizes $\tilde{\mathcal{L}}^{(t+1)}$ by $N_{\tilde{\mathcal{L}}^{(t+1)}}$, we have from Propositions 3 and 4 that the normalized map $\mathcal{L}^{(t+1)}$ obtained after the application of the Operator Sinkhorn iterative procedure to $\tilde{\mathcal{L}}^{(t+1)}$ can be expressed as:

$$\mathcal{L}^{(t+1)} = \mathcal{L}^\star \circ \left(\mathsf{I} - \frac{1}{n\Lambda(\mathfrak{X}^\star)} \sum_{j=1}^n \left(M^{(t)}(\hat{X}^{(j)}) - X^{(j)\star}\right) \boxtimes X^{(j)\star} + \mathsf{F}\right) \circ M^{(t)} \circ N_{\tilde{\mathcal{L}}^{(t+1)}},$$

where $\mathsf{F} \in \mathrm{End}(\mathbb{R}^{q \times q})$ is suitably bounded. As $M^{(t)}$ and $N_{\tilde{\mathcal{L}}^{(t+1)}}$ are both rank-preservers, we need to prove that the expression within parentheses $\mathsf{I} - \frac{1}{n\Lambda(\mathfrak{X}^\star)} \sum_{j=1}^n (M^{(t)}(\hat{X}^{(j)}) - X^{(j)\star}) \boxtimes X^{(j)\star} + \mathsf{F}$ is well approximated as a rank-preserver so that $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)})$ is suitably controlled. To make progress on this front, we note that $\mathsf{I} = I \otimes I$ is a rank-preserver. Therefore, if $-\frac{1}{n\Lambda(\mathfrak{X}^\star)} \sum_{j=1}^n (M^{(t)}(\hat{X}^{(j)}) - X^{(j)\star}) \boxtimes X^{(j)\star} + \mathsf{F}$ is small, a natural approach to characterizing how close $\mathsf{I} - \frac{1}{n\Lambda(\mathfrak{X}^\star)} \sum_{j=1}^n (M^{(t)}(\hat{X}^{(j)}) - X^{(j)\star}) \boxtimes X^{(j)\star} + \mathsf{F}$ is to a rank-preserver is to express this quantity in terms of the following *tangent space* at I with respect to the set of rank-preservers acting on the space of $q \times q$ matrices:

$$\mathcal{W} = \mathrm{span}\{I \otimes W_1 + W_2 \otimes I \mid W_1, W_2 \in \mathbb{R}^{q \times q}\}. \quad (12)$$

The next result gives such an expression.

Proposition 5 *Suppose* $\mathsf{D} : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$ *is a linear operator such that* $\|\mathsf{D}\|_{\ell_2} \le 1/10$ *and* $\mathsf{I} : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$ *is the identity operator. Then, we have that*

$$\mathsf{I} + \mathsf{D} = (\mathsf{I} + \mathcal{P}_{\mathcal{W}^\perp}(\mathsf{D}) + \mathsf{H}) \circ W$$

where $\mathsf{H} : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$ *is a linear operator such that* $\|\mathsf{H}\|_{\ell_2} \le 5\|\mathsf{D}\|_{\ell_2}^2/\sqrt{q}$ *and* $W : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$ *is a linear rank-preserver such that* $\|W - I\|_2 \le 3\|\mathsf{D}\|_{\ell_2}/\sqrt{q}$. *Here, the space* $\mathcal{W}$ *is as defined in* (12).

The proof of this proposition appears in "Appendix H". As detailed in the proof of Theorem 3 in Sect. 3.5, one can combine the preceding three results along with the observation that $c \, \mathcal{P}_{\mathcal{T}(X^\star)} \preceq [(\mathcal{L}^\star_{\mathcal{T}(X^\star)}{}' \mathcal{L}^\star_{\mathcal{T}(X^\star)})^{-1}]_{\mathbb{R}^{q \times q}} \preceq \tilde{c} \, \mathcal{P}_{\mathcal{T}(X^\star)}$ for suitable constants $c, \tilde{c} > 0$ (from Lemma 1 in Sect. 3.5 based on $\mathcal{L}^\star$ satisfying a suitable restricted isometry condition) to conclude that there exists an error term $\mathsf{E}^{(t+1)}$ at iteration $t + 1$ (corresponding to the error term $\mathsf{E}^{(t)}$ at iteration t that we fixed at the beginning of this argument) such that

$$\mathsf{E}^{(t+1)} = \mathcal{P}_{\mathcal{W}^\perp} \circ \left[\frac{1}{n \Lambda(\mathfrak{X}^\star)} \sum_{j=1}^n \left(X^{(j)\star} \boxtimes X^{(j)\star} \right) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)\star})} \right] \left(\mathcal{L}^{\star\prime} \mathcal{L}^\star \circ \mathsf{E}^{(t)} \right)$$
$$+ \mathcal{P}_{\mathcal{W}^\perp}(\mathsf{F}) + \mathcal{O}\left(\|\mathsf{E}^{(t)}\|_{\ell_2}^2 \right). \tag{13}$$

Thus, there are two 'significant' terms in this expression that govern the size of $\|\mathsf{E}^{(t+1)}\|_{\ell_2}$. To control the first term, we require a bound on the following operator norm:

$$\Omega(\mathfrak{X}^\star) := \left\| \mathcal{P}_{\mathcal{W}^\perp} \circ \left[\frac{1}{n} \sum_{j=1}^n \left(X^{(j)\star} \boxtimes X^{(j)\star} \right) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)\star})} \right] \right\|_2. \tag{14}$$

Note that this operator belongs to $\mathrm{End}(\mathrm{End}(\mathbb{R}^{q \times q}))$. In Sect. 3.5, we show that the first significant term in (13) is bounded as $\frac{2\|\mathcal{L}^\star\|_2^2 \Omega(\mathfrak{X}^\star)}{\Lambda(\mathfrak{X}^\star)}\|\mathsf{E}^{(t)}\|_{\ell_2}$. For the second term in (13), we show in Sect. 3.5 that $\|\mathsf{F}\|_{\ell_2} \lesssim \frac{q^2\|\mathcal{L}^\star\|_2 \Delta(\mathfrak{X}^\star)}{\Lambda(\mathfrak{X}^\star)}\|\mathsf{E}^{(t)}\|_{\ell_2}$ based on a bound on $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(0)})$ on the initial guess. Consequently, two of the key assumptions in Theorem 3 concern bounds on the quantities $\frac{\Omega(\mathfrak{X}^\star)}{\Lambda(\mathfrak{X}^\star)}$ and $\frac{\Delta(\mathfrak{X}^\star)}{\Lambda(\mathfrak{X}^\star)}$.

We note that the Operator Sinkhorn scaling procedure for normalization is crucial in our algorithm. Aside from addressing the identifiability issues as discussed in Sect. 2.1, the incorporation of this method also plays an important role in the convergence of Algorithm 3. Specifically, if we do not apply this procedure in each iteration of Algorithm 3, then the estimate of $\mathcal{L}^\star$ at the end of iteration $t + 1$ would be $\tilde{\mathcal{L}}^{(t+1)}$ from (10). In analyzing how close the image of the nuclear norm ball under $\tilde{\mathcal{L}}^{(t+1)}$ is to the image of the nuclear norm ball under $\mathcal{L}^\star$, we would need to consider how close $\mathbf{x}^\star \circ \hat{\mathbf{x}}^+$ is to an *orthogonal* rank-preserver as opposed to an arbitrary rank-preserver; in particular, we cannot apply Proposition 2 as $\tilde{\mathcal{L}}^{(t+1)}$ is not normalized. In analogy to the discussion preceding Proposition 5 and by noting that $I = I \otimes I$ is an orthogonal rank-preserver, we could attempt to express $\mathbf{x}^\star \circ \hat{\mathbf{x}}^+$ in terms of the following tangent space at I with respect to the set of orthogonal rank-preservers:

$$\mathcal{S} = \mathrm{span}\{I \otimes S_1 + S_2 \otimes I \mid S_1, S_2 \in \mathbb{R}^{q \times q} \text{ and skew-symmetric}\}. \tag{15}$$

Following similar reasoning as in the preceding paragraph, the convergence of our algorithm without normalization would be governed by $\|\mathcal{P}_{\mathcal{S}^\perp} \circ [\frac{1}{n}\sum_{j=1}^n (X^{(j)\star} \boxtimes X^{(j)\star}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)\star})}]\|_2$. This operator norm is, in general, much larger than the quantity $\Omega(\mathfrak{X}^\star)$ defined in (14) as $\mathcal{S} \subset \mathcal{W}$, which can in turn affect the convergence of our algorithm. In particular, for a natural random ensemble $\mathfrak{X}^\star$ of low-rank matrices described in Proposition 8 in Sect. 3.4, the condition on $\Omega(\mathfrak{X}^\star)$ in Theorem 3 is satisfied while the analogous condition on $\|\mathcal{P}_{\mathcal{S}^\perp} \circ [\frac{1}{n}\sum_{j=1}^n (X^{(j)\star} \boxtimes X^{(j)\star}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)\star})}]\|_2$ is violated (both of these conclusions hold with high probability), thus highlighting the importance of the inclusion of the normalization step for the convergence of our method; see the remarks following Proposition 8 for details.

3.3 Main Result

The following theorem gives the main result concerning the local convergence of our algorithm:

Theorem 3 *Let* $\mathbf{y}^{(j)} = \mathcal{L}^\star(X^{(j)\star})$, $j = 1, \ldots, n$, *where* $\mathcal{L}^\star : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ *is a linear map and* $\mathfrak{X}^\star := \{X^{(j)\star}\}_{j=1}^n \subset \mathbb{R}^{q \times q}$. *Suppose the collection* $\mathfrak{X}^\star$ *satisfies the following conditions:*

1. *There exists* $r < q$ *and* $s > 0$ *such that* $\mathrm{rank}(X^{(j)\star}) = r$ *and* $s \geq \sigma_1(X^{(j)\star}) \geq \sigma_r(X^{(j)\star}) \geq s/2$ *for each* $j = 1, \ldots, n$;
2. $\frac{\Omega(\mathfrak{X}^\star)}{\Lambda(\mathfrak{X}^\star)} \leq \frac{d}{40q^2}$; *and*
3. $\frac{\Delta(\mathfrak{X}^\star)}{\Lambda(\mathfrak{X}^\star)} \leq \frac{\sqrt{d}}{100q^3}$.

Suppose the linear map $\mathcal{L}^\star : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ *satisfies the following conditions:*

1. $\mathcal{L}^\star$ *satisfies the restricted isometry condition* $\delta_{4r}(\mathcal{L}^\star) \leq \frac{1}{20}$, *where* r *is the rank of each* $X^{(j)\star}$;
2. $\mathcal{L}^\star$ *is normalized and surjective; and*
3. $\|\mathcal{L}^\star\|_2^2 \leq \frac{5q^2}{d}$.

If we supply Algorithm 3 with a normalized initial guess $\mathcal{L}^{(0)} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ *with* $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(0)}) < \frac{1}{20,000q^{7/2}r^2\|\mathcal{L}^\star\|_2^2}$, *then the sequence* $\{\mathcal{L}^{(t)}\}$ *produced by the algorithm satisfies* $\limsup_{t\to\infty} \frac{\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)})}{\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})} \leq 2\|\mathcal{L}^\star\|_2^2 \frac{\Omega(\mathfrak{X}^\star)}{\Lambda(\mathfrak{X}^\star)} + 10q^2\|\mathcal{L}^\star\|_2 \frac{\Delta(\mathfrak{X}^\star)}{\Lambda(\mathfrak{X}^\star)} < 1$. *In other words,* $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)}) \to 0$ *with the rate of convergence bounded above by* $2\|\mathcal{L}^\star\|_2^2 \frac{\Omega(\mathfrak{X}^\star)}{\Lambda(\mathfrak{X}^\star)} + 10q^2\|\mathcal{L}^\star\|_2 \frac{\Delta(\mathfrak{X}^\star)}{\Lambda(\mathfrak{X}^\star)}$. *We assume here that Step 1 of Algorithm 3 is computed via Algorithm 2.*

Remark (i) In this result, the assumption that Step 1 of Algorithm 3 is computed via Algorithm 2 is made for the sake of concreteness. A similar result and proof are possible if Step 1 of Algorithm 3 is instead computed by solving (4) for a suitable choice of the regularization parameter. (ii) In conjunction with Proposition 2, this result implies that we obtain a linear map $\hat{\mathcal{L}}$ upon convergence of our algorithm such that the image of the nuclear norm ball in $\mathbb{R}^{q \times q}$ under $\hat{\mathcal{L}}$ is the same as it is under $\mathcal{L}^\star$.

The proof of this theorem is given in Sect. 3.5. In words, our result states that under a restricted isometry condition on the linear map $\mathcal{L}^\star$ and an isotropy condition on the low-rank matrices $\{X^{(j)\star}\}_{j=1}^n$, Algorithm 3 is locally linearly convergent to the appropriate semidefinite-representable regularizer that promotes the type of structure contained in the data $\{\mathcal{L}^\star\left(X^{(j)\star}\right)\}_{j=1}^n$. The restricted isometry condition on $\mathcal{L}^\star$ ensures that the geometry of the set of points $\{X^{(j)\star}\}_{j=1}^n$ in $\mathbb{R}^{q\times q}$ is (approximately) preserved in the lower-dimensional space $\mathbb{R}^d$. The isotropy condition on the collection $\{X^{(j)\star}\}_{j=1}^n$ ensures that we have observations that lie on most of the low-dimensional faces of the regularizer, which gives us sufficient information to reconstruct the regularizer.

Results of this flavor have previously been obtained in the classical dictionary learning literature [1,4], although our analysis is more challenging in comparison with this prior work for two reasons. First, two nearby sparse vectors with the same number of nonzero entries have the same support, while two nearby low-rank matrices with the same rank have different row/column spaces; geometrically, this translates to the point that two nearby sparse vectors have the same tangent space with respect to a suitably defined variety of sparse vectors, while two nearby low-rank matrices generically have different tangent spaces with respect to an appropriate variety of low-rank matrices. Second (and more significant), the normalization step in classical dictionary learning is simple—corresponding to scaling the columns of a matrix to have unit Euclidean norm, as discussed in Sect. 2.4—while the normalization step in our setting based on Operator Sinkhorn scaling is substantially more complicated. Indeed, one of the key aspects of our analysis is the relation between the stability properties of Operator Sinkhorn scaling and the tangent spaces to varieties of low-rank matrices, as is evident from the appearance of the parameter $\Omega\left(\{X^{(j)\star}\}_{j=1}^n\right)$ in Theorem 3.

The distance measure $\xi_{\mathcal{L}^\star}$ that appears in Theorem 3 is defined up to an equivalence relation, and with respect to the linear map $\mathcal{L}^\star$ to which we do not have access. In practice, it is useful to have a stopping criterion that only depends on the sequence of iterates. To this end, the next result states that under the same conditions as in Theorem 3, the sequence of iterates $\{\mathcal{L}^{(t)}\}$ obtained from our algorithm also converges (the limit point is generically different from $\mathcal{L}^\star$, although they specify the same regularizer):

Proposition 6 *Under the same setup and assumptions as in Theorem 3, the sequence of iterates $\{\mathcal{L}^{(t)}\}$ obtained from our algorithm is a Cauchy sequence.*

This result is proved in "Appendix I".

Extension to the noisy case. In practice, the data points $\mathbf{y}^{(j)}$ may be corrupted by noise, and it is of interest to investigate whether our algorithm is robust to noise. One can extend our analysis to demonstrate the robustness of our algorithm in a stylized setting in which the data points $\mathbf{y}^{(j)}$ in Theorem 3 are corrupted by additive noise. Briefly, such an extension comprises two key steps. First, one can show that there exists a normalized linear map $\check{\mathcal{L}}$ that is close to $\mathcal{L}^\star$ (up to composition by an orthogonal rank-preserver), and which is a fixed point of our algorithm. The key ingredient in

demonstrating this is to prove that each iteration of our algorithm is *contractive* in a neighborhood of $\mathcal{L}^\star$ and to appeal to a suitable fixed-point theorem. The proximity of the regularizer defined by $\check{\mathcal{L}}$ to the regularizer defined by $\mathcal{L}^\star$ is determined by the radius of contraction, which depends linearly (under the conditions of Theorem 3) on the size of the noise corrupting the measurements and inverse polynomially on the size of the data set. Second, one can show that our algorithm is locally linearly convergent to $\check{\mathcal{L}}$ (up to composition by an orthogonal rank-preserver). This step essentially follows the same sequence of arguments as in the proof of Theorem 3, and it relies on the radius of contraction from the first step being smaller than the basin of attraction defined in Theorem 3; this is true as long as the noise corruption is suitably small and the number of data points is sufficiently large.

3.4 Ensembles Satisfying the Conditions of Theorem 3

Theorem 3 gives deterministic conditions on the underlying data under which our algorithm recovers the correct regularizer. In this section, we demonstrate that these conditions are in fact satisfied with high probability by certain natural random ensembles. Our first result states that random Gaussian linear maps upon normalization satisfy the requirements on the linear map in Theorem 3:

Proposition 7 *Let $\tilde{\mathcal{L}} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ be a linear map in which each of the d component linear functionals are specified by matrices $\tilde{\mathcal{L}}_i \in \mathbb{R}^{q \times q}$ with i.i.d random Gaussian entries with mean zero and variance $1/d$. Let $\mathcal{L}$ represent a normalized map obtained by composing $\tilde{\mathcal{L}}$ with a positive-definite rank-preserver. Fix any $\delta < 1$. Then, there exist positive constants c_1, c_2, c_3 depending only on δ such that if $d \geq c_1 r q$, then (i) $\delta_{4r}(\mathcal{L}) \leq \delta$ and (ii) $\|\mathcal{L}\|_2 \leq \sqrt{\frac{5q^2}{d}}$ with probability greater than $1 - c_2 \exp(-c_3 d)$.*

The proof of this result is given in "Appendix D". As shown in [11] random Gaussian linear maps from $\mathbb{R}^{q \times q}$ to $\mathbb{R}^d$ satisfy the restricted isometry property for rank-$4r$ matrices if $d \gtrsim r q$ (and this bound is tight). Our result shows that under the same scaling assumption on d, 'most' linear maps satisfy the more restrictive requirements of Theorem 3. Next we consider families of random low-rank matrices:

Proposition 8 *Let $\mathfrak{X} := \{X^{(j)}\}_{j=1}^n$ be an ensemble of matrices generated as $X^{(j)} = \sum_{i=1}^r s_i^{(j)} \mathbf{u}_i^{(j)} \mathbf{v}_i^{(j)\prime}$ with each $U^{(j)} = [\mathbf{u}_1^{(j)} | \dots | \mathbf{u}_r^{(j)}]$, $V^{(j)} = [\mathbf{v}_1^{(j)} | \dots | \mathbf{v}_r^{(j)}] \in \mathbb{R}^{q \times r}$ being drawn independently from the Haar measure on $q \times r$ matrices with orthonormal columns, and each $s_i^{(j)}$ being drawn independently from $\mathcal{D}$, where $\mathcal{D}$ is any distribution supported on $[s/2, s]$ for some $s > 0$. Then for any $0 < t_1 \leq 1/4$ and $0 < t_2$, the conditions (i) $\frac{\Delta(\mathfrak{X})}{\Lambda(\mathfrak{X})} \leq t_1$ and (ii) $\frac{\Omega(\mathfrak{X})}{\Lambda(\mathfrak{X})} \leq 80\frac{r}{q} + t_2$, are satisfied with probability greater than $1 - 2q \exp(-\frac{nt_1^2}{200q^4}) - q \exp(-\frac{nt_2^2}{200q^4})$. In particular, the requirements in Theorem 3 for $d \gtrsim r q$ are satisfied with high probability by the ensemble $\mathfrak{X}$ provided $n \gtrsim \frac{q^{10}}{d}$.*

Considering the requirements of Theorem 3 in the regime $d \gtrsim r q$ is not restrictive as this condition is necessary for the restricted isometry assumptions of Theorem 3 on $\mathcal{L}^\star$

to hold. The proof of this result is given in "Appendix B". Thus, in some sense, 'most' (sufficiently large) sets of low-rank matrices satisfy the requirements of Theorem 3. We also note that for a collection of low-rank matrices $\mathfrak{X}$ generated according to the ensemble in this proposition, the ratio $\frac{\Delta(\mathfrak{X})}{\Lambda(\mathfrak{X})} \to 0$ as $n \to \infty$, while one can show that the ratio $\frac{\Omega(\mathfrak{X})}{\Lambda(\mathfrak{X})} \asymp \frac{r}{q}$ as $n \to \infty$. Based on Theorem 3, this observation implies that for data generated according to the ensemble in Proposition 8, the rate of convergence of Algorithm 3 improves with an increase in the amount of data, but only up to a certain point beyond which the convergence rate plateaus. We illustrate this property with a numerical experiment in Sect. 4.1.

Remark It is critical in the preceding result that we project onto the orthogonal complement of the subspace $\mathcal{W}$ from (14) in the definition of $\Omega(\mathfrak{X})$. For a set of low-rank matrices $\mathfrak{X}$ drawn from the same ensemble as in Proposition 8, one can show that $\|\mathcal{P}_{\mathcal{S}^\perp} \circ \frac{1}{n} \sum_{j=1}^{n} (X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})}\|_2 > c\Lambda(\mathfrak{X})$ for a constant $c > 0$ with high probability, where the subspace $\mathcal{S}$ is defined in (15). In the context of the discussion at the end of the preceding section, we have that the conditions of Theorem 3 are violated if we do not incorporate the normalization step via Operator Sinkhorn scaling, which in turn impacts the convergence of our algorithm.

3.5 Proof of Theorem 3

Before giving a proof of Theorem 3, we state two relevant lemmas that are proved in "Appendix A".

Lemma 1 *Suppose a linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ satisfies the restricted isometry condition $\delta_{2r}(\mathcal{L}) < 1$. For any $\mathcal{T} := \mathcal{T}(X)$ with $X \in \mathbb{R}^{q \times q}$ and $\mathrm{rank}(X) \leq r$, we have that (i) $1 - \delta_{2r} \leq \lambda_{\min}(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}}) \leq \lambda_{\max}(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}}) \leq 1 + \delta_{2r}$, (ii) $\|(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}})^{-1}\|_2 = \|[(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}})^{-1}]_{\mathbb{R}^{q \times q}}\|_2 \leq \frac{1}{1 - \delta_{2r}}$, (iii) $\|\mathcal{P}_{\mathcal{T}} \circ \mathcal{L}'\mathcal{L}\|_2 \leq \sqrt{1 + \delta_{2r}}\|\mathcal{L}\|_2$, and (iv) $\|[(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}})^{-1}]_{\mathbb{R}^{q \times q}} \circ \mathcal{L}'\mathcal{L}\|_2 \leq \frac{\sqrt{1 + \delta_{2r}}}{1 - \delta_{2r}}\|\mathcal{L}\|_2$. Here $\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}} : \mathcal{T} \to \mathcal{T}$ is a self-adjoint linear map.*

Lemma 2 *Let $\mathfrak{X} := \{X^{(j)}\}_{j=1}^{n} \subset \mathbb{R}^{q \times q}$ be a collection of matrices, and let $s_{\min} := \min_j \|X^{(j)}\|_{\ell_2}^2$ and $s_{\max} := \max_j \|X^{(j)}\|_{\ell_2}^2$. Then, $s_{\min}/q^2 - \Delta(\mathfrak{X}) \leq \Lambda(\mathfrak{X}) \leq s_{\max}/q^2 + \Delta(\mathfrak{X})$.*

Proof (Theorem 3) To simplify the presentation of our proof, we define the following quantities $\alpha_0 := 20{,}000q^{7/2}r^2\|\mathcal{L}^\star\|_2^2$, $\alpha_1 := 800r^{5/2}\|\mathcal{L}^\star\|_2^2$, $\alpha_2 := 2\sqrt{r}\|\mathcal{L}^\star\|_2$, $\alpha_3 := 10q^2\|\mathcal{L}^\star\|_2$, $\alpha_4 := 5(q^2/\sqrt{r})\alpha_1$, $\alpha_5 := 100q^3\alpha_2^2$, $\alpha_6 := 5(q^2/\sqrt{r})\alpha_2$, and $\alpha_7 := \alpha_3 + \alpha_6/6 + 1/4$. The specific interpretation of these quantities is not essential to the proof—the pertinent detail is that they only depend on q, r, $\|\mathcal{L}^\star\|_2$.

