Journal of Machine Learning Research 21 (2020) 1-68 Submitted 9/19; Revised 8/20; Published 8/20

Complete Dictionary Learning via ℓ^4-Norm Maximization over the Orthogonal Group

Yuexiang Zhai[†] YSZ@BERKELEY.EDU
Zitong Yang[†] ZITONG@BERKELEY.EDU
Zhenyu Liao[‡] LIAOZHENYU2004@GMAIL.COM
John Wright[◇] JW2966@COLUMBIA.EDU
Yi Ma[†] YIMA@EECS.BERKELEY.EDU

[†] *Department of Electrical Engineering and Computer Science*
University of California, Berkeley, CA 94720-1770
[‡] *Kuaishou Technology.*
[◇] *Department of Electrical Engineering*
Columbia University, New York, NY, 10027

Editor: Julien Mairal

Abstract

This paper considers the fundamental problem of learning a complete (orthogonal) dictionary from samples of sparsely generated signals. Most existing methods solve the dictionary (and sparse representations) based on heuristic algorithms, usually without theoretical guarantees for either optimality or complexity. The recent ℓ^1-minimization based methods do provide such guarantees but the associated algorithms recover the dictionary one column at a time. In this work, we propose a new formulation that *maximizes* the ℓ^4-norm over the orthogonal group, to learn the entire dictionary. We prove that under a random data model, with nearly minimum sample complexity, the global optima of the ℓ^4-norm are very close to signed permutations of the ground truth. Inspired by this observation, we give a conceptually simple and yet effective algorithm based on "*matching, stretching, and projection*" (MSP). The algorithm provably converges locally and cost per iteration is merely an SVD. In addition to strong theoretical guarantees, experiments show that the new algorithm is significantly more efficient and effective than existing methods, including KSVD and ℓ^1-based methods. Preliminary experimental results on mixed real imagery data clearly demonstrate advantages of so learned dictionary over classic PCA bases.

Keywords: sparse dictionary learning, ℓ^4-norm maximization, orthogonal group, measure concentration, fixed point algorithm

1. Introduction and Overview

1.1. Motivation

One of the most fundamental problems in signal processing or data analysis is that given an observed signal $\boldsymbol{y}$, either continuous or discrete, we would like to find a transform $\mathcal{F}$ such that after applying the transform, the resulting signal $\boldsymbol{x} = \mathcal{F}(\boldsymbol{y})$ becomes much more compact, sparse, or compressible. We believe such a compact representation $\boldsymbol{x}$ can help

reveal intrinsic structures of the observed signal $\boldsymbol{y}$ and is also more amenable to storage, processing, and transmission.

For computational purposes, the transforms considered are typically orthogonal linear transforms so that both $\mathcal{F}$ and $\mathcal{F}^{-1}$ are easy to represent and compute: In this case, we have $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}$ or $\boldsymbol{x} = \boldsymbol{D}^*\boldsymbol{y}$ for some orthogonal matrix (or linear operator) $\boldsymbol{D}$. Examples include the classical Fourier transform (Oppenheim, 1999; Vetterli et al., 2014) or various wavelets (Vetterli and Kovacevic, 1995). Conventionally, the best transform to use is typically by "*design*": By assuming the signals of interest have certain physical or mathematical properties (e.g. band-limited, piece-wise smooth, or scale-invariant), one may derive or design the optimal transforms associated with different classes of functions or signals. This classical approach has found its deep mathematical roots in functional analysis (Kreyszig, 1978) and harmonic analysis (Stanton and Weinstein, 1981; Katznelson, 2004) and has seen great empirical successes in digital signal processing (Oppenheim, 1999).