To simplify notation in the proof, we denote $\Delta := \Delta(\mathfrak{X})$, $\Lambda := \Lambda(\mathfrak{X})$, and $\Omega := \Omega(\mathfrak{X})$. In addition, we also denote $\mathcal{T}^{(j)} := \mathcal{T}(X^{(j)\star})$. Our proof proceeds by establishing the following assertion. Suppose that the tth iterate $\mathcal{L}^{(t)}$ is such that $\mathcal{L}^{(t)} = \mathcal{L}^\star \circ (I + E^{(t)}) \circ M^{(t)}$, where $M^{(t)}$ is a rank-preserver and $E^{(t)}$ is a linear operator that satisfies $\|E^{(t)}\|_{\ell_2} < 1/\alpha_0$. Then, the $t + 1$th iterate is of the form

$\mathcal{L}^{(t+1)} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t+1)}) \circ \mathsf{M}^{(t+1)}$ for some rank-preserver $\mathsf{M}^{(t+1)}$, and some linear operator $\mathsf{E}^{(t+1)}$ that satisfies

$$\|\mathsf{E}^{(t+1)}\|_{\ell_2} \le \gamma_0 \|\mathsf{E}^{(t)}\|_{\ell_2} + \gamma_1 \|\mathsf{E}^{(t)}\|_{\ell_2}^2, \tag{16}$$

where $\gamma_0 = 2\|\mathcal{L}^\star\|_2^2 (\Omega/\Lambda) + \alpha_6 (\Delta/\Lambda)$, and $\gamma_1 = \alpha_4 + \alpha_5 + 5\alpha_7^2/\sqrt{q}$.

Before we prove this assertion, we note how it allows us to conclude the result. By taking the infimum over $\mathsf{E}^{(t)}$ on the right- hand side of (16) and by noting that $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)}) \le \|\mathsf{E}^{(t+1)}\|_{\ell_2}$, we have

$$\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)}) \le \gamma_0 \xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)}) + \gamma_1 \xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})^2. \tag{17}$$

One can check based on the initial assumption on $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(0)})$ that $\gamma := \gamma_0 + \gamma_1 \xi_{\mathcal{L}^\star}(\mathcal{L}^{(0)}) < 1$. By employing an inductive argument, one can establish that $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)}) \le \gamma \xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})$. Thus, $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)}) \le \gamma^t \xi_{\mathcal{L}^\star}(\mathcal{L}^{(0)}) \to 0$ as $t \to \infty$. By dividing the expression in (17) throughout by $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})$ and subsequently taking the limit $t \to \infty$, we obtain the asymptotic rate of convergence

$$\limsup_{t\to\infty} \frac{\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)})}{\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})} \le \limsup_{t\to\infty} \left(\gamma_0 + \gamma_1 \xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)}) \right) = \gamma_0.$$

We proceed to prove the assertion.

[Applying Proposition 3]: Since $\|\mathsf{E}^{(t)}\|_{\ell_2} \le \min\{1/(50\sqrt{q}), 1/(120r^2\|\mathcal{L}^\star\|_2)\}$, by applying Proposition 3 with the choice of $X^\star = X^{(j)\star}$, $\mathsf{E} = \mathsf{E}^{(t)}$, $\mathsf{M} = \mathsf{M}^{(t)}$, and $\mathcal{L}^\star$, we have for each $j = 1, \ldots, n$ that

$$M^{(t)}(\hat{X}^{(j)}) - X^{(j)\star} = -\left[[(\mathcal{L}_{\mathcal{T}^{(j)}}^{\star\prime} \mathcal{L}_{\mathcal{T}^{(j)}}^\star)^{-1}]_{\mathbb{R}^{q\times q}} \circ \mathcal{L}^{\star\prime} \mathcal{L}^\star \circ \mathsf{E}^{(t)} \right](X^{(j)\star}) + G^{(j)}, \tag{18}$$

where $G^{(j)}$ is a matrix that satisfies $\|G^{(j)}\|_{\ell_2} \le \alpha_1 \|X^{(j)\star}\|_2 \|\mathsf{E}^{(t)}\|_{\ell_2}^2$.

[Applying Proposition 4]: The next step is to apply Proposition 4 to the collections of matrices $\{X^{(j)\star}\}_{j=1}^n$ and $\{\hat{X}^{(j)}\}_{j=1}^n$. Let $\mathbf{x}^\star, \hat{\mathbf{x}}$ denote the linear maps $\mathbf{x}^\star : \mathbf{z} \mapsto \sum_{j=1}^n X^{(j)\star} \mathbf{z}_j$, $\hat{\mathbf{x}} : \mathbf{z} \mapsto \sum_{j=1}^n \hat{X}^{(j)} \mathbf{z}_j$. First note that $\alpha_1 \|\mathsf{E}^{(t)}\|_{\ell_2} \le \alpha_1/\alpha_0 \le \sqrt{r}\|\mathcal{L}^\star\|_2$. Second from the assumptions, we have $\Delta/\Lambda \le 1/21$. Hence by Lemma 2, we have $\Lambda \le s^2 r/q^2 + \Delta \le s^2 r/q^2 + \Lambda/21$. It follows that $\Delta \le s^2 r/(20q^2)$, and thus by Lemma 2, we have $\Lambda \ge s^2 r/(5q^2)$. Third by applying these inequalities and Lemma 1 to (18), we have $\|M^{(t)}(\hat{X}^{(j)}) - X^{(j)\star}\|_{\ell_2} \le ((\sqrt{1+\delta_{4r}})/(1-\delta_{4r}))\|\mathcal{L}^\star\|_2 \|X^{(j)\star}\|_2 \|\mathsf{E}^{(t)}\|_{\ell_2} + \alpha_1 \|X^{(j)\star}\|_2 \|\mathsf{E}^{(t)}\|_{\ell_2}^2 \le s\alpha_2/\alpha_0 \le s\alpha_2\|\mathsf{E}^{(t)}\|_{\ell_2} \le \sqrt{\Lambda}/20$. Fourth note that the assumptions imply $\Delta/\Lambda \le 1/6$. Hence by Proposition 4 applied to $\{X^{(j)\star}\}_{j=1}^n$ and $\{\hat{X}^{(j)}\}_{j=1}^n$ with the choice of $\mathsf{Q} = M^{(t)}$, we have

$$\mathbf{x}^\star \circ \hat{\mathbf{x}}^+ = (\mathsf{I} + \mathsf{D}) \circ M^{(t)},$$

where

$$D := \frac{1}{n\Lambda} \sum_{j=1}^{n} ([[(\mathcal{L}_{\mathcal{T}^{(j)}}^{\star\prime} \mathcal{L}_{\mathcal{T}^{(j)}}^{\star})^{-1}]_{\mathbb{R}^{q \times q}} \circ \mathcal{L}^{\star\prime} \mathcal{L}^{\star} \circ \mathsf{E}^{(t)}] \left(X^{(j)^{\star}} \right)) \boxtimes X^{(j)^{\star}}$$

$$- \frac{1}{n\Lambda} \sum_{j=1}^{n} G^{(j)} \boxtimes X^{(j)^{\star}} + \mathsf{F},$$

and

$$\|\mathsf{F}\|_{\ell_2} \le 20q(s\alpha_2\|\mathsf{E}^{(t)}\|_{\ell_2})^2/\Lambda + 2q(s\alpha_2\|\mathsf{E}^{(t)}\|_{\ell_2})\Delta/\Lambda^{3/2}$$
$$\le \alpha_5\|\mathsf{E}^{(t)}\|_{\ell_2}^2 + \alpha_6(\Delta/\Lambda)\|\mathsf{E}^{(t)}\|_{\ell_2}. \tag{19}$$

[Applying Proposition 5]: We proceed to bound $\|D\|_{\ell_2}$. Given a collection $\{A^{(j)}\}_{j=1}^n, \{B^{(j)}\}_{j=1}^n \subset \mathbb{R}^{q \times q}$ one has $\frac{1}{n}\|\sum_{j=1} A^{(j)} \boxtimes B^{(j)}\|_{\ell_2} \le \max_j \|A^{(j)} \boxtimes B^{(j)}\|_{\ell_2} = \max_j \|A^{(j)}\|_{\ell_2}\|B^{(j)}\|_{\ell_2}$. By combining this inequality with Lemma 1, we obtain the bounds

$$\frac{1}{n\Lambda} \left\| \sum_{j=1}^{n} ([[(\mathcal{L}_{\mathcal{T}^{(j)}}^{\star\prime} \mathcal{L}_{\mathcal{T}^{(j)}}^{\star})^{-1}]_{\mathbb{R}^{q \times q}} \circ \mathcal{L}^{\star\prime} \mathcal{L}^{\star} \circ \mathsf{E}^{(t)}] \left(X^{(j)^{\star}} \right)) \boxtimes X^{(j)^{\star}} \right\|_{\ell_2}$$

$$\le \left(2s^2 r\|\mathcal{L}^{\star}\|_2/\Lambda \right) \|\mathsf{E}^{(t)}\|_{\ell_2} \le \alpha_3 \|\mathsf{E}^{(t)}\|_{\ell_2}, \tag{20}$$

and

$$(1/n\Lambda) \left\| \sum_{j=1}^{n} G^{(j)} \boxtimes X^{(j)^{\star}} \right\|_{\ell_2} \le (\alpha_1 s^2 \sqrt{r}/\Lambda)\|\mathsf{E}^{(t)}\|_{\ell_2}^2 \le \alpha_4\|\mathsf{E}^{(t)}\|_{\ell_2}^2. \tag{21}$$

Hence by combining (19), (20), and (21), we have $\|D\|_{\ell_2} \le \alpha_3\|\mathsf{E}^{(t)}\|_{\ell_2} + \alpha_4\|\mathsf{E}^{(t)}\|_{\ell_2}^2 + \alpha_5\|\mathsf{E}^{(t)}\|_{\ell_2}^2 + \alpha_6(\Delta/\Lambda)\|\mathsf{E}^{(t)}\|_{\ell_2} \le \alpha_7\|\mathsf{E}^{(t)}\|_{\ell_2} \le \alpha_7/\alpha_0 \le 1/10$. Consequently, by applying Proposition 5 with this choice of D, we have

$$\mathbf{x}^{\star} \circ \hat{\mathbf{x}}^+ = (\mathsf{I} + \mathcal{P}_{\mathcal{W}^\perp}(\mathsf{D}) + \mathsf{H}) \circ W \circ M^{(t)}, \quad \|\mathsf{H}\|_{\ell_2} \le (5\alpha_7^2/\sqrt{q})\|\mathsf{E}^{(t)}\|_{\ell_2}^2, \tag{22}$$

for some rank-preserver W.

[Conclusion]: Recall from the description of the algorithm that the next iterate is given by $\mathcal{L}^{(t+1)} = \mathcal{L}^{\star} \circ \mathbf{x}^{\star} \circ \hat{\mathbf{x}}^+ \circ N_{\mathcal{L}^{\star} \circ \mathbf{x}^{\star} \circ \hat{\mathbf{x}}^+}$, where $N_{\mathcal{L}^{\star} \circ \mathbf{x}^{\star} \circ \hat{\mathbf{x}}^+}$ is the unique positive-definite rank-preserver that normalizes $\mathcal{L}^{\star} \circ \mathbf{x}^{\star} \circ \hat{\mathbf{x}}^+$. We define $\mathsf{E}^{(t+1)} := \mathcal{P}_{\mathcal{W}^\perp}(\mathsf{D}) + \mathsf{H}$, and hence

$$\mathcal{L}^{(t+1)} = \mathcal{L}^{\star} \circ (\mathsf{I} + \mathsf{E}^{(t+1)}) \circ M^{(t+1)}, \tag{23}$$

where $M^{(t+1)} = W \circ M^{(t)} \circ N_{\mathcal{L}^{\star} \circ \mathbf{x}^{\star} \circ \hat{\mathbf{x}}^+}$ is a composition of rank-preservers and hence is also a rank-preserver. It remains to bound $\|\mathsf{E}^{(t+1)}\|_{\ell_2}$.

As $\|[(\mathcal{L}_{\mathcal{T}^{(j)}}^{\star\prime}\mathcal{L}_{\mathcal{T}^{(j)}}^{\star})^{-1}]_{\mathbb{R}^{q\times q}}\|_2 \leq 2$ from Lemma 1, we have $[(\mathcal{L}_{\mathcal{T}^{(j)}}^{\star\prime}\mathcal{L}_{\mathcal{T}^{(j)}}^{\star})^{-1}]_{\mathbb{R}^{q\times q}} \preceq 2\mathcal{P}_{\mathcal{T}^{(j)}}$, and hence $(X^{(j)\star}\boxtimes X^{(j)\star})\otimes[(\mathcal{L}_{\mathcal{T}^{(j)}}^{\star\prime}\mathcal{L}_{\mathcal{T}^{(j)}}^{\star})^{-1}]_{\mathbb{R}^{q\times q}} \preceq 2(X^{(j)\star}\boxtimes X^{(j)\star})\otimes\mathcal{P}_{\mathcal{T}^{(j)}}$. Moreover, since $(X^{(j)\star}\boxtimes X^{(j)\star})\otimes[(\mathcal{L}_{\mathcal{T}^{(j)}}^{\star\prime}\mathcal{L}_{\mathcal{T}^{(j)}}^{\star})^{-1}]_{\mathbb{R}^{q\times q}}$ and $2(X^{(j)\star}\boxtimes X^{(j)\star})\otimes\mathcal{P}_{\mathcal{T}^{(j)}}$ are Kronecker products of positive-semidefinite operators, they too are positive-semidefinite operators, and hence $\mathcal{P}_{\mathcal{W}^\perp} \circ (\frac{1}{n}\sum_{j=1}^n(X^{(j)\star}\boxtimes X^{(j)\star})\otimes[(\mathcal{L}_{\mathcal{T}^{(j)}}^{\star\prime}\mathcal{L}_{\mathcal{T}^{(j)}}^{\star})^{-1}]_{\mathbb{R}^{q\times q}})^2 \circ \mathcal{P}_{\mathcal{W}^\perp} \preceq \mathcal{P}_{\mathcal{W}^\perp} \circ (\frac{2}{n}\sum_{j=1}^n(X^{(j)\star}\boxtimes X^{(j)\star})\otimes\mathcal{P}_{\mathcal{T}^{(j)}})^2 \circ \mathcal{P}_{\mathcal{W}^\perp}$. This implies the bound

$$2\Omega \geq \left\| \mathcal{P}_{\mathcal{W}^\perp} \circ \left(\frac{1}{n} \sum_{j=1}^n (X^{(j)\star} \boxtimes X^{(j)\star}) \otimes [(\mathcal{L}_{\mathcal{T}^{(j)}}^{\star\prime}\mathcal{L}_{\mathcal{T}^{(j)}}^{\star})^{-1}]_{\mathbb{R}^{q\times q}} \right) \right\|_2.$$

Combining this bound with the identity $\mathsf{L}(X_1) \boxtimes X_2 = \mathsf{L} \circ (X_1 \boxtimes X_2)$, we obtain

$$\frac{1}{n\Lambda} \left\| \mathcal{P}_{\mathcal{W}^\perp} \left(\sum_{j=1}^n \left(\left[(\mathcal{L}_{\mathcal{T}^{(j)}}^{\star\prime}\mathcal{L}_{\mathcal{T}^{(j)}}^{\star})^{-1} \right]_{\mathbb{R}^{q\times q}} \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}^{(t)} \right] (X^{(j)\star}) \right) \boxtimes X^{(j)\star} \right) \right\|_{\ell_2}$$

$$= \frac{1}{n\Lambda} \left\| \left[\mathcal{P}_{\mathcal{W}^\perp} \circ \left(\sum_{j=1}^n (X^{(j)\star} \boxtimes X^{(j)\star}) \otimes [(\mathcal{L}_{\mathcal{T}^{(j)}}^{\star\prime}\mathcal{L}_{\mathcal{T}^{(j)}}^{\star})^{-1}]_{\mathbb{R}^{q\times q}} \right) \right] (\mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}^{(t)}) \right\|_{\ell_2}$$

$$\leq (2\Omega/\Lambda)\|\mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}^{(t)}\|_{\ell_2} \leq (2\Omega/\Lambda)\|\mathcal{L}^\star\|_2^2\|\mathsf{E}^{(t)}\|_{\ell_2}. \tag{24}$$

From the definition of $\mathsf{E}^{(t+1)}$, we have the relation

$$\mathsf{E}^{(t+1)} = \mathcal{P}_{\mathcal{W}^\perp} \left(\frac{1}{n\Lambda} \sum_{j=1}^n [[(\mathcal{L}_{\mathcal{T}^{(j)}}^{\star\prime}\mathcal{L}_{\mathcal{T}^{(j)}}^{\star})^{-1}]_{\mathbb{R}^{q\times q}} \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}^{(t)}](X^{(j)\star}) \boxtimes X^{(j)\star} \right.$$

$$\left. + \frac{1}{n\Lambda} \sum_{j=1}^n G^{(j)} \boxtimes X^{(j)\star} + \mathsf{F} \right) + \mathsf{H}. \tag{25}$$

Since $\mathcal{P}_{\mathcal{W}^\perp}$ defines a projection, we have $(1/n\Lambda)\|\mathcal{P}_{\mathcal{W}^\perp}(\sum_{j=1}^n G^{(j)} \boxtimes X^{(j)\star})\|_{\ell_2} \leq (1/n\Lambda)\|\sum_{j=1}^n G^{(j)} \boxtimes X^{(j)\star}\|_{\ell_2}$, and $\|\mathcal{P}_{\mathcal{W}^\perp}(\mathsf{F})\|_{\ell_2} \leq \|\mathsf{F}\|_{\ell_2}$. Hence, by applying bounds (19), (21), (22), and (24) to (25), we obtain

$$\|\mathsf{E}^{(t+1)}\|_{\ell_2} \leq ((2\Omega/\Lambda)\|\mathcal{L}^\star\|_2^2 + \alpha_6(\Delta/\Lambda))\|\mathsf{E}^{(t)}\|_{\ell_2} + (\alpha_4 + \alpha_5 + 5\alpha_7^2/\sqrt{q})\|\mathsf{E}^{(t)}\|_{\ell_2}^2$$

$$= \gamma_0\|\mathsf{E}^{(t)}\|_{\ell_2} + \gamma_1\|\mathsf{E}^{(t)}\|_{\ell_2}^2.$$

This completes the proof. $\qquad\square$

4 Numerical Experiments

4.1 Illustration with Synthetic Data

We begin with a demonstration of the utility of our algorithm in recovering a regularizer from synthetic data. Our experiment qualitatively confirms the predictions of Theorem 3 regarding the rate of convergence.

Setup. We generate a standard Gaussian linear map $\mathcal{L} : \mathbb{R}^{7 \times 7} \to \mathbb{R}^{30}$ and we normalize it; denote the normalized version as $\mathcal{L}^\star$. We generate data $\{\mathbf{y}^{(j)}\}_{j=1}^{1000}$ as $\mathbf{y}^{(j)} = \mathcal{L}^\star(\mathbf{u}^{(j)} \mathbf{v}^{(j)'})$, where each $\mathbf{u}^{(j)}, \mathbf{v}^{(j)}$ is drawn independently from the Haar measure on the unit sphere in $\mathbb{R}^7$. We generate standard Gaussian maps $\mathcal{E}^{(i)} : \mathbb{R}^{7 \times 7} \to \mathbb{R}^{30}$, $i = 1, \ldots, 20$ that are used to corrupt $\mathcal{L}^\star$ in providing the initial guess to our algorithm. Specifically, for each $\sigma \in \{0.125, 0.25, \ldots, 2.5\}$ and each $\mathcal{E}^{(i)}$, $i = 1, \ldots, 20$ we supply as initial guess to our algorithm the normalized version of $\mathcal{L}^\star + \sigma \mathcal{E}^{(i)}$. In addition, we supply the subset $\{\mathbf{y}^{(j)}\}_{j=1}^m$ for each $m \in \{50, 100, \ldots, 1000\}$ to our algorithm. The objective of this experiment is to investigate the role of the number of data points (denoted by m) and the size of the error in the initial guess (denoted by σ) on the performance of our algorithm.

Characterizing recovery of correct regularizer. Before discussing the results, we describe a technique assessing whether our algorithm recovers the correct regularizer. In particular, as we do not know of a tractable technique for computing the distance measure ξ between two linear maps (8), we consider an alternative approach for computing the 'distance' between two linear maps. For linear maps from $\mathbb{R}^{q \times q}$ to $\mathbb{R}^d$, we fix a set of unit Euclidean norm rank-one matrices $\{\mathbf{s}^{(k)} \mathbf{t}^{(k)'}\}_{k=1}^\ell$, where each $\mathbf{s}^{(k)}, \mathbf{t}^{(k)} \in \mathbb{R}^q$ is drawn uniformly from the Haar measure on the sphere and ℓ is chosen to be larger than q^2. Given an estimate $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ of a linear map $\mathcal{L}^\star : \mathbb{R}^{q \times q} \to \mathbb{R}^d$, we compute the following

$$\text{dist}_{\mathcal{L}^\star}(\mathcal{L}) := \frac{1}{\ell} \sum_{k=1}^{\ell} \inf_{\substack{X \in \mathbb{R}^{q \times q} \\ \text{rank}(X) \leq 1}} \left\| \mathcal{L}^\star \left(\mathbf{s}^{(k)} \mathbf{t}^{(k)'} \right) - \mathcal{L}(X) \right\|_{\ell_2}^2 . \tag{26}$$

To compute the minimum for each term in the sum, we employ the heuristic described in Algorithm 2. If $\mathcal{L}^\star$ satisfies a suitable restricted isometry condition for rank-one matrices and if $\mathcal{L}$ is specified as $\mathcal{L}^\star$ composed with a near-orthogonal rank-preserver, then we have that $\text{dist}_{\mathcal{L}^\star}(\mathcal{L}) \approx 0$; in the opposite direction, as $\ell > q^2$, we have that $\text{dist}_{\mathcal{L}^\star}(\mathcal{L}) \approx 0$ implies $\xi_{\mathcal{L}^\star}(\mathcal{L}) \approx 0$. In our setting with $q = 7$, we set $\ell = 100$. If our algorithm provides an estimate $\mathcal{L}$ such that $\text{dist}_{\mathcal{L}^\star}(\mathcal{L}) < 10^{-3}$, then we declare that our method has succeeded in recovering the correct regularizer.

Results. In Fig. 2, we plot for each $\sigma \in \{0.125, 0.25, \ldots, 2.5\}$ the average number of iterations—taken over the 20 different initial guesses specified by the normalized versions of $\mathcal{L}^\star + \sigma \mathcal{E}^{(i)}$, $i = 1, \ldots, 20$—required for Algorithm 3 [with Step 1 computed by solving (5) via Algorithm 2] to succeed in recovering the correct regularizer as a function of the number of data points m supplied as input. The different curves in the figure correspond to different noise levels (specified by σ) in the initial guess;

Fig. 2 Average number of iterations required to identify correct regularizer as a function of the number of observations; each line represents a fixed noise level σ denoting the amount of corruption in the initial guess (see Sect. 4.1 for details of the experimental setup)
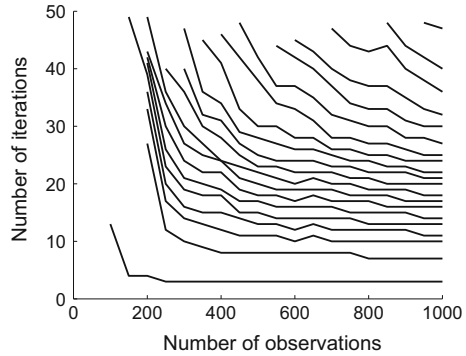
that is, the curves higher up in the figure are associated with larger noise levels. There are two main conclusions to be drawn from this result. First, the average number of iterations grows as the initial guess is of increasingly poorer quality. Second, and more interesting, is that the number of iterations required for convergence improves with an increase in the number of input data points, but only up to a certain stage beyond which the convergence rate seems to plateau (this is a feature at every noise level in this plot). This observation confirms the predictions of Theorem 3 and of Proposition 8 (specifically, see the discussion immediately following this proposition).

4.2 Illustration with Natural Images

4.2.1 Representing Natural Image Patches

The first stage of this experiment contrasts projections of low-rank matrices and projections of sparse vectors purely from the perspective of representing a collection of image patches.

Setup. We consider a data set $\{\mathbf{y}^{(j)}\}_{j=1}^{6480} \in \mathbb{R}^{64}$ of image patches. These data are obtained by taking 8×8 patches from larger images of seagulls and considering these patches as well as their rotations, as is common in the dictionary learning literature; Fig. 3 gives an example of a seagull image as well as several smaller patches. To ensure that we learned a centered and suitably isotropic norm, we center the entire data set to ensure that the average of the $\mathbf{y}^{(j)}$'s is the origin and then scale each data point so that it has unit Euclidean norm. We apply Algorithm 3 (with Step 1 computed by solving (5) via Algorithm 2) and the analog of this procedure for dictionary learning described in Sect. 2.4. We assess the quality of the description of the data set $\{\mathbf{y}^{(j)}\}_{j=1}^{6480}$ as a projection of low-rank matrices (obtained using our approach) as opposed to a projection of sparse vectors (obtained using dictionary learning).

Representation complexity. To assess the performance of each representation framework, we require a characterization of the number of parameters needed to specify an image patch in each representation as well as the resulting quality of approximation. Given a collection $\{\mathbf{y}^{(j)}\}_{j=1}^{n} \subset \mathbb{R}^{d}$, suppose we represent each point as $\mathbf{y}^{(j)} \approx \mathcal{L}(X^{(j)})$ for a linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^{d}$ and a rank-r matrix $X^{(j)} \in \mathbb{R}^{q \times q}$.
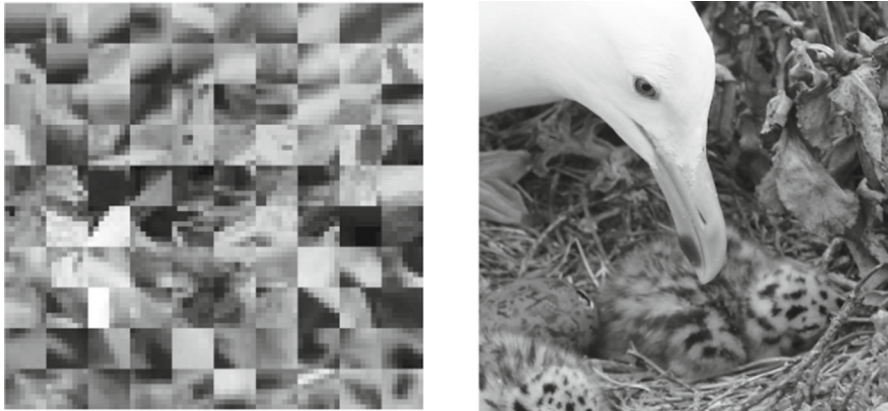
Fig. 3 Image patches (left) obtained from larger raw images (sample on the right)

The number of parameters required to specify each $X^{(j)}$ is $2qr - r^2$, and the number of parameters required to specify $\mathcal{L}$ is dq^2. Consequently, the average number of parameters required to specify each $\mathbf{y}^{(j)}$ is $2qr - r^2 + \frac{dq^2}{n}$. In a similar manner, if each $\mathbf{y}^{(j)} \approx L\mathbf{x}^{(j)}$ for a linear map $L : \mathbb{R}^p \times \mathbb{R}^d$ and a vector $\mathbf{x}^{(j)} \in \mathbb{R}^p$ with s nonzero coordinates, the average number of parameters required to each $\mathbf{y}^{(j)}$ is $2s + \frac{dp}{n}$. In each case, we assess the quality of the approximation by considering the average squared error over the entire set $\{\mathbf{y}^{(j)}\}_{j=1}^n$.

Results. We initialize both our algorithm and the dictionary learning method with random linear maps (suitably normalized in each case). Before contrasting the two approaches, we highlight the improvement in performance our method provides over a pure random linear map. Specifically, Fig. 4 shows for several random initializations that our algorithm (as well as the alternating update method in dictionary learning) provides a significant refinement in approximation quality as the number of iterations increases. Therefore, there is certainly value in employing our algorithm (even with a random initialization) to obtain better representations than pure random projections of low-rank matrices. Next we proceed to a detailed comparison of the two representation frameworks. We employ our approach to learn a representation of the image patch data set with $q \in \{9, 10, \ldots, 15\}$ and the values of the rank r chosen so that the overall representation complexity lies in the range [17, 33]. Similarly, we employ dictionary learning with $p \in \{100, 200, \ldots, 1400\}$ and the values of the sparsity level s chosen so that the overall representation complexity lies in the range [17, 33]. The left subplot in Fig. 6 gives a comparison of these two frameworks. (To interpret the y-axis of the plot, note that the each data point is scaled to have unit norm.) Our approach provides an improvement over dictionary learning for small levels of representation complexity and is comparable at larger levels.

Comparison of atoms. Figure 5 gives an illustration of the atoms obtained from classical dictionary learning (i.e., learning a polyhedral regularizer) as well as those learned using our approach. The left subplot shows the finite collection of atoms of a polyhedral regularizer (corresponding to the finite number of extreme points), and
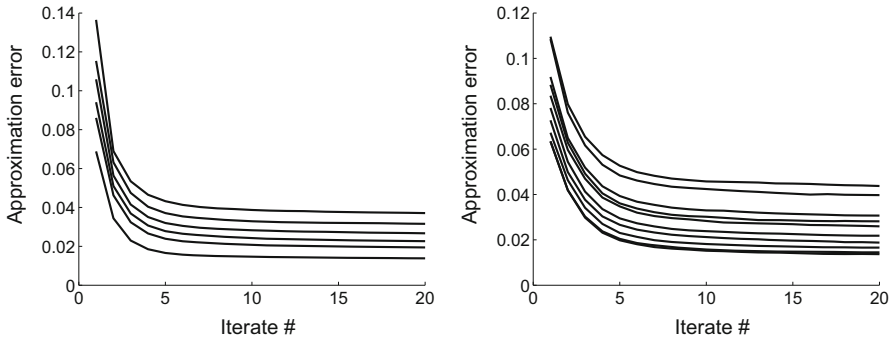
Fig. 4 Progression in mean-squared error with increasing number of iterations with random initializations for learning a semidefinite regularizer (left) and a polyhedral regularizer (right)
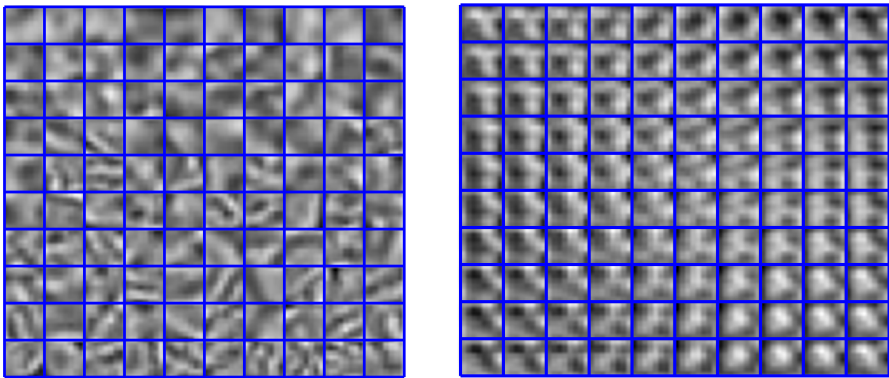

Fig. 5 Comparison between atoms learned from dictionary learning (left) and our algorithm (right)

the right subplot shows a finite subset of the infinite collection of atoms learned using our approach. The individual atoms in each case generally correspond to piecewise smooth regions separated by boundaries. However, the geometry of the *collection* of atoms is distinctly different in the two cases; in particular, the atoms learned using our approach better represent the transformations underlying natural images. As we discuss in the next set of experiments, our framework provides regularizers that lead to improved denoising performance on natural images in comparison with polyhedral regularizers.

4.2.2 Denoising Natural Image Patches

We compare the performance of polyhedral and semidefinite regularizers in denoising natural image patches corrupted by noise.

Setup. The 6480 data points from the previous experiment are designated as a training set. Here we consider an additional collection $\{\mathbf{y}_{\text{test}}^{(j)}\}_{j=1}^{720} \subset \mathbb{R}^{64}$ of 8×8 test image patches obtained from larger seagull images (as with the training set) and subsequently shifted by an average of the pre-centered training set. We corrupt each of

these test points by i.i.d. Gaussian noise to obtain $\mathbf{y}_{\text{obs}}^{(j)} = \mathbf{y}_{\text{test}}^{(j)} + \mathbf{w}^{(j)}, \ j = 1, \ldots, 720,$ where each $\mathbf{w}^{(j)} \sim \mathcal{N}(0, \sigma^2 I)$ with σ^2 chosen so that the average signal-to-noise ratio $\frac{1}{720} \sum_{j=1}^{n} \frac{\|\mathbf{y}_{\text{test}}^{(j)}\|_{\ell_2}^2}{64\sigma^2} \approx 18.$ Our objective is to investigate the denoising performance of the polyhedral and semidefinite regularizers (learned on the training set) on the data set $\{\mathbf{y}_{\text{obs}}^{(j)}\}_{j=1}^{720}$. Specifically, we analyze the following proximal denoising procedure:

$$\hat{\mathbf{y}}_{\text{denoise}} = \arg\min_{\mathbf{y} \in \mathbb{R}^{64}} \ \frac{1}{2}\|\mathbf{y}_{\text{obs}} - \mathbf{y}\|_{\ell_2}^2 + \lambda\|\mathbf{y}\|, \tag{27}$$

where $\|\cdot\|$ is a regularizer learned on the training set and $\lambda > 0$ is a regularization parameter.

Computational complexity of regularizer. To compare the performances of different regularizers, it is instructive to consider the cost associated with employing a regularizer for denoising. In particular, the regularizers learned on the training set have unit balls that are specified as linear images of the nuclear norm ball and the ℓ_1 ball. Consequently, the main cost associated with employing a regularizer is the computational complexity of solving the corresponding proximal denoising problem (27). Thus, we analyze the normalized mean-squared denoising error $\frac{1}{720} \sum_{j=1}^{n} \frac{\|\mathbf{y}_{\text{obs}}^{(j)} - \mathbf{y}_{\text{denoise}}^{(j)}\|_{\ell_2}^2}{64\sigma^2}$ of a regularizer as a function of the computational complexity of solving (27). For a polyhedral norm $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$ with unit ball specified as the image under a linear map $L : \mathbb{R}^p \to \mathbb{R}^d$ of the ℓ_1 ball in $\mathbb{R}^p$, we solve (27) as follows by representing the norm $\|\cdot\|$ in a lifted manner:

$$\hat{\mathbf{y}}_{\text{denoise}} = \arg\min_{\substack{\mathbf{x}, \mathbf{z} \in \mathbb{R}^p \\ s, t \in \mathbb{R}}} \ \frac{1}{2}s + \lambda t$$
$$\text{s.t.} \ \ \|\mathbf{y}_{\text{obs}} - L\mathbf{x}\|_{\ell_2}^2 \le s, \quad \sum_{i=1}^{p} \mathbf{z}_i \le t, \quad \begin{pmatrix} \mathbf{z} - \mathbf{x} \\ \mathbf{z} + \mathbf{x} \end{pmatrix} \ge 0. \tag{28}$$

To solve (28) to an accuracy ϵ using an interior-point method with the usual logarithmic barriers for the nonnegative orthant and the second-order cone, we have that the number of operations required is $\sqrt{2p + 2} \log(\frac{2p+2}{\epsilon\eta}((d + 2p + 2)^3 + (2p + 2)^3))$—this represents the number of outer loop iterations of the interior-point method—multiplied by $(d+2q+2)^3 + (2q+2)^3$—this represents the number of operations required to solve the associated linear system in the inner loop—for a barrier parameter η [45,51]. In a similar manner, for a semidefinite regularizer $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$ with unit ball specified as the image under a linear map $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ of the nuclear norm ball in $\mathbb{R}^{q \times q}$, we again solve (27) as follows by representing the norm $\|\cdot\|$ in an analogous lifted manner:
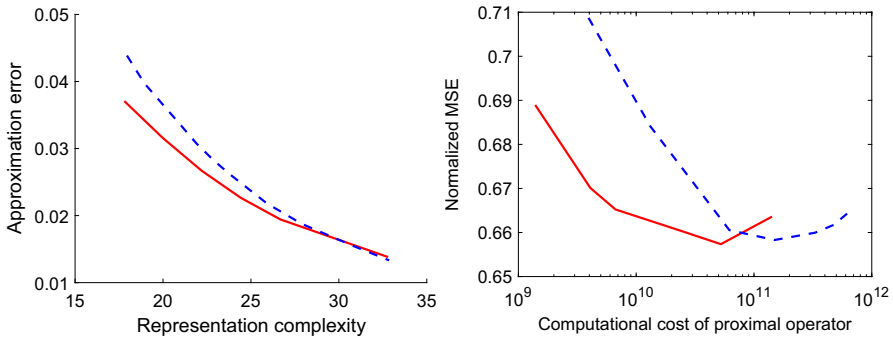
Fig. 6 Comparison between dictionary learning (blue; dashed) and our approach (red; solid) in representing natural image patches (left); comparison between polyhedral (blue; dashed) and semidefinite (red; solid) regularizers in denoising natural image patches (right)

$$\hat{\mathbf{y}}_{\text{denoise}} = \underset{\substack{X \in \mathbb{R}^{q \times q} \\ Z_1, Z_2 \in \mathbb{S}^q \\ s, t \in \mathbb{R}}}{\arg \min} \quad \tfrac{1}{2}s + \lambda t$$

$$\text{s.t.} \quad \|\mathbf{y}_{\text{obs}} - \mathcal{L}(X)\|_{\ell_2}^2 \le s, \ \tfrac{1}{2}\text{trace}(Z_1 + Z_2) \le t, \ \begin{pmatrix} Z_1 & X \\ X' & Z_2 \end{pmatrix} \succeq 0. \tag{29}$$

As before, to solve (29) to an accuracy ϵ using an interior-point method with the usual logarithmic barriers for the positive-semidefinite cone and the second-order cone, we have that the number of operations required is

$\sqrt{2q+2} \log \left(\frac{2q+2}{\epsilon \eta} \left(\left(d + 2\binom{q}{2} + 2\right)^3 + \left(2\binom{q}{2} + 2\right)^3 \right) \right)$ multiplied by

$\left(d + \binom{2q}{2} + 2\right)^3 + \left(\binom{2q}{2} + 2\right)^3$ for a barrier parameter η [51].

Results. We learn semidefinite regularizers on the training set using Algorithm 3 for $q \in \{9, \ldots, 20\}$ and for a rank of 1. We also learn polyhedral regularizers on the training set using dictionary learning for $p \in \{9^2, 10^2, \ldots, 20^2\}$ and with corresponding sparsity levels in the range $\{\sqrt{p} - 1, \sqrt{p}\}$ to ensure that the representation complexity matches the corresponding representation complexity of the images of rank-one matrices in the semidefinite case. As the lifted dimensions q^2 and p increase, the computational complexities of the associated proximal denoisers (with the learned regularizers) also increase. The right subplot in Fig. 6 gives the average normalized mean-squared error over the noisy test data (generated as described above). The optimal choice of the regularization parameter λ for each regularizer is obtained by sweeping over a range to obtain the best denoising performance, as we have access to the underlying uncorrupted image patches $\{\mathbf{y}_{\text{test}}^{(j)}\}_{j=1}^{720}$. For both types of regularizers, the denoising performance improves initially before degrading due to overfitting. More significantly, given a fixed computational budget, these experiments suggest that semidefinite regularizers provide better performance than polyhedral regularizers in denoising image patches in our data set. The denoising operation (27) is in fact a basic computational building block (often referred to as a proximal operator) in first-order

algorithms for solving convex programs that arise in a range of inverse problems [48]. As such, we expect the results of this section to be qualitatively indicative of the utility of our approach in other inferential tasks beyond denoising.

5 Discussion

Our paper describes an algorithmic framework for learning regularizers from data in settings in which prior domain-specific expertise is not directly available. We learn these regularizers by computing a structured factorization of the data matrix, which is accomplished by combining techniques for the affine rank minimization problem with the Operator Sinkhorn scaling procedure. The regularizers obtained using our method are convex, and they can be computed via semidefinite programming. Our approach may be viewed as a semidefinite analog of dictionary learning, which can be interpreted as a technique for learning polyhedral regularizers from data. We discuss next some directions for future work.

5.1 Algorithmic Questions

It would be of interest to better understand the question of initialization for our algorithm. Random initialization often works well in practice, and it would be useful to provide theoretical support for this approach by building on recent work on other factorization problems [29,58]. To this end, we describe two experimental setups on synthetic data showing instances where our algorithm recovers the true regularizer from random initialization. In the first setup, we generate a standard Gaussian linear map $\mathcal{L} : \mathbb{R}^{8\times 8} \to \mathbb{R}^{50}$ and normalize it. Let $\mathcal{L}^\star$ denote the resulting normalized map. We generate data $\{\mathbf{y}^{(j)}\}_{j=1}^{10^4}$ as $\mathbf{y}^{(j)} = \mathcal{L}^\star(\mathbf{u}^{(j)}\mathbf{v}^{(j)'})/\|\mathcal{L}^\star(\mathbf{u}^{(j)}\mathbf{v}^{(j)'})\|_{\ell_2}$, where each $\mathbf{u}^{(j)}, \mathbf{v}^{(j)}$ is drawn independently from the Haar measure on the unit sphere in $\mathbb{R}^8$. We apply our algorithm to the data, and we supply as initialization the normalization of a standard Gaussian linear map. The left subplot of Fig. 7 shows the progression of the mean-squared error over 10 different initializations. As the measurements do not contain any additional noise, the minimum attainable error is zero. We observe that our algorithm recovers the regularizer in all 10 random initializations; moreover, we observe local, linear convergence in the neighborhood of the global minimizer, which agrees with our analysis. Note that the progress of our algorithm reveals interesting behavior in that the global recovery of the regularizer is characterized by three distinct phases—(i) an initial phase in which progress is significant; (ii) an intermediate phase in which progress is incremental but stable; and (iii) a terminal phase that corresponds to local, linear convergence. In particular, these graphs indicate that global convergence to the underlying regularizer is *not* linear. The second setup is similar to the first one, with the two main differences being that we consider a linear map $\mathcal{L}^\star : \mathbb{R}^{8\times 8} \to \mathbb{R}^{60}$ of slightly different dimensions and that the data points $\{\mathbf{y}^{(j)}\}_{j=1}^{2\times 10^4}$ are images of rank-two matrices. The right subplot of Fig. 7 shows the progression of our algorithm over 10 different initializations. In contrast to the previous setup where every initialization led to a global minimum, in this case our algorithm attains a local minimum in 4 out of
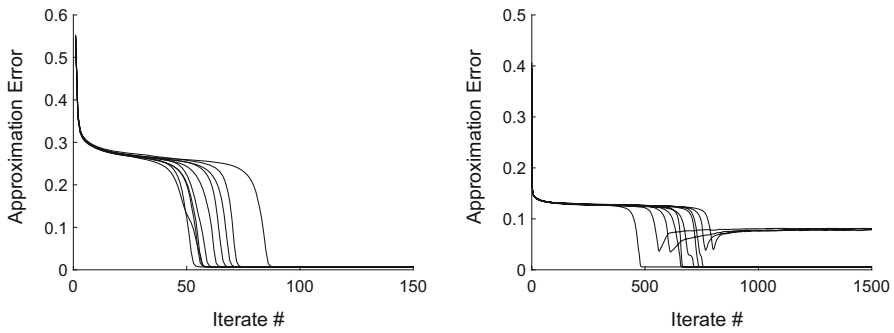
Fig. 7 Progression of our algorithm in recovering regularizers in a synthetic experimental setup; the horizontal axis represents the number of iterations, and each line corresponds to a different random initialization. The left subplot shows a problem instance in which all 10 different random initializations recover a global minimizer, while the right subplot shows a different problem instance in which 4 out of 10 random initializations lead to local minima

10 initializations and a global minimum in the remaining 6 initializations. In summary, our experiments suggest that random initialization may sometimes be effective, and understanding this effectiveness warrants further investigation.

Beyond random initialization, there have also been efforts on data-driven strategies for initialization in dictionary learning by reducing the question to a type of clustering/community detection problem [2,5]. While the relation between clustering and estimating the elements of a finite atomic set is conceptually natural, identifying an analog of the clustering problem for estimating the image of a variety of rank-one matrices (which is a structured but infinite atomic set) is less clear; we seek such a conceptual link in order to develop an initialization strategy for our algorithm. In a completely different direction, there is also recent work on a convex relaxation for the dictionary learning problem that avoids the difficulties associated with local minima [6]; while this technique is considerably more expensive computationally in comparison with alternating updates, developing analogous convex relaxation approaches for the problem of learning semidefinite regularizers may subsequently point the way to efficient global techniques that are different from alternating updates.

5.2 Approximation-Theoretic Questions

The focus of our paper has been on the algorithmic aspects of learning semidefinite regularizers from data. It is of interest to investigate the power of finite atomic sets in comparison with atomic sets specified as projections of determinantal varieties from a harmonic analysis perspective (for a fixed representation complexity; see Sect. 4.2.1 for a discussion on how these are defined). For example, what types of data are better described using one representation framework versus the other? As a simple preliminary illustration, we generate two sets of 400 points in $\mathbb{R}^{500}$, with the first set being a random projection of sparse vectors in $\mathbb{R}^{900}$ and the second set being a random projection of rank-one matrices in $\mathbb{R}^{900}$ of the form $(\cdots \cos(2\pi\alpha_j t_i), \sin(2\pi\alpha_j t_i), \cdots)'(\cdots \cos(2\pi\beta_j t_i), \sin(2\pi\beta_j t_i), \cdots)$ for ran-
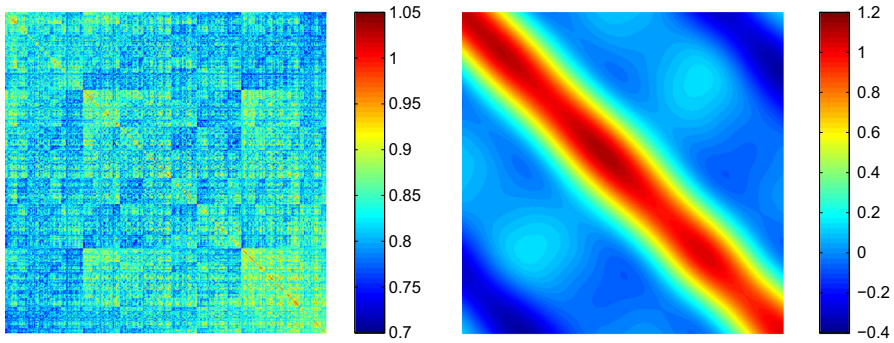
Fig. 8 Gram matrices of images of sparse vectors (left) and low-rank matrices (right)

domly chosen frequencies α_j, β_j; the representation complexities of both these sets is the same. Figure 8 gives the Gram matrices associated with these data sets. The data set of projections of sparse vectors appears to consist of 'clusters' of 'block' structure, while the data set of projections of low-rank matrices appears to consist of smoother 'toroidal' structure. We seek a better understanding of this phenomenon by analyzing the relative strengths of representations based on finite atomic sets versus projections of low-rank matrices. In a different direction, it is also of interest to explore other families of infinite atomic sets that yield tractable regularizers in other conic programming frameworks. Specifically, dictionary learning and our approach provide linear and semidefinite programming regularizers, but there are other families of computationally efficient convex cones such as the power cone and the exponential cone; learning atomic sets that are amenable to optimization in these frameworks would lead to a broader suite of data-driven approaches for identifying regularizers.