Nevertheless, in the modern big data era, both science and engineering are inundated with tremendous high-dimensional data, such as images, audios, languages, and genetics etc. Many of such data may or may not belong to the classes of functions or signals for which we know the optimal transforms. Assumptions about their intrinsic (low-dim) structures are not clear enough for us to derive any new transform either. Therefore, we are compelled to change our practice and ask whether we can "*learn*" an optimal transform (if exists) directly from the observed data. That is, if we observe many, say p, samples $\boldsymbol{y}_i \in \mathbb{R}^n$ from a model:

$$\boldsymbol{y}_i = \boldsymbol{D}_o\boldsymbol{x}_i, \quad i = 1, \ldots, p$$

where $\boldsymbol{x}_i \in \mathbb{R}^m$ is presumably much more compact or sparse, can we both learn $\boldsymbol{D}_o$ and recover the associated $\boldsymbol{x}_i$ from such $\boldsymbol{y}_i$? Here $\boldsymbol{D}_o$ is called the ground truth "dictionary" and the problem is known as "Dictionary Learning." Dictionary learning is a fundamental problem in data science since finding a sparse representation of data appears in different applications such as computational neural science (Olshausen and Field, 1996, 1997), machine learning (Argyriou et al., 2008; Ranzato et al., 2007), and computer vision (Elad and Aharon, 2006; Yang et al., 2010; Mairal et al., 2014). Note that here we know neither the dictionary $\boldsymbol{D}_o$ nor the hidden state $\boldsymbol{x}$. So in machine learning, dictionary learning belongs to the category of *unsupervised learning* problems, and a fundamental one that is.

In this work, we consider the problem of *learning a complete dictionary*[1] from sparsely generated sample signals. More precisely, an n-dimensional sample $\boldsymbol{y} \in \mathbb{R}^n$ is assumed to be a sparse superposition of columns of a non-singular complete dictionary $\boldsymbol{D}_o \in \mathbb{R}^{n\times n}$: $\boldsymbol{y} = \boldsymbol{D}_o\boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^n$ is a sparse (coefficient) vector. A typical statistical model for the sparse coefficient is that entries of $\boldsymbol{x}$ are i.i.d. Bernoulli-Gaussian $\{x_i\} \sim_{iid} \mathrm{BG}(\theta)$[2] (Spielman et al., 2012; Sun et al., 2015; Bai et al., 2018).

Suppose we are given a collection of sample signals $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p] \in \mathbb{R}^{n\times p}$, each of which is generated as $\boldsymbol{y}_i = \boldsymbol{D}_o\boldsymbol{x}_i$ for a nonsingular matrix $\boldsymbol{D}_o$. Write $\boldsymbol{X}_o = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p] \in \mathbb{R}^{n\times p}$. In this notation, we have:

$$\boldsymbol{Y} = \boldsymbol{D}_o\boldsymbol{X}_o. \tag{1}$$

1. In this work, we only consider the case the dictionary is complete. The more general setting in which the dictionary $\boldsymbol{D}_o$ is over-complete, $m > n$, is beyond the scope of this paper.
2. I.e., each entry x_i is a product of independent Bernoulli and standard normal random variables: $x_i = \Omega_i V_i$, where $\Omega_i \sim_{iid} \mathrm{Ber}(\theta)$ and $V_i \sim_{iid} \mathcal{N}(0,1)$.

Dictionary learning is the problem of recovering both the dictionary $\boldsymbol{D}_o$ and the sparse coefficients $\boldsymbol{X}_o$, given only the samples $\boldsymbol{Y}$. Equivalently, we wish to factorize $\boldsymbol{Y}$ as $\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X}$, where $\boldsymbol{D}$ is an estimate of the true dictionary $\boldsymbol{D}_o$ and $\boldsymbol{X}$ is the sparsest possible.

Under the Bernoulli-Gaussian assumption, the problem of learning an arbitrary complete dictionary can be reduced to that of learning an *orthogonal* dictionary: As shown in Sun et al. (2015), when the objective is smooth, the problem can be converted to the orthogonal case through a preconditioning:

$$\boldsymbol{Y} \leftarrow \left(\frac{1}{p\theta}\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-\frac{1}{2}} \boldsymbol{Y}.$$

So without loss of generality, we can assume that $\boldsymbol{D}_o$ is an orthogonal matrix: $\boldsymbol{D}_o \in \mathsf{O}(n;\mathbb{R})$.