Appendix A: Proofs of Lemmas 1 and 2

Proof (Lemma 1) Note that if $Z \in \mathcal{T}$, then Z has rank at most $2r$. As a consequence of the restricted isometry property, we have $(1 - \delta_{2r})\|Z\|_{\ell_2}^2 \leq \|[\mathcal{L} \circ \mathcal{P}_{\mathcal{T}}](Z)\|_{\ell_2}^2 \leq (1+\delta_{2r})\|Z\|_{\ell_2}^2$. Since $Z \in \mathcal{T}$ is arbitrary, we have $1-\delta_{2r} \leq \lambda(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}}) \leq 1+\delta_{2r}$, which proves (i). This immediately implies the bound in (ii). Moreover since $\|\mathcal{L} \circ \mathcal{P}_{\mathcal{T}}\|_2 = \|\mathcal{P}_{\mathcal{T}} \circ \mathcal{L}' \circ \mathcal{L} \circ \mathcal{P}_{\mathcal{T}}\|_2^{1/2} \leq \sqrt{1 + \delta_{2r}}$, we have $\|\mathcal{P}_{\mathcal{T}} \circ \mathcal{L}'\mathcal{L}\|_2 \leq \sqrt{1 + \delta_{2r}}\|\mathcal{L}\|_2$, which is (iii). Last we have $\|[(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}})^{-1}]_{\mathbb{R}^{q \times q}} \circ \mathcal{L}'\mathcal{L}\|_2 \leq \|[(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}})^{-1}]_{\mathbb{R}^{q \times q}}\|_2 \|\mathcal{P}_{\mathcal{T}} \circ \mathcal{L}'\mathcal{L}\|_2 \leq \frac{\sqrt{1+\delta_{2r}}}{1-\delta_{2r}}\|\mathcal{L}\|_2$, which proves (iv). $\qquad\square$

Proof (Lemma 2) To simplify notation, we omit $(\mathfrak{X})$. Since $\text{trace}(\Sigma) = \frac{1}{n}\sum_{j=1}^{n} \|X^{(j)}\|_{\ell_2}^2$, we have $s_{\min} \leq \text{trace}(\Sigma) \leq s_{\max}$. Next we have the inequalities $(\Lambda - \Delta)\mathsf{I} \preceq \Sigma \preceq (\Lambda + \Delta)\mathsf{I}$. The result follows by applying trace. $\qquad\square$

Appendix B: Proof of Proposition 8

In this section, we prove that the ensemble of random matrices $\mathfrak{X}$ described in Proposition 8 satisfy the deterministic conditions in Theorem 3 with high probability. We begin with computing $\mathbb{E}_{\mathcal{D}}[X^{(j)} \boxtimes X^{(j)}]$, and $\mathbb{E}_{\mathcal{D}}[(X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})}]$. Note that the random matrices $\{X^{(j)} \boxtimes X^{(j)}\}_{j=1}^n$ and the random operators $\{(X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})}\}_{j=1}^n$ are almost surely bounded above in spectral norm by construction. This allows us to conclude Proposition 8 with an application of the Matrix Hoeffding Inequality [63].

To simplify notation, we adopt the following. In the first two results, we omit the superscript j from $X^{(j)}$. In the remainder of the section, we let $\mathbb{E} = \mathbb{E}_{\mathcal{D}}$, $\bar{s}^2 := \mathbb{E}[s^2]$, $\{\mathbf{e}_i\}_{i=1}^q \subset \mathbb{R}^q$ be the set of standard basis vectors, and $\{E_{ij}\}_{i,j=1}^q \subset \mathbb{R}^{q \times q}$ be the set of matrices whose (i,j)th entry is 1 and is 0 everywhere else.

Proposition 9 *Suppose $X \sim \mathcal{D}$ as described in Proposition 8. Then, $\mathbb{E}[X \boxtimes X] = \bar{s}^2 (r/q^2)\mathsf{I}$.*

Proof It suffices to show that $\mathbb{E}\langle X \boxtimes X, \mathbf{e}_w \mathbf{e}_x' \boxtimes \mathbf{e}_y \mathbf{e}_z'\rangle = \mathbb{E}\langle X, \mathbf{e}_w \mathbf{e}_x'\rangle \langle X, \mathbf{e}_y \mathbf{e}_z'\rangle = \delta_{wy}\delta_{xz}\bar{s}^2 (r/q^2)$. Let $X = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i'$ as described in the statement of Proposition 8. Suppose we denote $\mathbf{u}_i = (u_{i1}, \ldots, u_{iq})'$, and $\mathbf{v}_i = (v_{i1}, \ldots, v_{iq})'$. By applying independence, we have $\mathbb{E}\langle X, \mathbf{e}_w \mathbf{e}_x'\rangle \langle X, \mathbf{e}_y \mathbf{e}_z'\rangle = \mathbb{E}[(\sum_{i=1}^r s_i u_{iw} v_{ix})(\sum_{k=1}^r s_k u_{ky} v_{kz})] = \sum_{i,k=1}^r \mathbb{E}[s_i s_k]\mathbb{E}[u_{iw} u_{ky}]\mathbb{E}[v_{ix} v_{kz}]$. There are two cases we need to consider.
[Case $w \neq y$ or $x \neq z$]: Without loss of generality, suppose that $w \neq y$. Then, $\mathbb{E}[u_{iw} u_{ky}] = 0$ for all $1 \leq i, k \leq q$, and hence $\mathbb{E}\langle X \boxtimes X, \mathbf{e}_w \mathbf{e}_x' \boxtimes \mathbf{e}_y \mathbf{e}_z'\rangle = 0$.
[Case $w = y$ and $x = z$]: Note that if $i \neq k$, then $\mathbb{E}[u_{iw} u_{ky}] = \mathbb{E}[u_{iw}]\mathbb{E}[u_{ky}] = 0$. Since $\mathbf{u}_i$ is a unit-norm vector distributed u.a.r., we have $\mathbb{E}[u_{ix}^2] = 1/q$. Hence $\mathbb{E}\langle X \boxtimes X, \mathbf{e}_w \mathbf{e}_x' \boxtimes \mathbf{e}_y \mathbf{e}_z'\rangle = \sum_{i=1}^r \mathbb{E}[s_i^2]\mathbb{E}[u_{iw}^2]\mathbb{E}[v_{ix}^2] = \bar{s}^2 r/q^2$. $\square$

Our next result requires the definition of certain subspaces of $\mathbb{R}^{q \times q}$ and $\mathrm{End}(\mathbb{R}^{q \times q})$.

We define the following subspaces in $\mathbb{R}^{q \times q}$: Let $\mathcal{G} := \{W : W = W', W \in I^{\perp}\}$ be the subspace of symmetric matrices that are orthogonal to the identity, $\mathcal{H} := \{W : W = -W'\}$ be the subspace of skew-symmetric matrices, and $\mathcal{I} = \mathrm{Span}(I)$. It is clear that $\mathbb{R}^{q \times q} = \mathcal{G} \oplus \mathcal{H} \oplus \mathcal{I}$.

In addition to the subspace $\mathcal{W}$ defined in (12), we define the following subspaces in $\mathrm{End}(\mathbb{R}^{q \times q})$:

1. $\mathcal{W}_{SS} := \mathrm{Span}(\{A \otimes B : A, B \in \mathcal{G}\})$,
2. $\mathcal{W}_{AA} := \mathrm{Span}(\{A \otimes B : A, B \in \mathcal{H}\})$,
3. $\mathcal{W}_{SA} := \mathrm{Span}(\{A \otimes B : A \in \mathcal{G}, B \in \mathcal{H}\})$,
4. $\mathcal{W}_{AS} := \mathrm{Span}(\{A \otimes B : A \in \mathcal{H}, B \in \mathcal{G}\})$.

Note that $\mathrm{End}(\mathbb{R}^{q \times q}) = \mathcal{W} \oplus \mathcal{W}_{SS} \oplus \mathcal{W}_{AA} \oplus \mathcal{W}_{SA} \oplus \mathcal{W}_{AS}$. To verify this, first express an arbitrary linear map $\mathsf{E} \in \mathrm{End}(\mathbb{R}^{q \times q})$ as a sum of Kronecker products $\mathsf{E} = \sum_{i=1} A_i \otimes B_i$, second decompose each matrix A_i, B_i into components in the subspaces $\{\mathcal{G}, \mathcal{H}, \mathcal{I}\}$, and third expand the expression. The orthogonality between subspaces is immediate from the identity $\langle A_i \otimes B_i, A_j \otimes B_j\rangle = \langle A_i, A_j\rangle \langle B_i, B_j\rangle$.

Proposition 10 *Suppose $X \sim \mathcal{D}$ as described in Proposition 8. Then,*

$$\mathbb{E}[(X \boxtimes X) \otimes \mathcal{P}_{\mathcal{T}(X)}] = c_{\mathcal{W}}\mathsf{I}_{\mathcal{W}} + c_{\mathcal{W}_{SS}}\mathsf{I}_{\mathcal{W}_{SS}} + c_{\mathcal{W}_{AA}}\mathsf{I}_{\mathcal{W}_{AA}} + c_{\mathcal{W}_{SA}}\mathsf{I}_{\mathcal{W}_{SA}} + c_{\mathcal{W}_{AS}}\mathsf{I}_{\mathcal{W}_{AS}},$$

where (i) $c_{\mathcal{W}} = \bar{s}^2 r(\frac{1}{q^2})$, *(ii)* $c_{\mathcal{W}_{SS}} = \bar{s}^2 r(\frac{1}{q^2} - \frac{(q-r)^2}{(q-1)^2(q+2)^2})$, *(iii)* $c_{\mathcal{W}_{AA}} = \bar{s}^2 r(\frac{1}{q^2} - \frac{(q-r)^2}{q^2(q-1)^2})$, *and (iv)* $c_{\mathcal{W}_{SA}} = c_{\mathcal{W}_{AS}} = \bar{s}^2 r(\frac{1}{q^2} - \frac{(q-r)^2}{q(q-1)^2(q+2)})$.

Proof The proof consists of two parts, namely (i) to prove that the mean, when restricted to the respective subspaces described above, has diagonal entries as specified, and (ii) to prove that the off-diagonal elements are zero with respect to any basis that obeys the specified decomposition of $\mathrm{End}(\mathbb{R}^{q \times q})$. In addition, it suffices to only consider linear maps that are Kronecker products since these maps generate the respective subspaces. The following identity for all matrices A_i, B_i, A_j, B_j is particularly useful

$$\langle (A_i' \otimes B_i) \boxtimes (A_j' \otimes B_j), \mathbb{E}[(X \boxtimes X) \otimes \mathcal{P}_{\mathcal{T}(X)}] \rangle = \mathbb{E}\langle \mathcal{P}_{\mathcal{T}(X)}(B_j X A_j), \mathcal{P}_{\mathcal{T}(X)}(B_i X A_i) \rangle. \tag{30}$$

One may equivalently describe the distribution of X as follows—let $X = U \Sigma_R V'$, where U, V are $q \times q$ matrices drawn from the Haar measure and Σ_R is a diagonal matrix whose first r entries are drawn from $\mathcal{D}$, and the remaining entries are 0 (to simplify notation, we omit the dependence on X in the matrices U, V). Let $I_N = \mathrm{diag}(0, \ldots, 0, 1, \ldots, 1)$ be a diagonal matrix consisting of $q - r$ ones. Under this notation, the projector is simply the map $\mathcal{P}_{\mathcal{T}(X)}(Z) = Z - U I_N U' Z V I_N V'$. The remainder of the proof is divided into the two parts outlined above.

[Part (i)]: The restriction to diagonal entries correspond to the case $i = j$, and hence, equation (30) simplifies to $\mathbb{E}[\|\mathcal{P}_{\mathcal{T}(X)}(BXA)\|_{\ell_2}^2]$. Consequently, we have

$$\mathbb{E}[\|\mathcal{P}_{\mathcal{T}(X)}(BXA)\|_{\ell_2}^2] = \mathbb{E}[\|BU\Sigma_R V'A\|_{\ell_2}^2] - \mathbb{E}[\|I_N U'AU\Sigma_R V'BV I_N\|_{\ell_2}^2].$$

First, we compute $\mathbb{E}[\|I_N U'AU\Sigma_R V'BV I_N\|_{\ell_2}^2]$. By the cyclicity of trace and iterated expectations, we have

$$\mathbb{E}[\|I_N U'AU\Sigma_R V'BV I_N\|_{\ell_2}^2]$$
$$= \mathbb{E}[\mathrm{trace}(\Sigma_R^{1/2} U'A'U I_N U'AU\Sigma_R V'BV I_N V'B'V\Sigma_R^{1/2})]$$
$$= \mathbb{E}_U[\mathbb{E}_V[\mathrm{trace}(\Sigma_R^{1/2} U'A'U I_N U'AU\Sigma_R V'BV I_N V'B'V\Sigma_R^{1/2})]].$$

It suffices to compute $\mathbb{E}[\Sigma_R^{1/2} V'BV I_N V'B'V\Sigma_R^{1/2}] = \Sigma_R^{1/2} \mathbb{E}[V'BV I_N V'B'V]\Sigma_R^{1/2}$ in the three cases corresponding to $B \in \{\mathcal{G}, \mathcal{H}, \mathcal{I}\}$, respectively. Using linearity and symmetry, it suffices to compute $\mathbb{E}[V'BV E_{11} V'B'V]$. We split this computation into the following three separate cases.

[Case $B \in \mathcal{I}$]: We have $I_N \Sigma_R^{1/2} = 0$, and hence the mean is the zero matrix.

[Case $B \in \mathcal{H}$]: Claim: If $B \in \mathcal{H}$, and $\|B\|_{\ell_2} = 1$, then $\mathbb{E}[V'BV E_{11} V'B'V] = (I - E_{11})/(q(q-1))$.

Proof Denote $V = [\mathbf{v}_1 | \ldots | \mathbf{v}_q]$. The off-diagonal entries vanish as $\mathbb{E}\langle E_{ij}, V'BV E_{11} V'B'V \rangle = \mathbb{E}(\mathbf{v}_i'B\mathbf{v}_i)(\mathbf{v}_1'B\mathbf{v}_j) = 0$ whenever $i \neq j$, as one of the indices i, j appears exactly once. By a symmetry argument, we have $\mathbb{E}[V'BV E_{11} V'B'V] = \alpha I + \beta E_{11}$ for some α, β. First $\mathbb{E}[\mathrm{trace}(V'BV E_{11} V'B'V)] = \mathbb{E}[\mathrm{trace}(BV E_{11} V'B')] =$

trace$(B\mathbb{E}[VE_{11}V']B') = $ trace$(B(I/q)B') = 1/q$, which gives $\alpha q + \beta = 1/q$. Second since B is asymmetric, $V'BV$ is also asymmetric and hence is 0 on the diagonals. Thus, $\langle V'BVE_{11}V'B'V, E_{11}\rangle = 0$, which gives $\alpha + \beta = 0$. The two equations yield the values of α and β.

[Case: $B \in \mathcal{G}$]: Claim: If $B \in \mathcal{G}$, and $\|B\|_{\ell_2} = 1$, then $\mathbb{E}[V'BVE_{11}V'B'V] = (I + (1 - 2/q)E_{11})/((q-1)(q+2))$.

Proof With an identical argument as the previous claim, one has $\mathbb{E}[V'BVE_{11}V'B'V] = \alpha I + \beta E_{11}$, where $\alpha q + \beta = 1/q$. Next $\mathbb{E}[\langle V'BVE_{11}V'B'V, E_{11}\rangle] = \mathbb{E}[(\mathbf{v}_1'B\mathbf{v}_1)^2]$, where $\mathbf{v}_1$ is a unit-norm vector distributed u.a.r. Since conjugation by orthogonal matrices preserves trace, and $\mathbf{v}_1$ has the same distribution as $Q\mathbf{v}_1$ for any orthogonal Q, we may assume that $B = \text{diag}(b_{11}, \ldots, b_{qq})$ is diagonal without loss of generality. Suppose we let $\mathbf{v}_1 = (v_1, \ldots, v_q)'$. Then, $\mathbb{E}[(\mathbf{v}_1'B\mathbf{v}_1)^2] = \mathbb{E}[\sum b_{ii}^2 v_i^4 + \sum_{i \neq j} b_{ii}b_{jj}v_i^2 v_j^2] = \mu_1(\sum b_{ii}^2) + \mu_2(\sum_{i \neq j} b_{ii}b_{jj})$, where $\mu_1 = \mathbb{E}[v_1^4]$, and $\mu_2 = \mathbb{E}[v_1^2 v_2^2]$. Since trace$(B) = 0$, we have $\sum b_{ii}^2 = -\sum_{i \neq j} b_{ii}b_{jj}$. Last from Theorem 2 of [18], we have $\mu_1 = 3/(q(q+2))$, and $\mu_2 = 1/(q(q+2))$, which gives $\mathbb{E}[(\mathbf{v}_1'B\mathbf{v}_1)^2] = 2/(q(q+2))$, and hence $\alpha + \beta = 2/(q(q+2))$. The two equations yield the values of α and β.

With a similar set of computations, one can show that $\mathbb{E}[\|BU\Sigma_R V'A\|_{\ell_2}^2] = \bar{s}^2 r/q^2$ for arbitrary unit-norm A, B. An additional set of computations yields the diagonal entries, which completes the proof. We omit these computations.

[Part (ii)]: We claim that it suffices to show that $\mathbb{E}[V'A_i V E_{11}V'A_j'V]$ is the zero matrix whenever $A_i, A_j \in \{\mathcal{G}, \mathcal{H}, \mathcal{I}\}$, and satisfy $\langle A_i, A_j\rangle = 0$. We show how this proves the result. Suppose $A_i \otimes B_i, A_j \otimes B_j$ satisfy $\langle A_i \otimes B_i, A_j \otimes B_j\rangle = \langle A_i, A_j\rangle\langle B_i, B_j\rangle = 0$. Without loss of generality, we may assume that $\langle A_i, A_j\rangle = 0$. From equation (30), we have

$$\mathbb{E}\langle \mathcal{P}_{\mathcal{T}(X)}(B_j X A_j), \mathcal{P}_{\mathcal{T}(X)}(B_i X A_i)\rangle = \mathbb{E}[\text{trace}(A_j'V\Sigma_R U'B_j'B_i U\Sigma_R V'A_i)]$$
$$- \mathbb{E}[\text{trace}(A_j'V\Sigma_R U'B_j'U I_N U'B_i U\Sigma_R V'A_i V I_N V')].$$

By cyclicity of trace and iterated expectations, we have

$$\mathbb{E}[\text{trace}(A_j'V\Sigma_R U'B_j'U I_N U'B_i U\Sigma_R V'A_i V I_N V')]$$
$$= \mathbb{E}_U[\text{trace}(\Sigma_R^{1/2}U'B_j'U I_N U'B_i U\Sigma_R^{1/2}$$
$$(\mathbb{E}_V[\Sigma_R^{1/2}V'A_i V I_N V'A_j'V\Sigma_R^{1/2}]))] = 0,$$

which proves part (ii) of the proof. It leaves to prove the claim. We do so by verifying that the matrix $\mathbb{E}[V'A_i V E_{11}V'A_j'V]$ is 0 in every coordinate, which is equivalent to showing that $\mathbb{E}(\mathbf{v}_m'A_i\mathbf{v}_1)(\mathbf{v}_n'A_j\mathbf{v}_1) = 0$ for all m, n. There are three cases.

[Case $m \neq n$]: Without loss of generality, suppose that $m \neq 1$. Then, $\mathbb{E}(\mathbf{v}_m'A_i\mathbf{v}_1)(\mathbf{v}_n'A_j\mathbf{v}_1) = \mathbb{E}[\mathbb{E}[(\mathbf{v}_m'A_i\mathbf{v}_1)(\mathbf{v}_n'A_j\mathbf{v}_1)|\mathbf{v}_1, \mathbf{v}_n]] = 0$.

[Case $m = n = 1$]: We divide into further subcases depending on the subspaces A_i, A_j belong to. If $A_i \in \mathcal{H}$, then $\mathbf{v}_1'A_i\mathbf{v}_1 = 0$ since it is a scalar. Hence, we eliminate the case where either matrix is in $\mathcal{H}$. Since $\langle A_i, A_j\rangle = 0$ it cannot be that both $A_i, A_j \in \mathcal{I}$. Suppose that $A_i = I/\sqrt{q}$ and $A_j \in \mathcal{G}$.

Then, $\mathbb{E}[(\mathbf{v}_1' A_i \mathbf{v}_1)(\mathbf{v}_1' A_j \mathbf{v}_1)] = \mathbb{E}[(\mathbf{v}_1' A_j \mathbf{v}_1)]/\sqrt{q} = \mathbb{E}[\text{trace}(A_j \mathbf{v}_1 \mathbf{v}_1')]/\sqrt{q} = 0$. Our remaining case is when $A_i, A_j \in \mathcal{G}$, and $\langle A_i, A_j \rangle = 0$. As before, we let $\mathbf{v}_1 = (v_1, \ldots, v_q)'$. Then,

$$\mathbb{E}[(\mathbf{v}_1' A_i \mathbf{v}_1)(\mathbf{v}_1' A_j \mathbf{v}_1)] = \mathbb{E}[\sum_{pqrs} A_{i,pq} A_{j,rs} v_p v_q v_r v_s]$$

$$= \sum_p A_{i,pp} A_{j,pp} \mathbb{E}[v_p^4] + \sum_{p \neq r} A_{i,pp} A_{j,rr} \mathbb{E}[v_p^2 v_r^2] + 2 \sum_{p \neq q} A_{i,pq} A_{j,pq} \mathbb{E}[v_p^2 v_q^2],$$

where in the second equality we used the fact that A_i, A_j are symmetric to obtain a factor of 2 in the last term. Next we apply the relations $\mathbb{E}[v_p^4] = 3/(q(q+2))$, $\mathbb{E}[v_p^2 v_r^2] = 1/(q(q+2))$, as well as the relations $0 = \langle A_i, I \rangle \langle A_j, I \rangle = \sum_p A_{i,pp} A_{j,pp} + \sum_{p \neq r} A_{i,pp} A_{j,rr}$, and $0 = \langle A_i, A_j \rangle = \sum_p A_{i,pp} A_{j,pp} + \sum_{p \neq q} A_{i,pq} A_{j,pq}$ to conclude that the mean is zero.

[Case $m = n \neq 1$]: We have

$$\mathbb{E}[(\mathbf{v}_m' A_i \mathbf{v}_1)(\mathbf{v}_m' A_j \mathbf{v}_1)] = \mathbb{E}[\mathbb{E}[\text{trace}(A_i \mathbf{v}_1 \mathbf{v}_1' A_j' \mathbf{v}_m \mathbf{v}_m')]|\mathbf{v}_1]$$

$$= \mathbb{E}[\text{trace}(A_i \mathbf{v}_1 \mathbf{v}_1' A_j' (I - \mathbf{v}_1 \mathbf{v}_1')/(q-1))|\mathbf{v}_1]$$

$$= \mathbb{E}[\text{trace}(A_i \mathbf{v}_1 \mathbf{v}_1' A_j'/(q-1))]$$

$$= \mathbb{E}[\text{trace}(A_i I A_j'/(q(q-1)))] = 0,$$

where the first equality applies the fact that conditioned on $\mathbf{v}_1$, $\mathbb{E}[\mathbf{v}_m \mathbf{v}_m']$ is the identity matrix in the subspace $\mathcal{T}(\mathbf{v}_1 \mathbf{v}_1')^\perp$ suitably scaled, and the second inequality applies the previous case. □

Proof (Proposition 8) First, we have $0 \preceq X^{(j)} \boxtimes X^{(j)} \preceq s^2 r I$. By Proposition 9, we have $\mathbb{E}[X^{(j)} \boxtimes X^{(j)}] = (\bar{s}^2 r/q^2) I$. Since $(X^{(j)} \boxtimes X^{(j)} - (\bar{s}^2 r/q^2) I)^2 \preceq s^4 r^2 I$, we have $\mathbb{P}(\|(1/n) \sum_{i=1}^n X^{(j)} \boxtimes X^{(j)} - (\bar{s}^2 r/q^2) I\| > tr s^2) \leq 2q \exp(-t^2 n/8)$ via an application of the Matrix Hoeffding Inequality (Theorem 1.3 in [63]).

Second we have $\|X^{(j)} \boxtimes X^{(j)}\|_2 \leq s^2 r$, and $\|\mathcal{P}_{\mathcal{T}(X^{(j)})}\|_2 = 1$, and hence $(X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})} \preceq s^2 r I \otimes I =: s^2 r \mathbf{I}$. From Proposition 10, we have

$$\mathbb{E}[(X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})}] \preceq \frac{\bar{s}^2 r}{q^2} \mathbf{I}_\mathcal{W} + \frac{16 \bar{s}^2 r^2}{q^3} \mathbf{I}_{\mathcal{W}^\perp}.$$

Since $((X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})} - r\mathbf{I})^2 \preceq s^4 r^2 \mathbf{I}$, we have

$$\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n} \sum_{i=1}^n (X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})} - \mathbb{E}[(X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})}]\right) \geq tr s^2\right)$$

$$\leq q \exp(-t^2 n/8)$$

by an application of the Matrix Hoeffding Inequality.

Let $t = t_1/(5q^2)$ in the first concentration bound, and $t = t_2/(5q^2)$ in the second concentration bound. Then, $\Delta(\mathfrak{X}) \leq t_1 s^2 r/(5q^2)$, and $\Omega(\mathfrak{X}) \leq$

$16s^2r^2/q^3 + t_2s^2r/(5q^2)$, with probability greater than $1 - 2q\exp(-nt_1^2/(200q^4)) - q\exp(-nt_2^2/(200q^4))$. We condition on the event that both inequalities hold. Since $\Delta(\mathfrak{X}) \leq t_1 s^2 r/(5q^2) \leq s^2 r/(20q^2)$, by Lemma 2, we have $\Lambda(\mathfrak{X}) \geq s^2r/(5q^2)$, and hence $\Delta(\mathfrak{X})/\Lambda(\mathfrak{X}) \leq t_1$, and $\Omega(\mathfrak{X})/\Lambda(\mathfrak{X}) \leq 80r/q + t_2$. □

Appendix C: Stability of Matrix and Operator Scaling

In this section, we prove a stability property of Sinkhorn scaling and Operator Sinkhorn scaling. For Sinkhorn scaling, we show that if a matrix is close to being doubly stochastic and has entries that are suitably bounded away from 0, then the resulting row and column scalings are close to $\mathbf{1} := (1, \ldots, 1)'$. We also prove the operator analog of this result. These results are subsequently used to prove Propositions 2 and 7. We note that there is an extensive literature on the stability of matrix scaling, with results of a similar flavor to ours. However, Proposition 11 in this section is stated in a manner that is directly suited to our analysis, and we include it for completeness.

5.3 Main Results

Proposition 11 (Local stability of Matrix Scaling) *Let $T \in \mathbb{R}^{q \times q}$ be a matrix such that*

1. $|\langle \mathbf{e}_i, T(\mathbf{e}_j) \rangle - 1/q| \leq 1/(2q)$ *for all standard basis vectors* $\mathbf{e}_i, \mathbf{e}_j$; *and*
2. $\epsilon := \max\{\|T\mathbf{1} - \mathbf{1}\|_\infty, \|T'\mathbf{1} - \mathbf{1}\|_\infty\} \leq 1/(48\sqrt{q})$.

Let D_1, D_2 be diagonal matrices such that $D_2 T D_1$ is doubly stochastic. Then,

$$\|D_2 \otimes D_1 - \mathsf{I}\|_2 \leq 96\sqrt{q}\epsilon.$$

Proposition 12 (Local stability of Operator Scaling) *Let $\mathsf{T} : \mathbb{S}^q \to \mathbb{S}^q$ be a rank-indecomposable linear operator such that*

1. $|\langle \mathbf{v}\mathbf{v}', \mathsf{T}(\mathbf{u}\mathbf{u}') \rangle - 1/q| \leq 1/(2q)$ *for all unit-norm vectors* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^q$; *and*
2. $\epsilon := \max\{\|\mathsf{T}(I) - I\|_2, \|\mathsf{T}'(I) - I\|_2\} \leq 1/(48\sqrt{q})$.

Let $N_1, N_2 \in \mathbb{S}^q$ be positive-definite matrices such that $(N_2 \otimes N_2) \circ \mathsf{T} \circ (N_1 \otimes N_1)$ is doubly stochastic. Then, $\|N_2^2 \otimes N_1^2 - \mathsf{I}\|_2 \leq 96\sqrt{q}\epsilon$. Furthermore, we have $\|N_2 \otimes N_1 - \mathsf{I}\|_2 \leq 96\sqrt{q}\epsilon$.

5.4 Proofs

The proof of Proposition 11 relies on the fact that matrix scaling can be cast as the solution of a convex program; specifically, we utilize the correspondence between diagonal matrices D_1, D_2 such that $D_2 T D_1$ is doubly stochastic, and the vectors $\boldsymbol{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_q)', \boldsymbol{\eta} := (\eta_1, \ldots, \eta_q)'$ that minimize the following convex function

$$F(\boldsymbol{\varepsilon}, \boldsymbol{\eta}) = \sum_{ij} T_{ij} \exp(\varepsilon_i + \eta_j) - \sum_i \varepsilon_i - \sum_j \eta_j$$

via the maps $(D_2)_{ii} = \exp(\varepsilon_i)$ and $(D_1)_{jj} = \exp(\eta_j)$ [31] (see also [39])—this holds for all matrices T with positive entries. We remark that one can derive the above relationship from first-order optimality. In the following, we prove bounds on the minima of F (see Lemma 5).

The proof of Proposition 12 relies on a reduction to the setup in Proposition 11.

We begin with a lower estimate of the sum of exponential functions. We use the estimate to prove Proposition 11.

Definition 3 Let $\alpha \geq 0$. Define the function $c_\alpha : \mathbb{R} \to \mathbb{R}$

$$c_\alpha(x) = \begin{cases} \frac{1}{2}\exp(-\alpha)x^2 & \text{if } |x| \leq \alpha \\ \frac{1}{2}\exp(-\alpha)\alpha|x| & \text{if } |x| \geq \alpha \end{cases} .$$

Remark Note that the function $c_\alpha(\cdot)$ is continuous.

Lemma 3 *For all x*

$$\exp(x) \geq 1 + x + c_\alpha(x).$$

Proof (Lemma 3) The second derivative of $\exp(x)$ is $\exp(x)$, and it is greater than $\exp(-\alpha)$ over all x such that $|x| \leq \alpha$. Hence, by strong convexity of $\exp(x)$, we have $\exp(x) \geq 1 + x + (1/2)\exp(-\alpha)x^2$ over the interval $[-\alpha, \alpha]$.

It follows that $\exp(\alpha) \geq 1 + \alpha + c_\alpha(\alpha)$, and $\exp(-\alpha) \geq 1 - \alpha + c_\alpha(-\alpha)$. Since the function $\exp(x)$ is convex, and c_α is linear in the intervals $(-\infty, -\alpha]$ and $[\alpha, \infty)$, respectively, it suffices to check that (i) the gradient of $\exp(x)$ at $x = \alpha$, which is $\exp(\alpha)$, exceeds that of $c_\alpha(\cdot)$, and (ii) the gradient of $c_\alpha(\cdot)$ exceeds that of $\exp(x)$ at $x = -\alpha$, which is $\exp(-\alpha)$.

First, we prove (i). Since $\alpha \geq 0$, we have $1 + 2\alpha \geq \sqrt{1 + 2\alpha}$. Hence, $2\exp(\alpha) \geq 2 + 2\alpha \geq 1 + \sqrt{1 + 2\alpha}$. By noting that the quadratic $2z^2 - 2z - \alpha = 0$ has roots $(1/2) \pm (1/2)\sqrt{1 + 2\alpha}$, we have the inequality $\exp(\alpha) \geq 1 + (1/2)\exp(-\alpha)\alpha$, from which (i) follows.

Next we prove (ii). Since $\alpha \geq 0$, we have $\exp(\alpha) \geq 1 + \alpha \geq 1 + \alpha/2$, and hence $1 - (1/2)\exp(-\alpha)\alpha \geq \exp(-\alpha)$ from which (ii) follows. $\qquad\square$

Lemma 4 *Let $\{\varepsilon_i\}_{i=1}^q$ and $\{\eta_j\}_{j=1}^q$ be a collection of reals satisfying $(\sum_i \varepsilon_i) + (\sum_j \eta_j) \geq -2q$. Then, there is a constant $d \in \mathbb{R}$ for which*

$$\frac{1}{q}\sum_{ij}\exp(\varepsilon_i + \eta_j) \geq q + \left(\sum_i(\varepsilon_i + c_\alpha(\varepsilon_i + d))\right) + \left(\sum_j(\eta_j + c_\alpha(\eta_j - d))\right).$$

Proof Consider the function

$$f(d) := \sum_i \left(\varepsilon_i + d + c_\alpha(\varepsilon_i + d)\right) - \sum_j \left(\eta_j - d + c_\alpha(\eta_j - d)\right).$$

Then, $f(\cdot)$ is continuous in d, and $f(d) \to \pm\infty$ as $d \to \pm\infty$. By the Intermediate Value Theorem, there is a $d^\star$ for which $f(d^\star) = 0$. Then,

$$\sum_i \left(1 + \varepsilon_i + d^\star + c_\alpha(\varepsilon_i + d^\star)\right) = \sum_j \left(1 + \eta_j - d^\star + c_\alpha(\eta_j - d^\star)\right).$$

By summing both sides and noting that $c_\alpha(\cdot) \geq 0$, we have that each side of the above equation is nonnegative. It follows that

$$\frac{1}{q} \sum_{ij} \exp(\varepsilon_i + \eta_j) = \frac{1}{q} \left(\sum_i \exp(\varepsilon_i + d^\star)\right) \left(\sum_j \exp(\eta_j - d^\star)\right)$$

$$\geq \frac{1}{q} \left(\sum_i \left(1 + \varepsilon_i + d^\star + c_\alpha(\varepsilon_i + d^\star)\right)\right)$$

$$\left(\sum_j \left(1 + \eta_j - d^\star + c_\alpha(\eta_j - d^\star)\right)\right)$$

$$\geq q + \left(\sum_i (\varepsilon_i + c_\alpha(\varepsilon_i + d))\right) + \left(\sum_j (\eta_j + c_\alpha(\eta_j - d))\right).$$

$\square$

Lemma 5 *Given vectors* $\boldsymbol{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_q)$ *and* $\boldsymbol{\eta} := (\eta_1, \ldots, \eta_q)$ *define*

$$F(\boldsymbol{\varepsilon}, \boldsymbol{\eta}) = \sum_{ij} T_{ij} \exp(\varepsilon_i + \eta_j) - \sum_i \varepsilon_i - \sum_j \eta_j, \tag{31}$$

and $\epsilon_{ij} := T_{ij} - 1/q$. *Suppose* $(i)\, |\epsilon_{ij}| \leq 1/2q$, *and* $(ii)\, \epsilon := \max\{|\sum_i \epsilon_{ij}|, |\sum_j \epsilon_{ij}|\} \leq 1/(24\sqrt{q})$. *Let* $\boldsymbol{\varepsilon}^\star, \boldsymbol{\eta}^\star$ *be a minimizer of* F. *Then,* $|\varepsilon_i^\star + \eta_j^\star| \leq 48\sqrt{q}\epsilon$, *for all* i, j.

Proof Suppose $|\varepsilon_i + \eta_j| > 48\sqrt{q}\epsilon$ for some (i, j). We show that $\boldsymbol{\varepsilon}, \boldsymbol{\eta}$ cannot be a minimum. We split the analysis to two cases.

$[(\sum_i \varepsilon_i) + (\sum_j \eta_j) < -2q]$: Since $T_{ij} > 0$, we have $F(\boldsymbol{\varepsilon}, \boldsymbol{\eta}) > -(\sum_i \varepsilon_i) - (\sum_j \eta_j) \geq 2q$. Then, $F(\mathbf{0}, \mathbf{0}) = \sum_i (\sum_j T_{ij}) = \sum_i (1 + \sum_j \epsilon_{ij}) \leq q(1 + 1/(24\sqrt{q})) \leq 2q < F(\boldsymbol{\varepsilon}, \boldsymbol{\eta})$.

$[(\sum_i \varepsilon_i) + (\sum_j \eta_j) \geq -2q]$: Let $\alpha = 24\sqrt{q}\epsilon$, and define the sets

1. $\mathfrak{S}(\boldsymbol{\varepsilon}) = \{i : |\varepsilon_i| \geq \alpha\}$;
2. $\mathfrak{T}(\boldsymbol{\varepsilon}) = \{i : \alpha > |\varepsilon_i| \geq 4\epsilon \exp(\alpha)\}$; and
3. $\mathfrak{U}(\boldsymbol{\varepsilon}) = \{i : 4\epsilon \exp(\alpha) > |\varepsilon_i|\}$.

Similarly define the sets $\mathfrak{S}(\boldsymbol{\eta}), \mathfrak{T}(\boldsymbol{\eta}), \mathfrak{U}(\boldsymbol{\eta})$.

First since $\alpha \leq 1$, we have $\alpha \geq \alpha \exp(\alpha)/3 \geq 8\sqrt{q}\epsilon \exp(\alpha) \geq 8\epsilon \exp(\alpha)$, and hence

$$\frac{1}{4} \left(\sum_{i \in \mathfrak{S}(\boldsymbol{\varepsilon})} c_\alpha(\varepsilon_i) + \sum_{j \in \mathfrak{S}(\boldsymbol{\eta})} c_\alpha(\eta_j)\right) \geq \epsilon \left(\sum_{i \in \mathfrak{S}(\boldsymbol{\varepsilon})} |\varepsilon_i| + \sum_{j \in \mathfrak{S}(\boldsymbol{\eta})} |\eta_j|\right).$$

Second

$$\frac{1}{2}\left(\sum_{i \in \mathfrak{T}(\boldsymbol{\varepsilon})} c_\alpha(\varepsilon_i) + \sum_{j \in \mathfrak{T}(\boldsymbol{\eta})} c_\alpha(\eta_j)\right) = \sum_{i \in \mathfrak{T}(\boldsymbol{\varepsilon})} \frac{1}{4}\exp(-\alpha)\varepsilon_i^2 + \sum_{j \in \mathfrak{T}(\boldsymbol{\eta})} \frac{1}{4}\exp(-\alpha)\eta_j^2$$

$$\geq \epsilon\left(\sum_{i \in \mathfrak{T}(\boldsymbol{\varepsilon})} |\varepsilon_i| + \sum_{j \in \mathfrak{T}(\boldsymbol{\eta})} |\eta_j|\right).$$

Third since there is an index (i, j) such that $|\varepsilon_i + \eta_j| > 48\sqrt{q}\epsilon$, one of the sets $\mathfrak{S}(\boldsymbol{\varepsilon})$, $\mathfrak{S}(\boldsymbol{\eta})$ is nonempty. By noting that $\alpha\exp(-\alpha) \geq 8\sqrt{q}\epsilon$, we have

$$\frac{1}{4}\left(\sum_{i \in \mathfrak{S}(\boldsymbol{\varepsilon})} c_\alpha(\varepsilon_i) + \sum_{j \in \mathfrak{S}(\boldsymbol{\eta})} c_\alpha(\eta_j)\right) > \epsilon \times 2q \times 4\epsilon\exp(\alpha) \geq \epsilon\left(\sum_{i \in \mathfrak{U}(\boldsymbol{\varepsilon})} |\varepsilon_i| + \sum_{j \in \mathfrak{U}(\boldsymbol{\eta})} |\eta_j|\right).$$

We have $\epsilon(\sum_i |\varepsilon_i| + \sum_j |\eta_j|) \geq \sum_i (\varepsilon_i(\sum_j \epsilon_{ij})) + \sum_j (\eta_j(\sum_i \epsilon_{ij})) = \sum_{ij} \epsilon_{ij}(\varepsilon_i + \eta_j)$. By combining the above inequalities with Lemma 4, we have

$$\frac{1}{2q}\sum\left(\exp(\varepsilon_i + \eta_j) - (\varepsilon_i + \eta_j) - 1\right) \geq \frac{1}{2}\left(\sum_i c_\alpha(\varepsilon_i) + \sum_j c_\alpha(\eta_j)\right)$$

$$> \sum_{ij} \epsilon_{ij}(\varepsilon_i + \eta_j). \tag{32}$$

Also, since $\exp(\varepsilon_i + \eta_j) - (\varepsilon_i + \eta_j) - 1 \geq 0$ for all i, j, and $|\epsilon_{ij}| \leq 1/(2q)$, we have

$$\frac{1}{2q}\sum_{ij}(\exp(\varepsilon_i + \eta_j) - (\varepsilon_i + \eta_j) - 1)$$

$$\geq \max_{ij}|\epsilon_{ij}| \times \sum_{ij}\left|\exp(\varepsilon_i + \eta_j) - (\varepsilon_i + \eta_j) - 1\right|$$

$$\geq \sum_{ij} \epsilon_{ij}(\exp(\varepsilon_i + \eta_j) - (\varepsilon_i + \eta_j) - 1). \tag{33}$$

By combining equations (32) and (33), we have

$$\frac{1}{q}\sum_{ij}(\exp(\varepsilon_i + \eta_j) - (\varepsilon_i + \eta_j) - 1) > -\sum_{ij} \epsilon_{ij}(\exp(\varepsilon_i + \eta_j) - 1),$$

which implies $F(\boldsymbol{\varepsilon}, \boldsymbol{\eta}) > F(\mathbf{0}, \mathbf{0})$. $\qquad \square$

Proof (Proposition 11) By Lemma 5 any minimum $\boldsymbol{\varepsilon}^\star$, $\boldsymbol{\eta}^\star$ satisfies $|\varepsilon_i^\star + \eta_j^\star| \leq 48\sqrt{q}\epsilon$. Hence by the one-to-one correspondence between the minima of F and the diagonal scalings D_1, D_2 [31], we have $\|D_2 \otimes D_1 - \mathsf{I}\|_2 \leq \exp(48\sqrt{q}\epsilon) - 1 \leq 96\sqrt{q}\epsilon$. $\qquad\square$

Proof (Proposition 12) Without loss of generality, we may assume that N_1, N_2 are diagonal matrices, say D_1, D_2, respectively. Define the matrix $T_{ij} = \langle \mathbf{e}_i \mathbf{e}'_i, \mathsf{T}(\mathbf{e}_j \mathbf{e}'_j) \rangle$. It is straightforward to check that T satisfies the conditions of Proposition 11; moreover, the condition that $(N_2 \otimes N_2) \circ \mathsf{T} \circ (N_1 \otimes N_1)$ is a doubly stochastic operator implies that $D_2^2 T D_1^2$ is a doubly stochastic matrix. By Proposition 11, we have $\|D_1^2 \otimes D_2^2 - \mathsf{I}\|_2 \leq 96\sqrt{q}\epsilon$, and hence $\|N_1^2 \otimes N_2^2 - \mathsf{I}\|_2 \leq 96\sqrt{q}\epsilon$. Since N_1, N_2 are self-adjoint, we also have $\|N_1 \otimes N_2 - \mathsf{I}\|_2 \leq 96\sqrt{q}\epsilon$. $\qquad\square$

Appendix D: Proof of Proposition 7

In this section, we prove that Gaussian linear maps that are subsequently normalized satisfy the deterministic conditions in Theorem 3 concerning the linear map $\mathcal{L}^\star$ with high probability. There are two steps to our proof. First, we state sufficient conditions for linear maps such that, when normalized, satisfy the deterministic conditions. Second we show that Gaussian maps satisfy these sufficient conditions with high probability.

We introduce the following parameter that measures how close a linear map $\mathcal{L}$ is to being normalized.

Definition 4 Let $\mathcal{L} \in \mathbb{R}^{q \times q} \to \mathbb{R}^d$ be a linear map. The *nearly normalized parameter* of $\mathcal{L}$ is defined as

$$\epsilon(\mathcal{L}) := \max\{\|\mathsf{T}_\mathcal{L}(I) - I\|_2, \|\mathsf{T}'_\mathcal{L}(I) - I\|_2\}.$$

Proposition 13 *Let $\mathcal{L} : \mathbb{R}^{q \times q} \to \mathbb{R}^d$ be a linear map that satisfies (i) the restricted isometry condition $\delta_r(\mathcal{L}) \leq 1/2$, and (ii) whose nearly normalized parameter satisfies $\epsilon(\mathcal{L}) \leq 1/(650\sqrt{q})$. Let $\mathcal{L} \circ N_\mathcal{L}$ be the normalized linear map where $N_\mathcal{L}$ is a positive-definite rank-preserver. Then, $\mathcal{L} \circ N_\mathcal{L}$ satisfies the restricted isometry condition $\delta_r(\mathcal{L} \circ N) \leq \bar{\delta}_r := (1 + \delta_r(\mathcal{L}))(1 + 96\sqrt{q}\epsilon(\mathcal{L}))^2 - 1 < 1$. Moreover, $\|\mathcal{L} \circ N_\mathcal{L}\|_2 \leq (1 + 96\sqrt{q}\epsilon(\mathcal{L}))\|\mathcal{L}\|_2$.*

Proof (Proposition 13) Since $\mathcal{L}$ satisfies the restricted isometry condition $\delta_1(\mathcal{L}) \leq 1/2$, we have $|\langle \mathbf{vv}', \mathsf{T}_\mathcal{L}(\mathbf{uu}') \rangle - 1/q| \leq 1/(2q)$ for all unit-norm vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^q$. In addition, the linear map $\mathcal{L}$ has nearly normalized parameter $\epsilon(\mathcal{L}) \leq 1/(650\sqrt{q})$. Hence by applying Proposition 12 to the linear map $\mathsf{T}_\mathcal{L}$, any pair of positive-definite matrices Q_2, Q_1 such that $Q_2 \otimes Q_2 \circ \mathsf{T}_\mathcal{L} \circ Q_1 \otimes Q_1$ is doubly stochastic satisfies $\|Q_2 \otimes Q_1 - \mathsf{I}\|_2 \leq 96\sqrt{q}\epsilon(\mathcal{L})$. By noting the correspondence between such matrices with the positive-definite rank-preserver $N_\mathcal{L}$ such that $\mathcal{L} \circ N_\mathcal{L}$ is normalized via the relation $N_\mathcal{L} = Q_2 \otimes Q_1$ (see Corollary 2), we have $\|N_\mathcal{L}\|_2 \leq 1 + 96\sqrt{q}\epsilon(\mathcal{L})$.

Let X be a matrix with rank at most r. Then,

$$\|\mathcal{L}(N_\mathcal{L}(X))\|_{\ell_2} \leq \sqrt{1 + \delta_r(\mathcal{L})}\|N_\mathcal{L}\|_2\|X\|_{\ell_2} \leq \sqrt{1 + \delta_r(\mathcal{L})}(1 + 96\sqrt{q}\epsilon(\mathcal{L}))\|X\|_{\ell_2},$$

and hence $\|\mathcal{L}(N_{\mathcal{L}}(X))\|_{\ell_2}^2 \leq (1 + \bar{\delta}_r)\|X\|_{\ell_2}^2$. A similar set of steps show that $\|\mathcal{L}(N_{\mathcal{L}}(X))\|_{\ell_2}^2 \geq (1-\bar{\delta}_r)\|X\|_{\ell_2}^2$. Last $\|\mathcal{L} \circ N_{\mathcal{L}}\|_2 \leq \|\mathcal{L}\|_2\|N_{\mathcal{L}}\|_2 \leq (1+96\sqrt{q}\epsilon)\|\mathcal{L}\|_2$. $\qquad\square$

Proposition 14 [20, Theorem II.13] *Let $t > 0$ be fixed. Suppose $\mathcal{L} \sim \mathcal{N}(0, 1/d)$. Then, with probability greater than $1 - \exp(-t^2 d/2)$, we have $\|\mathcal{L}\|_2 \leq \sqrt{q^2/d} + 1 + t$.*

Proposition 15 [11, Theorem 2.3] *Let $0 < \delta < 1$ be fixed. There exists constants c_1, c_2 such that for $d \geq c_1 q r$, if $\mathcal{L} \sim \mathcal{N}(0, 1/d)$, then with probability greater than $1 - 2\exp(-c_2 d)$ the linear map $\mathcal{L}$ satisfies the restricted isometry condition $\delta_r(\mathcal{L}) \leq \delta$.*

Proposition 16 (Gaussian linear maps are nearly normalized) *Suppose $3/\sqrt{d} \leq \epsilon \leq 3$. Suppose $\mathcal{L} \sim \mathcal{N}(0, 1/d)$. Then with probability greater than $1 - 4\exp(-q(-1 + \sqrt{d}\epsilon/3)^2/2)$, the nearly normalized parameter of $\mathcal{L}$ is smaller than ϵ.*

Bounding the nearly normalized parameter of a Gaussian linear map exactly corresponds to computing the deviation of the sum of independent Wishart matrices from its mean in spectral norm. To do so, we appeal to the following concentration bound.

Proposition 17 (Concentration of sum of Wishart matrices) *Suppose $3/\sqrt{d} \leq t \leq 3$. Let $\{X^{(j)}\}_{j=1}^d$, $X^{(j)} = G^{(j)}G^{(j)\prime}$, where $G^{(j)} \in \mathbb{R}^{q\times q}$, $G^{(j)} \sim \mathcal{N}(0, 1/q)$, be a collection of independent Wishart matrices. Then, $\mathbb{P}(\|\frac{1}{d}\sum_{j=1}^d X^{(j)} - I\|_2 \geq t) \leq 2\exp(-q(-1 + \sqrt{d}t/3)^2/2)$.*

Proof (Proposition 17) Consider the linear map $G = [G^{(1)}|\ldots|G^{(d)}]$. Then, $\sum_{j=1}^d X^{(j)} = GG'$, and $\|\frac{1}{d}\sum_{j=1}^d X^{(j)} - I\|_2 \leq t$ if and only if $\sigma(G) \in [\sqrt{d(1-t)}, \sqrt{d(1+t)}]$. By [20, Theorem II.13], we have $\sigma(G) \in [\sqrt{d} - 1 - \tilde{t}, \sqrt{t} + 1 + \tilde{t}]$ with probability greater than $1 - 2\exp(-q\tilde{t}^2/2)$. The result follows with the choice of $\tilde{t} = -1 + \sqrt{d}t/3$. $\qquad\square$

Proof (Proposition 16) This is a direct application of Proposition 17 with $G^{(j)} = \sqrt{q/d}\mathcal{L}^{(j)}$ and $G^{(j)\prime} = \sqrt{q/d}\mathcal{L}^{(j)}$, followed by a union bound. $\qquad\square$

Proof (Proposition 7) We choose $t = 1/50$ in Proposition 14, $\delta = \delta_{4r}/2$ in Proposition 15, and $\epsilon = \delta/(960\sqrt{q})$ in Proposition 16. Then, there are constants c_1, c_2, c_3 such that if $d \geq c_1 r q$, then (i) $\|\tilde{\mathcal{L}}\|_2 \leq \sqrt{q^2/d} + 51/50 \leq (101/50)\sqrt{q^2/d}$, (ii) $\tilde{\mathcal{L}}$ satisfies the restricted isometry condition $\delta_{4r}(\tilde{\mathcal{L}}) \leq \delta_{4r}/2$, and (iii) $\tilde{\mathcal{L}}$ is nearly normalized with parameter $\epsilon(\tilde{\mathcal{L}}) \leq \delta_{4r}/960\sqrt{q}$, with probability greater than $1 - c_2\exp(-c_3 d)$.

By applying Proposition 13, we conclude that the linear map $\mathcal{L}$ satisfies the restricted isometry condition $\delta_{4r}(\mathcal{L}) \leq (1+\delta_{4r}/2)(1+\delta_{4r}/10)^2 - 1 \leq \delta_{4r}$, and $\|\mathcal{L}\|_2 \leq \sqrt{5q^2/d}$. $\qquad\square$

Appendix E: Proof of Proposition 2

Proof (Proposition 2) First, we check that the linear map $\mathcal{L}^\star \circ (I + E)$ satisfies the restricted isometry condition $\delta_1(\mathcal{L}^\star \circ (I+E)) \leq 1/2$. For any rank-one unit-norm matrix X, we have $\|[\mathcal{L}^\star \circ (I + E)](X)\|_{\ell_2} \leq \|\mathcal{L}^\star(X)\|_{\ell_2} + \|\mathcal{L}^\star(E(X))\|_{\ell_2} \leq \sqrt{1 + 1/10} +$

$1/150 \le \sqrt{1 + 1/2}$. A similar set of inequalities show that $\|[\mathcal{L}^\star \circ (I + E)](X)\|_{\ell_2} \ge \sqrt{1 - 1/2}$.

Second we check that the nearly normalized parameter of $\mathcal{L}^\star \circ (I + E)$ satisfies $\epsilon(\mathcal{L}^\star \circ (I + E)) \le 1/48\sqrt{q}$. Denote $\mathcal{E} := \mathcal{L}^\star \circ E$. For all unit-norm rank-one matrices E, we have $\|\mathcal{E}(E)\|_2^2 \le \|\mathcal{L}^\star\|_2^2 \|E\|_{\ell_2}^2$. Hence for any unit-norm $\mathbf{u} \in \mathbb{R}^q$, we have

$$\frac{1}{q} \sum_{j=1}^{d} \langle \mathcal{E}_j \mathcal{E}'_j, \mathbf{u}\mathbf{u}' \rangle = \frac{1}{q} \sum_{j=1}^{d} \sum_{k=1}^{q} (\mathcal{E}'_j \mathbf{u})_k^2 = \frac{1}{q} \sum_{k=1}^{q} \|\mathcal{E}(\mathbf{u}\mathbf{e}'_k)\|_{\ell_2}^2 \le \|\mathcal{L}^\star\|_2^2 \|E\|_{\ell_2}^2.$$

Using the fact that $\mathcal{L}^\star$ is normalized, we have

$$\frac{1}{q} \sum_{j=1}^{d} \langle \mathcal{L}^\star_j \mathcal{L}^{\star'}_j, \mathbf{u}\mathbf{u}' \rangle = 1.$$

By combining the previous inequalities with an application of Cauchy–Schwarz, we have

$$\begin{aligned}
&\langle \mathsf{T}_{\mathcal{L}^\star \circ (I+E)}(I) - I, \mathbf{u}\mathbf{u}' \rangle \\
&= \langle \mathsf{T}_{\mathcal{L}^\star + \mathcal{E}}(I) - \mathsf{T}_{\mathcal{L}^\star}(I), \mathbf{u}\mathbf{u}' \rangle \\
&= \frac{1}{q} \sum_{j=1}^{d} \langle \mathcal{E}_j \mathcal{E}'_j, \mathbf{u}\mathbf{u}' \rangle + \frac{1}{q} \sum_{j=1}^{d} \langle \mathcal{L}^\star_j \mathcal{E}'_j, \mathbf{u}\mathbf{u}' \rangle + \frac{1}{q} \sum_{j=1}^{d} \langle \mathcal{E}_j \mathcal{L}^{\star'}_j, \mathbf{u}\mathbf{u}' \rangle \\
&\le 3\|\mathcal{L}^\star\|_2 \|E\|_{\ell_2},
\end{aligned}$$

Furthermore, since $\mathbf{u}$ is arbitrary, it follows that

$$\|\mathsf{T}_{\mathcal{L}^\star + \mathcal{E}}(I) - I\|_2 \le 3\|\mathcal{L}^\star\|_2 \|E\|_{\ell_2}.$$

Using a similar sequence of steps, one can show that $\|\mathsf{T}'_{\mathcal{L}^\star + \mathcal{E}}(I) - I\|_2 \le 3\|\mathcal{L}^\star\|_2 \|E\|_{\ell_2}$. Thus, $\epsilon(\mathcal{L}^\star \circ (I + E)) \le 3\|\mathcal{L}^\star\|_2 \|E\|_{\ell_2} \le 1/(48\sqrt{q})$.

The result follows by applying Proposition 12 to the linear map $\mathsf{T}_{\mathcal{L}^\star \circ (I+E)}$. □

Appendix F: Proof of Proposition 3

The proof of Proposition 3 is based on the following result concerning affine rank minimization, which may be of independent interest.

Proposition 18 *Suppose $X^\star$ is a $q \times q$ rank-r matrix satisfying $\sigma_r(X^\star) \ge 1/2$. Let $\mathbf{y} = \mathcal{L}(X^\star) + \mathbf{z}$, where the linear map $\mathcal{L}$ satisfies the restricted isometry condition $\delta_{4r}(\mathcal{L}) \le 1/10$, and $\|\mathcal{L}'\mathbf{z}\|_2 =: \epsilon \le 1/(80r^{3/2})$. Let $\hat{X}$ be the optimal solution to*

$$\hat{X} = \underset{X}{\mathrm{argmin}} \; \|\mathbf{y} - \mathcal{L}(X)\|_{\ell_2}^2 \quad \text{s.t.} \quad \mathrm{rank}(X) \le r.$$

Then, (i) $\|\hat{X} - X^\star\|_2 \le 4\sqrt{r}\epsilon$, and (ii) $\hat{X} - X^\star = [(\mathcal{L}'_{\mathcal{T}(X^\star)}\mathcal{L}_{\mathcal{T}(X^\star)})^{-1}]_{\mathbb{R}^{q\times q}}(\mathcal{L}'_{\mathcal{T}(X^\star)}\mathbf{z}) + G$, where $\|G\|_{\ell_2} \le 340 r^{3/2}\epsilon^2$.

The proof of Proposition 18 requires two preliminary results which we state and prove first. Our development relies on results from matrix perturbation theory; we refer the reader to [38,57] for detailed expositions. Several of our results are minor modifications of analogous results in [15].

The following result and the accompanying proof is a minor modification of Proposition 2.2 in supplementary material (s.m.) of [15] and its proof. The modification allows us to provide a bound that does not scale with the ambient dimension.

Proposition 19 *Let $X_1, X_2 \in \mathbb{R}^{q\times q}$ be rank-r matrices. Let σ be the smallest nonzero singular value of X_1, and suppose that $\|X_1 - X_2\|_2 \le \sigma/8$. Then, $\|\mathcal{P}_{\mathcal{T}(X_1)^\perp}(X_2)\|_{\ell_2} \le \sqrt{r}\|X_1 - X_2\|_2^2/(3\sigma)$, and $\|\mathcal{P}_{\mathcal{T}(X_1)^\perp}(X_2)\|_2 \le \|X_1 - X_2\|_2^2/(5\sigma)$.*

In the following proof, given a matrix $X \in \mathbb{R}^{q\times q}$, we denote $\tilde{X} := \begin{pmatrix} 0 & X' \\ X & 0 \end{pmatrix}$.

Proof (Proposition 19) Let $\tilde{\Delta} = \tilde{X}_2 - \tilde{X}_1$, and let $\kappa = \sigma/4$. By combining equation (1.5) in the s.m. of [15] with the proofs of Propositions 1.2 and 2.2 in the s.m. of [15] it can be shown that $\mathcal{P}_{\mathcal{T}(\tilde{X}_1)^\perp}(\tilde{X}_2) = (1/(2\pi i))\oint_{\mathcal{C}_\kappa} \zeta[\tilde{X}_1 - \zeta I]^{-1}\tilde{\Delta}[\tilde{X}_1 - \zeta I]^{-1}\tilde{\Delta}[\tilde{X}_2 - \zeta I]^{-1}d\zeta$, where the contour integral is taken along $\mathcal{C}_\kappa$ defined as the circle centered at the origin with radius κ.

By a careful use of the inequality $\|AB\|_{\ell_2} \le \|A\|_2\|B\|_{\ell_2}$, we have $\|[\tilde{X}_1 - \zeta I]^{-1}\tilde{\Delta}[\tilde{X}_1 - \zeta I]^{-1}\tilde{\Delta}[\tilde{X}_2 - \zeta I]^{-1}\|_{\ell_2} \le \|[\tilde{X}_1 - \zeta I]^{-1}\|_2\|\tilde{\Delta}\|_{\ell_2}\|[\tilde{X}_1 - \zeta I]^{-1}\|_2\|\tilde{\Delta}\|_2 \|[\tilde{X}_2 - \zeta I]^{-1}\|_2$. Since $\tilde{\Delta}$ is a matrix with rank at most $4r$, we have $\|\tilde{\Delta}\|_{\ell_2} \le \sqrt{4r}\|\tilde{\Delta}\|_2$. We proceed to apply the same bounds as those used in the proof of Proposition 1.2 in the s.m. of [15] to obtain $\|\mathcal{P}_{\mathcal{T}(\tilde{X}_1)^\perp}(\tilde{X}_2)\|_{\ell_2} \le 2\sqrt{r}\kappa^2\|\tilde{\Delta}\|_2^2/((\sigma - \kappa)^2(\sigma - 3\kappa/2)) \le \sqrt{2r}\|\tilde{X}_1 - \tilde{X}_2\|_2^2/(3\sigma)$. The first inequality follows by noting that $\sqrt{2}\|\mathcal{P}_{\mathcal{T}(X_1)^\perp}(X_2)\|_{\ell_2} = \|\mathcal{P}_{\mathcal{T}(\tilde{X}_1)^\perp}(\tilde{X}_2)\|_{\ell_2}$ and that $\|X_1 - X_2\|_2 = \|\tilde{X}_1 - \tilde{X}_2\|_2$.

The proof of the second inequality follows from a similar argument. $\square$

We define the following distance measure between two subspaces $\mathcal{T}_1$ and $\mathcal{T}_2$ [15]

$$\rho(\mathcal{T}_1, \mathcal{T}_2) := \sup_{\|N\|_2 \le 1} \|\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2}(N)\|_2.$$

This definition is useful for quantifying the distance between tangent spaces with respect to the variety of low-rank matrices for pairs of nearby matrices.

Lemma 6 *Let $X_1, X_2 \in \mathbb{R}^{q\times q}$ be matrices with rank at most r, and satisfy $\|X_1 - X_2\|_2 \le \sigma/8$, where σ is the smallest nonzero singular value of X_2. Let $\mathcal{T}_1 := \mathcal{T}(X_1)$ and $\mathcal{T}_2 := \mathcal{T}(X_2)$ be tangent spaces on the variety of matrices with rank at most r at the points X_1 and X_2, respectively. Let $\mathcal{L}$ be a linear map satisfying the restricted isometry condition $\delta_{4r}(\mathcal{L}) \le 1/10$. If $Z_i \in \mathcal{T}_i$, $i \in \{1, 2\}$, then $\|[(\mathcal{L}'_{\mathcal{T}_1}\mathcal{L}_{\mathcal{T}_1})^{-1}]_{\mathbb{R}^{q\times q}}(Z_1) - [(\mathcal{L}'_{\mathcal{T}_2}\mathcal{L}_{\mathcal{T}_2})^{-1}]_{\mathbb{R}^{q\times q}}(Z_2)\|_{\ell_2} \le (43/10)\sqrt{r}\|Z_1 - Z_2\|_2 + 16r\|X_1 - X_2\|_2\|Z_2\|_2/\sigma$.*

Proof (Lemma 6) To simplify notation, we denote $Y_i = [(\mathcal{L}'_{\mathcal{T}_i}\mathcal{L}_{\mathcal{T}_i})^{-1}]_{\mathbb{R}^{q \times q}}(Z_i)$, $i \in \{1, 2\}$. From the triangle inequality, we have $\|Y_1 - Y_2\|_{\ell_2} \leq \|\mathcal{P}_{\mathcal{T}_1^\perp}(Y_1 - Y_2)\|_{\ell_2} + \|\mathcal{P}_{\mathcal{T}_1}(Y_1 - Y_2)\|_{\ell_2}$. We bound both components separately.

$[\|\mathcal{P}_{\mathcal{T}_1^\perp}(Y_1 - Y_2)\|_{\ell_2}]$: From Proposition 2.1 of the s.m. of [15], we have $\rho(\mathcal{T}_1, \mathcal{T}_2) \leq \frac{2}{\sigma}\|X_1 - X_2\|_2$. From Lemma 1, we have $\|Y_2 - Z_2\|_{\ell_2} \leq \delta_{4r}\|Y_2\|_{\ell_2} \leq \frac{\delta_{4r}}{1-\delta_{4r}}\|Z_2\|_{\ell_2} \leq \frac{\sqrt{2r}\delta_{4r}}{1-\delta_{4r}}\|Z_2\|_2$. Hence

$$
\begin{aligned}
\|\mathcal{P}_{\mathcal{T}_1^\perp}(Y_2 - Z_2)\|_{\ell_2} &= \|[I - \mathcal{P}_{\mathcal{T}_1}]([\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2}](Y_2 - Z_2))\|_{\ell_2} \\
&\leq 2\sqrt{r}\|[\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2}](Y_2 - Z_2)\|_2 \\
&\leq 2\sqrt{r}\rho(\mathcal{T}_1, \mathcal{T}_2)\|Y_2 - Z_2\|_2 \\
&\leq \frac{4\sqrt{2}r}{\sigma}\frac{\delta_{4r}}{1-\delta_{4r}}\|X_1 - X_2\|_2\|Z_2\|_2.
\end{aligned}
$$

Here the first inequality follows by noting that $[I - \mathcal{P}_{\mathcal{T}_1}]([\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2}](Y_2 - Z_2))$ has rank at most $4r$. Next

$$
\|\mathcal{P}_{\mathcal{T}_1^\perp}(Z_2)\|_{\ell_2} = \|\mathcal{P}_{\mathcal{T}_1^\perp}(Z_1 - Z_2)\|_{\ell_2} \leq \|Z_1 - Z_2\|_{\ell_2} \leq 2\sqrt{r}\|Z_1 - Z_2\|_2.
$$

By combining both bounds with the triangle inequality, we obtain

$$
\begin{aligned}
\|\mathcal{P}_{\mathcal{T}_1^\perp}(Y_1 - Y_2)\|_{\ell_2} = \|\mathcal{P}_{\mathcal{T}_1^\perp}(Y_2)\|_{\ell_2} &\leq \|\mathcal{P}_{\mathcal{T}_1^\perp}(Z_2)\|_{\ell_2} + \|\mathcal{P}_{\mathcal{T}_1^\perp}(Y_2 - Z_2)\|_{\ell_2} \\
&\leq 2\sqrt{r}\|Z_1 - Z_2\|_2 + \frac{4\sqrt{2}r}{\sigma}\frac{\delta_{4r}}{1-\delta_{4r}}\|X_1 \\
&\quad - X_2\|_2\|Z_2\|_2.
\end{aligned}
$$

$[\|\mathcal{P}_{\mathcal{T}_1}(Y_1 - Y_2)\|_{\ell_2}]$: Define the linear map $\mathsf{G} = \mathcal{L}'_{\mathcal{T}_1 \cup \mathcal{T}_2}\mathcal{L}_{\mathcal{T}_1 \cup \mathcal{T}_2}$. First $\|[\mathcal{P}_{\mathcal{T}_2} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2) - [\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2)\|_{\ell_2} \leq 2\sqrt{r}\|[\mathcal{P}_{\mathcal{T}_2} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2) - [\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2)\|_2 \leq 2\sqrt{r}\rho(\mathcal{T}_1, \mathcal{T}_2)\|\mathsf{G}(Y_2)\|_2$, where $\|\mathsf{G}(Y_2)\|_2 \leq \|\mathsf{G}(Y_2)\|_{\ell_2} \leq (1 + \delta_{4r})\|Y_2\|_{\ell_2} \leq \frac{1+\delta_{4r}}{1-\delta_{4r}}\|Z_2\|_{\ell_2} \leq \sqrt{2r}\frac{1+\delta_{4r}}{1-\delta_{4r}}\|Z_2\|_2$. Second $\|[\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2) - [\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_1}](Y_2)\|_{\ell_2} = \|[\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ (\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2})](Y_2)\|_{\ell_2} \leq \|[\mathsf{G} \circ (\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2})](Y_2)\|_{\ell_2} \leq (1 + \delta_{4r})\|[\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2}](Y_2)\|_{\ell_2} \leq 2\sqrt{r}(1 + \delta_{4r})\|[\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2}](Y_2)\|_2 \leq 2\sqrt{r}(1 + \delta_{4r})\rho(\mathcal{T}_1, \mathcal{T}_2)\|Y_2\|_2$, where $\|Y_2\|_2 \leq \|Y_2\|_{\ell_2} \leq \frac{\sqrt{2r}}{1-\delta_{4r}}\|Z\|_2$. Third by combining these bounds with an application of Lemma 1 and the triangle inequality, we obtain

$$
\begin{aligned}
&\|\mathcal{P}_{\mathcal{T}_1}(Y_1 - Y_2)\|_{\ell_2} \\
&\leq \frac{1}{1-\delta_{4r}}\|[\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_1}](Y_1 - Y_2)\|_{\ell_2} \\
&\leq \frac{1}{1-\delta_{4r}}(\|[\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_1}](Y_1) - [\mathcal{P}_{\mathcal{T}_2} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2)\|_{\ell_2} \\
&\quad + \|[\mathcal{P}_{\mathcal{T}_2} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2) - [\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2)\|_{\ell_2} \\
&\quad + \|[\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2) - [\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_1}](Y_2)\|_{\ell_2})
\end{aligned}
$$

$$\leq \frac{1}{1-\delta_{4r}}(2\sqrt{r}\|Z_1 - Z_2\|_2 + 4\sqrt{2}r\rho(\mathcal{T}_1, \mathcal{T}_2)\frac{1+\delta_{4r}}{1-\delta_{4r}}\|Z\|_2).$$

$\square$

Proof (Proposition 18) We prove (i) and (ii) in sequence.

[(i)]: Let $\hat{X}_o$ be the optimal solution to the following

$$\hat{X}_o = \underset{X}{\arg\min} \|\mathbf{y} - \mathcal{L}(X)\|_{\ell_2}^2 \quad \text{s.t.} \quad \text{rank}(X) \leq r, \quad \|X - X^\star\|_2 \leq 4\sqrt{r}\epsilon.$$

Since $4\sqrt{r}\epsilon < 1/2 \leq \sigma_r(X^\star)$, $\hat{X}_o$ has rank exactly r, and hence is a smooth point with respect to the variety of matrices with rank at most r. Define the tangent space $\hat{\mathcal{T}} := \mathcal{T}(\hat{X}_o)$, and the matrix $\hat{X}_c$ as the solution to the following optimization instance

$$\hat{X}_c = \underset{X}{\arg\min} \|\mathbf{y} - \mathcal{L}(X)\|_2^2 \quad \text{s.t.} \quad X \in \hat{\mathcal{T}}, \quad \|X - X^\star\|_2 \leq 4\sqrt{r}\epsilon.$$

Here $\hat{X}_c$ is the solution to the optimization instance where the constraint $X \in \hat{\mathcal{T}}$, which is convex, replaces the only non-convex constraint in the previous optimization instance. Hence $\hat{X}_c = \hat{X}_o$. Define $\hat{X}_{\hat{\mathcal{T}}}$ as the solution to the following optimization instance

$$\hat{X}_{\hat{\mathcal{T}}} = \underset{X}{\arg\min} \|\mathbf{y} - \mathcal{L}(X)\|_{\ell_2}^2 \quad \text{s.t.} \quad X \in \hat{\mathcal{T}}.$$

The first-order condition is given by $\mathcal{L}'\mathcal{L}(\hat{X}_{\hat{\mathcal{T}}} - X^\star) - \mathcal{L}'\mathbf{z} + Q_{\hat{\mathcal{T}}^\perp} = 0$, where $Q_{\hat{\mathcal{T}}^\perp} \in \hat{\mathcal{T}}^\perp$ is the Lagrange multiplier associated with the constraint $X \in \hat{\mathcal{T}}$. Project the above equation onto the subspace $\hat{\mathcal{T}}$ to obtain $[\mathcal{P}_{\hat{\mathcal{T}}} \circ \mathcal{L}'\mathcal{L} \circ \mathcal{P}_{\hat{\mathcal{T}}}](\hat{X}_{\hat{\mathcal{T}}} - X^\star) = [\mathcal{P}_{\hat{\mathcal{T}}} \circ \mathcal{L}'\mathcal{L} \circ \mathcal{P}_{\hat{\mathcal{T}}^\perp}](X^\star) + \mathcal{P}_{\hat{\mathcal{T}}}(\mathcal{L}'\mathbf{z})$, and hence

$$\hat{X}_{\hat{\mathcal{T}}} - X^\star = [(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{q\times q}} \circ \left([\mathcal{L}'\mathcal{L} \circ \mathcal{P}_{\hat{\mathcal{T}}^\perp}](X^\star) + \mathcal{L}'\mathbf{z}\right) - \mathcal{P}_{\hat{\mathcal{T}}^\perp}(X^\star).$$

We proceed to bound $\|\hat{X}_{\hat{\mathcal{T}}} - X^\star\|_2$. First, we have $\|\hat{X}_c - X^\star\|_2 \leq 4\sqrt{r}\epsilon \leq 1/20$, and hence $\sigma_r(\hat{X}_c) \geq 9/20$. Second by applying Proposition 19, we have $\|\mathcal{P}_{\hat{\mathcal{T}}^\perp}(X^\star)\|_2 = \|\mathcal{P}_{\hat{\mathcal{T}}^\perp}(\hat{X}_c - X^\star)\|_2 \leq (4\sqrt{r}\epsilon)^2/(5\sigma_r(\hat{X}_c)) \leq (64/9)r\epsilon^2$, and $\|\mathcal{P}_{\hat{\mathcal{T}}^\perp}(X^\star)\|_{\ell_2} \leq (320/27)r^{3/2}\epsilon^2$. Third by Lemma 1 and noting the inequality $\|\cdot\|_2 \leq \|\cdot\|_{\ell_2}$, we have

$$\|[(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{q\times q}}(\mathcal{L}'\mathbf{z})\|_2 \leq \|[(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{q\times q}}\|_2 \|\mathcal{P}_{\hat{\mathcal{T}}}(\mathcal{L}'\mathbf{z})\|_{\ell_2}$$
$$\leq 2\sqrt{2r}\|\mathcal{L}'\mathbf{z}\|_2/(1-\delta_{4r}) \leq (16/5)\sqrt{r}\epsilon.$$

Fourth by Proposition 2.7 in [30], we have

$$\|[[(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{q\times q}} \circ \mathcal{L}'\mathcal{L} \circ \mathcal{P}_{\hat{\mathcal{T}}^\perp}](X^\star)\|_2$$
$$\leq \|[(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{q\times q}}\|_2 \|[\mathcal{P}_{\hat{\mathcal{T}}} \circ \mathcal{L}'\mathcal{L} \circ \mathcal{P}_{\hat{\mathcal{T}}^\perp}](X^\star_{\hat{\mathcal{T}}^\perp})\|_{\ell_2}$$
$$\leq \delta_{4r}\|\mathcal{P}_{\hat{\mathcal{T}}^\perp}(X^\star)\|_{\ell_2}/(1-\delta_{4r}) \leq (320/243)r^{3/2}\epsilon^2.$$

Last, we combine the bounds to obtain $\|\hat{X}_{\hat{\mathcal{T}}} - X^\star\|_2 \le 8r\epsilon^2 + (16/5)\sqrt{r}\epsilon + 2r^{3/2}\epsilon^2 < 4\sqrt{r}\epsilon$. This implies that the constraint $\|X - X^\star\|_2 \le 4\sqrt{r}\epsilon$ for $\hat{X}_c$ and $\hat{X}_o$ are inactive, and hence $\hat{X} = \hat{X}_o = \hat{X}_c = \hat{X}_{\hat{\mathcal{T}}}$.

[(ii)]: We have

$$G = [(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{q \times q}}(\mathcal{L}'\mathbf{z}) - [(\mathcal{L}'_{\mathcal{T}^\star}\mathcal{L}_{\mathcal{T}^\star})^{-1}]_{\mathbb{R}^{q \times q}}(\mathcal{L}'\mathbf{z})$$
$$+ [[(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{q \times q}} \circ \mathcal{L}'\mathcal{L} \circ \mathcal{P}_{\hat{\mathcal{T}}\perp}](X^\star) - \mathcal{P}_{\hat{\mathcal{T}}\perp}(X^\star).$$

We deal with the contributions of each term separately.

First $\|[\mathcal{P}_{\mathcal{T}^\star} - \mathcal{P}_{\hat{\mathcal{T}}}](\mathcal{L}'\mathbf{z})\|_2 \le \rho(\hat{\mathcal{T}}, \mathcal{T}^\star)\|\mathcal{L}'\mathbf{z}\|_2 \le (2\epsilon/\sigma_r(X^\star))\|\hat{X} - X^\star\|_2 \le 16\sqrt{r}\epsilon^2$, where the second inequality applies Proposition 2.1 of the s.m. of [15]. Second $\|\mathcal{P}_{\mathcal{T}^\star}(\mathcal{L}'\mathbf{z})\|_2 \le 2\|\mathcal{L}'\mathbf{z}\|_2 = 2\epsilon$. Hence by applying Lemma 6 with the choice of $Z_1 = \mathcal{P}_{\hat{\mathcal{T}}}(\mathcal{L}'\mathbf{z})$ and $Z_2 = \mathcal{P}_{\mathcal{T}^\star}(\mathcal{L}'\mathbf{z})$, we obtain $\|[(\mathcal{L}'_{\mathcal{T}^\star}\mathcal{L}_{\mathcal{T}^\star})^{-1}]_{\mathbb{R}^{q \times q}}(\mathcal{L}'\mathbf{z}) - [(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{q \times q}}(\mathcal{L}'\mathbf{z})\|_{\ell_2} \le 70r\epsilon^2 + 256r^{3/2}\epsilon^2$. Third, we have $\|[(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{q \times q}} \circ \mathcal{L}'\mathcal{L} \circ \mathcal{P}_{\hat{\mathcal{T}}\perp}](X^\star)\|_{\ell_2} \le (320/243)r^{3/2}\epsilon^2$, and $\|\mathcal{P}_{\hat{\mathcal{T}}\perp}(X^\star)\|_{\ell_2} \le (320/27)r^{3/2}\epsilon^2$.

The bound follows by summing up these bounds. $\qquad\square$

The proof of Proposition 3 requires two additional preliminary results; in particular, the first establishes the restricted isometry condition for linear maps that are near linear maps that already satisfy the restricted isometry condition.

Proposition 20 *Suppose $\mathcal{L}^\star$ is a linear map that satisfies the restricted isometry condition $\delta_r(\mathcal{L}^\star) \le 1/20$. Let E be a linear operator such that $\|\mathsf{E}\|_2 \le 1/(50\|\mathcal{L}^\star\|_2)$. Then, $\mathcal{L} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E})$ satisfies the restricted isometry condition $\delta_r(\mathcal{L}) \le 1/10$.*

Proof (Proposition 20) Let X be a matrix with rank at most r. Then,

$$\|\mathcal{L}(X)\|_{\ell_2} \le \|\mathcal{L}^\star(X)\|_{\ell_2} + \|\mathcal{L}^\star(\mathsf{E}(X))\|_{\ell_2} \le (\sqrt{1 + \delta_r(\mathcal{L}^\star)} + 1/50)\|X\|_{\ell_2}$$
$$\le \sqrt{1 + 1/10}\|X\|_{\ell_2}.$$

A similar argument also proves the lower bound $\|\mathcal{L}(X)\|_{\ell_2} \ge \sqrt{1 - 1/10}\|X\|_{\ell_2}$. $\quad\square$

Lemma 7 *Suppose $\mathcal{L}$ satisfies the restricted isometry condition $\delta_1(\mathcal{L}) < 1$. Then, $\|\mathcal{L}'\mathcal{L}\|_{\ell_2,2} \le \sqrt{2(1 + \delta_1(\mathcal{L}))}\|\mathcal{L}\|_2$.*

Proof Let $Z \in \arg\max_{X:\|X\|_{\ell_2} \le 1}\|\mathcal{L}'\mathcal{L}(X)\|_2$, and let $\mathcal{T}$ be the tangent space of the rank-one matrix corresponding to the largest singular value of Z. Then, $\sup_{X:\|X\|_{\ell_2} \le 1}\|\mathcal{L}'\mathcal{L}(X)\|_2 \le \sup_{X:\|X\|_{\ell_2} \le 1}\|[\mathcal{P}_{\mathcal{T}} \circ \mathcal{L}'\mathcal{L}](X)\|_2 \le \sqrt{2}\sup_{X:\|X\|_{\ell_2} \le 1}\|[\mathcal{P}_{\mathcal{T}} \circ \mathcal{L}'\mathcal{L}](X)\|_{\ell_2} \le \sqrt{2}\|\mathcal{P}_{\mathcal{T}} \circ \mathcal{L}'\mathcal{L}\|_2$. By Lemma 1, we have $\sqrt{2}\|\mathcal{P}_{\mathcal{T}} \circ \mathcal{L}'\mathcal{L}\|_2 \le \sqrt{2(1 + \delta_1(\mathcal{L}))}\|\mathcal{L}\|_2$. $\qquad\square$

Proof (Proposition 3) To simplify notation, we denote $\mathcal{T} := \mathcal{T}(X^\star)$. Without loss of generality, we may assume that $\|X^\star\|_2 = 1$. By the triangle inequality, we have

$$
\|(X^\star - \mathcal{M}(\hat{X})) - [(\mathcal{L}_{\mathcal{T}}^{\star\prime}\mathcal{L}_{\mathcal{T}}^\star)^{-1}]_{\mathbb{R}^{q\times q}} \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}(X^\star)\|_{\ell_2}
$$
$$
\leq \ \|(X^\star - \mathcal{M}(\hat{X})) - [((\mathsf{I} + \mathsf{E}') \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ
$$
$$
(\mathsf{I} + \mathsf{E})|_{\mathcal{T}})^{-1}]_{\mathbb{R}^{q\times q}} \circ (\mathsf{I} + \mathsf{E}') \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}(X^\star)\|_{\ell_2}
$$
$$
+ \ \|([((\mathsf{I} + \mathsf{E}') \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E})|_{\mathcal{T}})^{-1}]_{\mathbb{R}^{q\times q}}
$$
$$
- [(\mathcal{L}_{\mathcal{T}}^{\star\prime}\mathcal{L}_{\mathcal{T}}^\star)^{-1}]_{\mathbb{R}^{q\times q}}) \circ (\mathsf{I} + \mathsf{E}') \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}(X^\star)\|_{\ell_2}
$$
$$
+ \ \|[(\mathcal{L}_{\mathcal{T}}^{\star\prime}\mathcal{L}_{\mathcal{T}}^\star)^{-1}]_{\mathbb{R}^{q\times q}} \circ (\mathsf{I} + \mathsf{E}') \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}(X^\star)
$$
$$
- [(\mathcal{L}_{\mathcal{T}}^{\star\prime}\mathcal{L}_{\mathcal{T}}^\star)^{-1}]_{\mathbb{R}^{q\times q}} \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}(X^\star)\|_{\ell_2}
$$

We bound each term separately.

[First term]: Let $\tilde{\mathbf{z}} := [\mathcal{L}^\star \circ \mathsf{E}](X^\star)$. First by Proposition 20, the linear map $\mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E})$ satisfies the restricted isometry condition $\delta_{4r}(\mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E})) \leq 1/10$. Second, we have $\|\mathsf{I} + \mathsf{E}\|_{2,2} \leq 1 + \sqrt{q}\|\mathsf{E}\|_{\ell_2} \leq 51/50$. Third from Lemma 7, we have $\|\mathcal{L}^{\star\prime}\mathcal{L}^\star\|_{\ell_2,2} \leq \sqrt{2(1 + \delta_{4r}(\mathcal{L}^\star))}\|\mathcal{L}^\star\|_2$. Fourth $\|\mathsf{E}(X^\star)\|_{\ell_2} \leq \sqrt{r}\|\mathsf{E}\|_{\ell_2}$. Hence

$$
\|(\mathsf{I} + \mathsf{E}') \circ \mathcal{L}^{\star\prime}\tilde{\mathbf{z}}\|_2 \leq \|\mathsf{I} + \mathsf{E}\|_{2,2}\|\mathcal{L}^{\star\prime}\mathcal{L}^\star\|_{\ell_2,2}\|\mathsf{E}\|_{\ell_2}\|X^\star\|_2 \leq (3/2)\sqrt{r}\|\mathcal{L}^\star\|_2\|\mathsf{E}\|_{\ell_2}.
$$

By the initial conditions, we have that the above quantity is at most $1/(80r^{3/2})$. Consequently, by applying Proposition 18 to optimization instance (9) with the choice of linear map $\mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E})$ and error term $\tilde{\mathbf{z}}$ we have

$$
\|(X^\star - \mathcal{M}(\hat{X})) - [((\mathsf{I} + \mathsf{E}') \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E})|_{\mathcal{T}})^{-1}]_{\mathbb{R}^{q\times q}} \circ (\mathsf{I} + \mathsf{E}') \circ \mathcal{L}^{\star\prime}\tilde{\mathbf{z}}\|_{\ell_2}
$$
$$
\leq 765 r^{5/2}\|\mathcal{L}^\star\|_2^2\|\mathsf{E}\|_{\ell_2}^2.
$$

[Second term]: First by Lemma 1, we have $\|[(\mathcal{L}_{\mathcal{T}}^{\star\prime}\mathcal{L}_{\mathcal{T}}^\star)^{-1}]_{\mathbb{R}^{q\times q}}\|_2 \leq 20/19$. Second by the triangle inequality, we have $\|\mathcal{P}_{\mathcal{T}} \circ (\mathsf{I} + \mathsf{E}') \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}) \circ \mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}} \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathcal{P}_{\mathcal{T}}\|_2 \leq 3\|\mathcal{L}^\star\|_2\|\mathsf{E}\|_{\ell_2}$. Third by utilizing the identity $(A + B)^{-1} = A^{-1} - A^{-1} \circ B \circ A^{-1} + A^{-1} \circ B \circ A^{-1} \circ B \circ A^{-1} - \ldots$ with the choice of $A = \mathcal{L}_{\mathcal{T}}^{\star\prime}\mathcal{L}_{\mathcal{T}}^\star$ and $B = \mathcal{P}_{\mathcal{T}} \circ (\mathsf{I} + \mathsf{E})' \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}) \circ \mathcal{P}_{\mathcal{T}} - A$, we obtain

$$
\|[((\mathsf{I} + \mathsf{E}') \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E})|_{\mathcal{T}})^{-1}]_{\mathbb{R}^{q\times q}} - [(\mathcal{L}_{\mathcal{T}}^{\star\prime}\mathcal{L}_{\mathcal{T}}^\star)^{-1}]_{\mathbb{R}^{q\times q}}\|_2 \leq 4\|\mathcal{L}^\star\|_2\|\mathsf{E}\|_{\ell_2}.
$$

Fourth $\|\mathcal{P}_{\mathcal{T}} \circ (\mathsf{I} + \mathsf{E}') \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}(X^\star)\|_{\ell_2} \leq (11/10)\sqrt{r}\|\mathcal{L}^\star\|_2\|\mathsf{E}\|_{\ell_2}$. Hence

$$
\|([((\mathsf{I} + \mathsf{E}') \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E})|_{\mathcal{T}})^{-1}]_{\mathbb{R}^{q\times q}}
$$
$$
- [(\mathcal{L}_{\mathcal{T}}^{\star\prime}\mathcal{L}_{\mathcal{T}}^\star)^{-1}]_{\mathbb{R}^{q\times q}}) \circ (\mathsf{I} + \mathsf{E}') \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}(X^\star)\|_{\ell_2}
$$
$$
\leq 5\sqrt{r}\|\mathcal{L}^\star\|_2^2\|\mathsf{E}\|_{\ell_2}^2.
$$

[Third Term]: We have

$$\|[(\mathcal{L}_{\mathcal{T}}^{\star\prime}\mathcal{L}_{\mathcal{T}}^{\star})^{-1}]_{\mathbb{R}^{q\times q}} \circ (\mathsf{I}+\mathsf{E}') \circ \mathcal{L}^{\star\prime}\mathcal{L}^{\star} \circ \mathsf{E}(X^{\star})$$
$$-[(\mathcal{L}_{\mathcal{T}}^{\star\prime}\mathcal{L}_{\mathcal{T}}^{\star})^{-1}]_{\mathbb{R}^{q\times q}} \circ \mathcal{L}^{\star\prime}\mathcal{L}^{\star} \circ \mathsf{E}(X^{\star})\|_{\ell_2}$$
$$\leq \|[(\mathcal{L}_{\mathcal{T}}^{\star\prime}\mathcal{L}_{\mathcal{T}}^{\star})^{-1}]_{\mathbb{R}^{q\times q}}\|_2 \|\mathsf{E}'\|_2 \|\mathcal{L}^{\star}\|_2^2 \|\mathsf{E}(X^{\star})\|_{\ell_2} \leq 2\sqrt{r}\|\mathcal{L}^{\star}\|_2^2 \|\mathsf{E}\|_{\ell_2}^2.$$

[Conclude]: The result follows by summing each bound and applying Lemma 7. $\square$

Appendix G: Proof of Proposition 4

Proof (Proposition 4) To simplify notation, we let $\Lambda := \Lambda(\{A^{(j)}\}_{j=1}^n)$, $\Delta := \Delta(\{A^{(j)}\}_{j=1}^n)$, and $\mathbf{D}$ be the linear map defined as $\mathbf{D}: \mathbf{z} \mapsto \sum_{j=1}^n (\mathsf{Q}(B^{(j)}) - A^{(j)})\mathbf{z}_j$. In addition we define $\tau := (1/\sqrt{n\Lambda})\|\mathbf{D}\|_2$. Note that by the Cauchy–Schwarz inequality, we have $\tau \leq \omega/\sqrt{\Lambda} \leq 1/20$.

We begin by noting that since $\|(1/n\Lambda)\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime} - \mathsf{I}\|_2 \leq \Delta/\Lambda \leq 1/6$, we have $\|((1/n\Lambda)\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1}\|_2$, and $\|(1/n\Lambda)\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime}\|_2 \leq 6/5$.

Next we compute the following bounds. First $\|\mathbf{D} \circ \mathbf{D}' \circ (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1}\|_2 \leq \|\mathbf{D}\|_2^2 \|(\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1}\|_2 \leq (6/5)\tau^2$. Second $\|\mathbf{D} \circ \mathbf{X}^{\star\prime} \circ (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1}\|_2 \leq \|\mathbf{D}\|_2 \|\mathbf{X}^{\star\prime}\|_2 \|(\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1}\|_2 \leq \tau(6/5)^{3/2}$. Third $\|\mathbf{X}^{\star} \circ \mathbf{D}' \circ (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1}\|_2 \leq \tau(6/5)^{3/2}$. By applying these bounds to the following expansion, we obtain

$$\left((\mathbf{X}^{\star}+\mathbf{D}) \circ (\mathbf{X}^{\star}+\mathbf{D})'\right)^{-1}$$
$$= \left((\mathsf{I}+\mathbf{D} \circ \mathbf{X}^{\star\prime} \circ (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1} + \mathbf{X}^{\star} \circ \mathbf{D}' \circ (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1} + \mathsf{E}_1) \circ \mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime}\right)^{-1}$$
$$= (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1}\left(\mathsf{I} - \mathbf{D} \circ \mathbf{X}^{\star\prime} \circ (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1} - \mathbf{X}^{\star} \circ \mathbf{D}' \circ (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1} + \mathsf{E}_2\right),$$

where $\|\mathsf{E}_1\|_2 \leq (6/5)\tau^2$, and $\|\mathsf{E}_2\|_2 = \|-\mathsf{E}_1 + (\mathbf{D} \circ \mathbf{X}^{\star\prime} \circ (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1} + \mathbf{X}^{\star} \circ \mathbf{D}' \circ (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1} + \mathsf{E}_1)^2 - (\ldots)^3\|_2 \leq (\|\mathsf{E}_1\|_2 + \|\mathbf{D} \circ \mathbf{X}^{\star\prime} \circ (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1} + \mathbf{X}^{\star} \circ \mathbf{D}' \circ (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1} + \mathsf{E}_1\|_2^2 + \ldots) \leq (6/5)\tau^2 + (\tau(6/5)(\tau + 2\sqrt{6/5}))^2 + \ldots \leq 1.2\tau^2 + (3\tau)^2 + (3\tau)^3 + \ldots \leq 12\tau^2$.

We apply the above expansion to derive the following approximation of $\mathbf{X}^{\star} \circ (\mathbf{X}^{\star} + \mathbf{D})^+$

$$\mathbf{X}^{\star} \circ (\mathbf{X}^{\star}+\mathbf{D})^+$$
$$= \mathbf{X}^{\star} \circ (\mathbf{X}^{\star}+\mathbf{D})' \circ \left((\mathbf{X}^{\star}+\mathbf{D}) \circ (\mathbf{X}^{\star}+\mathbf{D})'\right)^{-1}$$
$$= (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime} + \mathbf{X}^{\star} \circ \mathbf{D}') \circ (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1}\left(\mathsf{I} - \mathbf{D} \circ \mathbf{X}^{\star\prime} \circ (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1}\right.$$
$$\left. - \mathbf{X}^{\star} \circ \mathbf{D}' \circ (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1} + \mathsf{E}_2\right)$$
$$= (\mathsf{I} - \mathbf{D} \circ \mathbf{X}^{\star+} + \mathsf{E}_3),$$

where E_3 satisfies $\|\mathsf{E}_3\|_2 \leq 2(\tau(6/5)^{3/2})(2(\tau(6/5)^{3/2}) + \|\mathsf{E}_2\|_2) + \|\mathsf{E}_2\|_2 \leq 20\tau^2$.

Next we write $((1/(n\Lambda))\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1} = \mathsf{I} + \mathsf{E}_4$, where $\|\mathsf{E}_4\|_2 \leq (6/5)\Delta/\Lambda$. Then,

$$\mathbf{X}^{\star} \circ (\mathbf{X}^{\star}+\mathbf{D})^+ = \mathsf{I} - \mathbf{D} \circ \mathbf{X}^{\star\prime} \circ (\mathbf{X}^{\star} \circ \mathbf{X}^{\star\prime})^{-1} + \mathsf{E}_3 = \mathsf{I} - (1/n\Lambda)\mathbf{D} \circ \mathbf{X}^{\star\prime} + \mathsf{F},$$

where $\|\mathsf{F}\|_2 \leq \|\mathsf{E}_3\|_2 + \|\mathbf{D} \circ \mathbf{X}^{\star\prime} \circ \mathsf{E}_4\|_2/(n\Lambda) \leq \|\mathsf{E}_3\|_2 + \tau(6/5)^{1/2}\|\mathsf{E}_4\|_2 \leq 20\tau^2 + 2\tau\Delta/\Lambda$. The result follows by noting that $\|\mathsf{F}\|_{\ell_2} \leq q\|\mathsf{F}\|_2$, $\tau \leq \omega/\sqrt{\Lambda}$ and that $\mathbf{X}^\star \circ \hat{\mathbf{X}}^+ = \mathbf{X}^\star \circ (\mathbf{X}^\star + \mathbf{D})^+ \circ \mathsf{Q}$. $\qquad\square$

Appendix H: Proof of Proposition 5

Proposition 21 *Given an operator* $\mathsf{E} : \mathbb{R}^{q \times q} \to \mathbb{R}^{q \times q}$, *there exists matrices* E_L, E_R *such that* $\mathcal{P}_{\mathcal{W}}(\mathsf{E}) = I \otimes E_L + E_R \otimes I$, *and* $\|E_L\|_{\ell_2}, \|E_R\|_{\ell_2} \leq \|\mathsf{E}\|_{\ell_2}/\sqrt{q}$.

Proof (Proposition 21) Define the subspaces $\mathcal{W}_R := \{S \otimes I : S \in \mathbb{R}^{q \times q}\}$ and $\mathcal{W}_L := \{I \otimes S : S \in \mathbb{R}^{q \times q}\}$. Note that $\mathcal{W}_R \cap \mathcal{W}_L = \mathrm{Span}(\mathsf{I})$, and hence $\mathcal{P}_{\mathcal{W}} = \mathcal{P}_{\mathcal{W}_R \cap \mathrm{Span}(\mathsf{I})^\perp} + \mathcal{P}_{\mathcal{W}_L \cap \mathrm{Span}(\mathsf{I})^\perp} + \mathcal{P}_{\mathrm{Span}(\mathsf{I})}$.

Define E_L and E_R to be matrices such that $E_R \otimes I = \mathcal{P}_{\mathcal{W}_R \cap \mathrm{Span}(\mathsf{I})^\perp}(\mathsf{E}) + (1/2)\mathcal{P}_{\mathrm{Span}(\mathsf{I})}(\mathsf{E})$, and $I \otimes E_L = \mathcal{P}_{\mathcal{W}_L \cap \mathrm{Span}(\mathsf{I})^\perp}(\mathsf{E}) + (1/2)\mathcal{P}_{\mathrm{Span}(\mathsf{I})}(\mathsf{E})$. For $i \in \{L, R\}$, we have the following. Since $\mathcal{P}_{\mathcal{W}_i \cap \mathrm{Span}(\mathsf{I})^\perp}$ and $(1/2)\mathcal{P}_{\mathrm{Span}(\mathsf{I})}$ are projectors onto orthogonal subspaces with spectral norm 1 and 1/2, respectively, we have $\|E_i \otimes I\|_{\ell_2} \leq \|\mathsf{E}\|_{\ell_2}$. Moreover, since $\|E_i \otimes I\|_{\ell_2} = \|E_i\|_{\ell_2}\|I\|_{\ell_2}$, we have $\|E_i\|_{\ell_2} \leq \|\mathsf{E}\|_{\ell_2}/\sqrt{q}$. $\square$

Proof (Proposition 5) By applying Proposition 21 to the operator D, we have $\mathcal{P}_{\mathcal{W}}(\mathsf{D}) = I \otimes E_L + E_R \otimes I$ for a pair of matrices $E_L, E_R \in \mathbb{R}^{q \times q}$ satisfying $\|E_L\|_{\ell_2}, \|E_R\|_{\ell_2} \leq \|\mathsf{D}\|_{\ell_2}/\sqrt{q}$. Moreover, since $\|E_L\|_2, \|E_R\|_2 < 1$, it follows that the matrices $I + E_R$ and $I + E_L$ are invertible. Consider the following identity

$$\mathsf{I} + \mathsf{D} = \Big(\mathsf{I} + \big(\mathcal{P}_{\mathcal{W}^\perp}(\mathsf{D}) - E_R \otimes E_L\big) \circ (I + E_R)^{-1}$$
$$\otimes (I + E_L)^{-1}\Big) \circ (I + E_R) \otimes (I + E_L).$$

We define $\mathsf{H} = (\mathcal{P}_{\mathcal{W}^\perp}(\mathsf{D}) - E_R \otimes E_L) \circ (I + E_R)^{-1} \otimes (I + E_L)^{-1} - \mathcal{P}_{\mathcal{W}^\perp}(\mathsf{D})$, and we define $W = (I + E_R) \otimes (I + E_L)$. By the triangle inequality, we have $\|W - \mathsf{I}\|_2 \leq 3\|\mathsf{D}\|_{\ell_2}/\sqrt{q}$.

Next we note that $\|(I + E_i)^{-1}\|_2 \leq 10/9$, $i \in \{L, R\}$ and that $\|(I + E_R)^{-1} \otimes (I + E_L)^{-1}\|_2 \leq 100/81$. We also have $\|E_R \otimes E_L\|_{\ell_2} = \|E_R\|_{\ell_2}\|E_L\|_{\ell_2} \leq \|\mathsf{D}\|_{\ell_2}^2/q$. By noting that $\|(I + E_i)^{-1} - I\|_2 \leq (10/9)\|E_i\|_2$, $i \in \{L, R\}$, we have $\|(I + E_R)^{-1} \otimes (I + E_L)^{-1} - I \otimes I\|_2 \leq 3\|\mathsf{D}\|_{\ell_2}/\sqrt{q}$. By combining these bounds, we obtain $\|\mathsf{H}\|_{\ell_2} \leq \|\mathcal{P}_{\mathcal{W}^\perp}(\mathsf{D})\|_{\ell_2}\|(I + E_R)^{-1} \otimes (I + E_L)^{-1} - I \otimes I\|_2 + \|E_R \otimes E_L\|_{\ell_2}\|(I + E_R)^{-1} \otimes (I + E_L)^{-1}\|_2 \leq 5\|\mathsf{D}\|_{\ell_2}^2/\sqrt{q}$. $\qquad\square$

Appendix I: Proof of Proposition 6

Proof (Proposition 6) To simplify notation in the proof, we denote $\alpha_8 := \alpha_8(q, \mathcal{L}^\star) = 96\sqrt{q}\|\mathcal{L}^\star\|_2$. We show that

$$\|\mathcal{L}^{(t)} - \mathcal{L}^{(t+1)}\|_2 \leq \alpha_9 \xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)}), \tag{34}$$

for some function $\alpha_9 := \alpha_9(q, r, \mathcal{L}^\star)$ that we specify later. In the proof of Theorem 3, we showed that $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)}) \le \gamma^t \xi_{\mathcal{L}^\star}(\mathcal{L}^{(0)})$ for some $\gamma < 1$. Hence, establishing (34) immediately implies that the sequence $\{\mathcal{L}^{(t)}\}_{t=1}^\infty$ is Cauchy.

Our proof builds on the proof of Theorem 3. Let

$$\mathcal{L}^{(t)} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t)}) \circ M$$

where $\mathsf{E}^{(t)}$ is a linear map that satisfies $\|\mathsf{E}^{(t)}\|_{\ell_2} < 1/\alpha_0$. In the proof of Theorem 3, we show that

$$\mathcal{L}^{(t+1)} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t+1)}) \circ W \circ M \circ N,$$

where $\|\mathsf{E}^{(t+1)}\|_{\ell_2} \le \|\mathsf{E}^{(t)}\|_{\ell_2}$, W is a rank-preserver and N is a positive-definite rank-preserver. Moreover, as a consequence of applying Proposition 5 to establish (23) in the proof, we obtain the bound $\|W - \mathsf{I}\|_2 \le 3\alpha_7 \|\mathsf{E}^{(t)}\|_{\ell_2}$. We use these bounds and relations to prove (34).

By the triangle inequality, we have

$$\|\mathcal{L}^{(t)} - \mathcal{L}^{(t+1)}\|_2 \le \|\mathcal{L}^\star \circ \mathsf{E}^{(t)} \circ M\|_2 + \|\mathcal{L}^\star \circ \mathsf{E}^{(t+1)} \circ W \circ M \circ N\|_2$$
$$+ \|\mathcal{L}^\star \circ M \circ (N - \mathsf{I})\|_2 + \|\mathcal{L}^\star \circ (W - \mathsf{I}) \circ M \circ N\|_2. \tag{35}$$

By Proposition 2 applied to the pairs of linear maps $\mathcal{L}^{(t)}, \mathcal{L}^\star$ and $\mathcal{L}^{(t+1)}, \mathcal{L}^\star$, we have $\|M - Q_1\|_2, \|W \circ M \circ N - Q_2\|_2 \le \alpha_8 \|\mathsf{E}^{(t)}\|_{\ell_2}$, for some pair of orthogonal rank-preservers Q_1, Q_2. Since $\alpha_8/\alpha_0 \le 1$, we have $\|M\|_2 \le 2$ and $\|W \circ M \circ N\|_2 \le 2$. Consequently, $\|\mathcal{L}^\star \circ \mathsf{E}^{(t)} \circ M\|_2, \|\mathcal{L}^\star \circ \mathsf{E}^{(t+1)} \circ W \circ M \circ N\|_2 \le 2\|\mathcal{L}^\star\|_2 \|\mathsf{E}^{(t)}\|_{\ell_2}$.

Next we bound $\|N - \mathsf{I}\|_2$. By utilizing $\|W \circ M \circ N - Q_2\|_2 \le \alpha_8/\alpha_0$, $\|M - Q_1\|_2 \le \alpha_8 \|\mathsf{E}^{(t)}\|_{\ell_2}$, and $\|W - \mathsf{I}\|_2 \le 3\alpha_7 \|\mathsf{E}^{(t)}\|_{\ell_2}$, one can show that $\|N - Q_3\|_2 \le (6\alpha_7 + 2\alpha_8 + 2)\|\mathsf{E}^{(t)}\|_{\ell_2}$, where $Q_3 = Q_1' \circ Q_2$ is an orthogonal rank-preserver. Since N is self-adjoint, we have $\|N^2 - \mathsf{I}\|_2 \le 3(6\alpha_7 + 2\alpha_8 + 2)\|\mathsf{E}^{(t)}\|_{\ell_2}$, and hence $\|N - \mathsf{I}\|_2 \le 3(6\alpha_7 + 2\alpha_8 + 2)\|\mathsf{E}^{(t)}\|_{\ell_2}$. This also implies the bound $\|N\|_2 \le 3$.

We apply these bounds to obtain $\|\mathcal{L}^\star \circ M \circ (N - \mathsf{I})\|_2 \le 6(6\alpha_7 + 2\alpha_8 + 2)\|\mathcal{L}^\star\|_2 \|\mathsf{E}^{(t)}\|_{\ell_2}$, and $\|\mathcal{L}^\star \circ (W - \mathsf{I}) \circ M \circ N\|_2 \le 9\alpha_7 \|\mathcal{L}^\star\|_2 \|\mathsf{E}^{(t)}\|_2$.

We define $\alpha_9 := (4 + 6(6\alpha_7 + 2\alpha_8 + 2) + 9\alpha_7)\|\mathcal{L}^\star\|_2$ (these are exactly the sum of the coefficients of $\|\mathsf{E}^{(t)}\|_{\ell_2}$ in the above bounds). The result follows by adding these bounds and subsequently taking the infimum over $\mathsf{E}^{(t)}$ in (35). $\qquad\square$

References

1. Agarwal, A., Anandkumar, A., Jain, P., Netrapalli, P.: Learning Sparsely Used Overcomplete Dictionaries via Alternating Minimization. SIAM Journal on Optimization **26**(4), 2775–2799 (2016). https://doi.org/10.1137/140979861

2. Agarwal, A., Anandkumar, A., Netrapalli, P.: A Clustering Approach to Learning Sparsely Used Overcomplete Dictionaries. IEEE Transactions on Information Theory **63**(1), 575–592 (2017). https://doi.org/10.1109/TIT.2016.2614684

3. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. IEEE Transactions on Signal Processing **54**(11), 4311–4322 (2006). https://doi.org/10.1109/TSP.2006.881199

4. Arora, S., Ge, R., Ma, T., Moitra, A.: Simple, Efficient, and Neural Algorithms for Sparse Coding. In: Conference on Learning Theory (2015)

5. Arora, S., Ge, R., Moitra, A.: New Algorithms for Learning Incoherent and Overcomplete Dictionaries. Journal of Machine Learning Research: Workshop and Conference Proceedings **35**, 1–28 (2014)

6. Barak, B., Kelner, J.A., Steurer, D.: Dictionary Learning and Tensor Decomposition via the Sum-of-Squares Method. In: Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing. ACM (2015). https://doi.org/10.1145/2746539.2746605

7. Barron, A.R.: Universal Approximation Bounds for Superpositions of a Sigmoidal Function. IEEE Transactions on Information Theory **39**(3), 930–945 (1993). https://doi.org/10.1109/18.256500

8. Bhaskar, B.N., Tang, G., Recht, B.: Atomic Norm Denoising with Applications to Line Spectral Estimation. IEEE Transactions on Signal Processing **61**(23), 5987–5999 (2013)

9. Blumensath, T., Davies, M.E.: Iterative Hard Thresholding for Compressed Sensing. Applied and Computational Harmonic Analysis **27**, 265–274 (2009). https://doi.org/10.1016/j.acha.2009.04.002

10. Bruckstein, A.M., Donoho, D.L., Elad, M.: From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images. SIAM Review **51**(1), 34–81 (2009). https://doi.org/10.1137/060657704

11. Candès, E.J., Plan, Y.: Tight Oracle Inequalities for Low-Rank Matrix Recovery From a Minimal Number of Noisy Random Measurements. IEEE Transactions on Information Theory **57**(4), 2342–2359. https://doi.org/10.1109/TIT.2011.2111771

12. Candès, E.J., Recht, B.: Exact Matrix Completion via Convex Optimization. Foundations of Computational Mathematics **9**(6), 717–772 (2009). https://doi.org/10.1007/s10208-009-9045-5

13. Candès, E.J., Romberg, J., Tao, T.: Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. IEEE Transactions on Information Theory **52**(2), 489–509 (2006). https://doi.org/10.1109/TIT.2005.862083

14. Candès, E.J., Tao, T.: Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies? IEEE Transactions on Information Theory **52**(12), 5406–5425 (2006). https://doi.org/10.1109/TIT.2006.885507

15. Chandrasekaran, V., Parillo, P., Willsky, A.S.: Latent Variable Graphical Model Selection via Convex Optimization. The Annals of Statistics **40**(4), 1935–1967 (2012). https://doi.org/10.1214/11-AOS949

16. Chandrasekaran, V., Recht, B., Parrilo, P.A., Willsky, A.S.: The Convex Geometry of Linear Inverse Problems. Foundations of Computational Mathematics **12**(6), 805–849 (2012). https://doi.org/10.1007/s10208-012-9135-7

17. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic Decomposition by Basis Pursuit. SIAM Journal on Scientific Computing **20**(1), 33–61 (1998). https://doi.org/10.1137/S1064827596304010

18. Cho, E.: Inner Products of Random Vectors on S^n. Journal of Pure and Applied Mathematics: Advances and Applications **9**(1), 63–68 (2013)

19. Cuturi, M.: Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. In: Advances in Neural Information Processing Systems (2013)

20. Davidson, K.R., Szarek, S.J.: Local Operator Theory, Random Matrices and Banach Spaces. In: W.B. Johnson, J. Lindenstrauss (eds.) Handbook of the Geometry of Banach Spaces, chap. 8, pp. 317–366. Elsevier B. V. (2011)

21. DeVore, R.A., Temlyakov, V.N.: Some Remarks on Greedy Algorithms. Advances in Computational Mathematics **5**(1), 173–187 (1996). https://doi.org/10.1007/BF02124742

22. Donoho, D.L.: Compressed Sensing. IEEE Transactions on Information Theory **52**(4), 1289–1306 (2006). https://doi.org/10.1109/TIT.2006.871582

23. Donoho, D.L.: For Most Large Underdetermined Systems of Linear Equations the Minimal ℓ_1-norm Solution Is Also the Sparsest Solution. Communications on Pure and Applied Mathematics **59**(6), 797–829 (2006). https://doi.org/10.1002/cpa.20132

24. Donoho, D.L., Huo, X.: Uncertainty Principles and Ideal Atomic Decomposition. IEEE Transactions on Information Theory **47**(7), 2845–2862

25. Elad, M.: Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer (2010). https://doi.org/10.1007/978-1-4419-7011-4

26. Fazel, M.: Matrix Rank Minimization with Applications. Ph.D. thesis, Department of Electrical Engineering, Stanford University (2002)

27. Fazel, M., Candès, E., Recht, B., Parrilo, P.: Compressed Sensing and Robust Recovery of Low Rank Matrices. In: 42nd IEEE Asilomar Conference on Signals, Systems and Computers (2008)

28. Garg, A., Gurvits, L., Oliveira, R., Wigderson, A.: A Deterministic Polynomial Time Algorithm for Non-Commutative Rational Identity Testing with Applications. In: IEEE 57th Annual Symposium on Foundations of Computer Science (2016). https://doi.org/10.1109/FOCS.2016.95

29. Ge, R., Lee, J.D., Ma, T.: Matrix Completion has No Spurious Local Minimum. In: Advances in Neural Information Processing Systems (2016)
30. Goldfarb, D., Ma, S.: Convergence of Fixed-Point Continuation Algorithms for Matrix Rank Minimization. Foundations of Computational Mathematics **11**, 183–210 (2011). https://doi.org/10.1007/s10208-011-9084-6
31. Gorman, W.M.: Estimating Trends in Leontief Matrices. Unpublished note, referenced in Bacharach (1970) (1963)
32. Gouveia, J., Parrilo, P.A., Thomas, R.R.: Lifts of Convex Sets and Cone Factorizations. Mathematics of Operations Research **38**(2), 248–264 (2013). https://doi.org/10.1287/moor.1120.0575
33. Gribonval, R., Jenatton, R., Bach, F., Kleinsteuber, M., Seibert, M.: Sample Complexity of Dictionary Learning and Other Matrix Factorizations. IEEE Transactions on Information Theory **61**(6), 3469–3486 (2015). https://doi.org/10.1109/TIT.2015.2424238
34. Gurvits, L.: Classical Complexity and Quantum Entanglement. Journal of Computer and Systems Sciences **69**(3), 448–484 (2004). https://doi.org/10.1016/j.jcss.2004.06.003
35. Idel, M.: A Review of Matrix Scaling and Sinkhorn's Normal Form for Matrices and Positive Maps. CoRR arXiv:1609.06349 (2016)
36. Jain, P., Meka, R., Dhillon, I.S.: Guaranteed Rank Minimization via Singular Value Projection. In: Advances in Neural Information Processing Systems (2009)
37. Jones, L.K.: A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training. The Annals of Statistics **20**(1), 608–613 (1992). https://doi.org/10.1214/aos/1176348546
38. Kato, T.: Perturbation Theory for Linear Operators. Springer-Verlag (1966)
39. Khachiyan, L., Kalantari, B.: Diagonal Matrix Scaling and Linear Programming. SIAM Journal on Optimization **2**(4), 668–672 (1991). https://doi.org/10.1137/0802034
40. Linial, N., Samorodnitsky, A., Wigderson, A.: A Deterministic Strongly Polynomial Algorithm for Matrix Scaling and Approximate Permanents. Combinatorica **20**(4), 545–568 (2000). https://doi.org/10.1007/s004930070007
41. Mairal, J., Bach, F., Ponce, J.: Sparse Modeling for Image and Vision Processing. Foundations and Trends in Computer Graphics and Vision **8**(2–3), 85–283 (2014). https://doi.org/10.1561/0600000058
42. Marcus, M., Moyls, B.N.: Transformations on Tensor Product Spaces. Pacific Journal of Mathematics **9**(4), 1215–1221 (1959)
43. Meinhausen, N., Bühlmann, P.: High-Dimensional Graphs and Variable Selection with the Lasso. The Annals of Statistics **34**(3), 1436–1462 (2006). https://doi.org/10.1214/009053606000000281
44. Natarajan, B.K.: Sparse Approximate Solutions to Linear Systems. SIAM Journal on Computing **24**(2), 227–234 (1993). https://doi.org/10.1137/S0097539792240406
45. Nesterov, Y., Nemirovskii, A.: Interior-Point Polynomial Algorithms in Convex Programming. SIAM Studies in Applied and Numerical Mathematics (1994). https://doi.org/10.1137/1.9781611970791
46. Olshausen, B.A., Field, D.J.: Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. Nature **381**, 607–609 (1996). https://doi.org/10.1038/381607a0
47. Oymak, S., Hassibi, B.: Sharp MSE Bounds for Proximal Denoising. Foundations of Computational Mathematics **16**(4), 965–1029 (2016). https://doi.org/10.1007/s10208-015-9278-4
48. Parikh, N., Boyd, S.: Proximal Algorithms. Foundations and Trends in Optimization **1**(3), 127–239 (2014). https://doi.org/10.1561/2400000003
49. Pisier, G.: Remarques sur un résultat non publié de B. Maurey. Séminaire Analyse fonctionnelle (dit "Maurey-Schwartz") pp. 1–12 (1981)
50. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. SIAM Review **52**(3), 471–501 (2010). https://doi.org/10.1137/070697835
51. Renegar, J.: A Mathematical View of Interior-Point Methods in Convex Optimization. MOS-SIAM Series on Optimization (2001). https://doi.org/10.1137/1.9780898718812
52. Schnass, K.: On the Identifiability of Overcomplete Dictionaries via the Minimisation Principle Underlying K-SVD. Applied and Computational Harmonic Analysis **37**(3), 464–491 (2014). https://doi.org/10.1016/j.acha.2014.01.005
53. Schnass, K.: Convergence Radius and Sample Complexity of ITKM Algorithms for Dictionary Learning. Applied and Computational Harmonic Analysis (2016). https://doi.org/10.1016/j.acha.2016.08.002

54. Shah, P., Bhaskar, B.N., Tang, G., Recht, B.: Linear System Identification via Atomic Norm Regularization. In: 51st IEEE Conference on Decisions and Control (2012)
55. Sinkhorn, R.: A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices. The Annals of Mathematical Statistics **35**(2), 876–879 (1964). https://doi.org/10.1214/aoms/1177703591
56. Spielman, D.A., Wang, H., Wright, J.: Exact Recovery of Sparsely-Used Dictionaries. Journal on Machine Learning and Research: Workshop and Conference Proceedings **23**(37), 1–18 (2012)
57. Stewart, G., Sun, J.: Matrix Perturbation Theory. Academic Press (1990)
58. Sun, J., Qu, Q., Wright, J.: A Geometric Analysis of Phase Retrieval. Foundations of Computational Mathematics (2017). https://doi.org/10.1007/s10208-017-9365-9
59. Sun, J., Qu, Q., Wright, J.: Complete Dictionary Recovery over the Sphere I: Overview and the Geometric Picture. IEEE Transactions on Information Theory **63**(2), 853–884 (2017). https://doi.org/10.1109/TIT.2016.2632162
60. Sun, J., Qu, Q., Wright, J.: Complete Dictionary Recovery over the Sphere II: Recovery by Riemannian Trust-region Method. IEEE Transactions on Information Theory **63**(2), 885–914 (2017). https://doi.org/10.1109/TIT.2016.2632149
61. Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, Series B **58**, 267–288 (1994)
62. Toh, K.C., Todd, M.J., Tütüncü, R.H.: SDPT3 – a MATLAB Software Package for Semidefinite Programming. Optimization Methods and Software **11**, 545–581 (1999). https://doi.org/10.1080/10556789908805762
63. Tropp, J.A.: User-Friendly Tail Bounds for Sums of Random Matrices. Foundations of Computational Mathematics **12**(4), 389–434 (2012). https://doi.org/10.1007/s10208-011-9099-z
64. Tunçel, L.: Potential Reduction and Primal-Dual Methods. In: H. Wolkowicz, R. Saigal, L. Vandenberghe (eds.) Handbook of Semidefinite Programming – Theory, Algorithms, and Applications, chap. 9. Kluwer's International Series in Operations Research and Management Science (2000). https://doi.org/10.1007/978-1-4615-4381-7
65. Vainsencher, D., Mannor, S., Bruckstein, A.M.: The sample complexity of dictionary learning. Journal of Machine Learning Research **12** (2011)
66. Yannakakis, M.: Expressing Combinatorial Optimization Problems by Linear Programs. Journal of Computer and System Sciences **43**, 441–466 (1991). https://doi.org/10.1016/0022-0000(91)90024-Y