Because $\boldsymbol{Y}$ is sparsely generated, the optimal estimate $\boldsymbol{D}_\star$ should make the associated coefficients $\boldsymbol{X}_\star$ maximally sparse. In other words, ℓ^0-norm, the number of non-zero entries, of $\boldsymbol{X}_\star$ should be as small as possible, therefore, one may formulate the following optimization program to find $\boldsymbol{X}_\star$:

$$\min_{\boldsymbol{X},\boldsymbol{D}} \|\boldsymbol{X}\|_0, \quad \text{subject to} \quad \boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X},\ \boldsymbol{D} \in \mathsf{O}(n;\mathbb{R}). \tag{2}$$

Under fairly mild conditions, the global minimizer of the ℓ^0-norm recovers the true dictionary $\boldsymbol{D}_o$ Spielman et al. (2012). However, global minimization of the ℓ^0-norm is a challenging NP-hard problem (Donoho, 2006; Candes and Tao, 2005; Natarajan, 1995). Traditionally, one resorts to local heuristics such as orthogonal matching pursuit, as in the KSVD algorithm (Aharon et al., 2006; Rubinstein et al., 2010).[3] This approach has been widely practiced but does not give any strong guarantees for optimality of the algorithm nor correctness of the solution, as we will see through experiments compared with our method.

Since ℓ^1-norm minimization promotes sparsity (Candès, 2014) and the ℓ^1-norm is convex and continuous, one may reformulate dictionary learning as an ℓ^1-minimization problem:

$$\min_{\boldsymbol{X},\boldsymbol{D}} \|\boldsymbol{X}\|_1, \quad \text{subject to} \quad \boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X},\ \boldsymbol{D} \in \mathsf{O}(n;\mathbb{R}). \tag{3}$$

This reformulation (3) remains a nonsmooth optimization with nonconvex constraints, which is in general still NP-hard (Murty and Kabadi, 1987). Nevertheless, many heuristic algorithms (Mairal et al., 2008, 2009, 2012) have attempted to reformulate optimization (3) as an unconstrained optimization problem with ℓ^1-regularization by removing the constraint $\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X}$.

Although the ℓ^0- or ℓ^1-minimization has been widely practiced in dictionary learning, rigorous justification for optimality and correctness is only provided recently. Geng and Wright (2014) is the first to show the local optimality of the ℓ^1-minimization. Spielman et al. (2012) further proves that a complete (square and invertible) $\boldsymbol{D}$ can be recovered from $\boldsymbol{Y}$, when each column of $\boldsymbol{X}$ contains no more than $O(\sqrt{n})$ of nonzero entries. Subsequent works (Agarwal et al., 2013, 2014; Arora et al., 2014, 2015) have provided provable algorithms for overcomplete dictionary learning, under the assumption that each column of X has $\tilde{O}(\sqrt{n})$ nonzero entries.[4]

3. See also (Ravishankar and Bresler, 2015) algorithms for learning orthogonal sparsifying transformations.
4. $\tilde{O}$ suppresses logarithm factors.

In this work, we mainly focus on *complete dictionary learning*, which implies $\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X})$. Spielman et al. (2012) has proposed to find the sparsest vector $\boldsymbol{d}^*\boldsymbol{Y}$ in $\mathrm{row}(\boldsymbol{Y})$ one by one via solving the following optimization:

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} \|\boldsymbol{d}^*\boldsymbol{Y}\|_1, \quad \text{subject to} \quad \boldsymbol{d} \neq \boldsymbol{0} \tag{4}$$

n times, instead of solving the hard nonconvex optimization (3) directly. (4) is easier to solve, since it can be further reduced to a linear programming. Sun et al. (2015) has proposed a new formulation with spherical constraint that finds one column of the dictionary via solving:

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} \|\boldsymbol{d}^*\boldsymbol{Y}\|_1, \quad \text{subject to} \quad \|\boldsymbol{d}\|_2 = 1. \tag{5}$$

For escaping saddle points, Sun et al. (2015) has proposed a provably correct second-order Riemannian Trust Region method (Absil et al., 2009) to solve (5) and improved the sparsity level of $\boldsymbol{X}$ to constant.[5] However, as addressed by Gilboa et al. (2018), the computational complexity of a second-order algorithm is high, not to mention one needs to solve (5) n times! To mitigate the computation complexity, Gilboa et al. (2018) suggests that the first-order gradient descent method with random initialization has the same performance as a second-order one, and Bai et al. (2018) shows that a randomly initialized first-order projected subgradient descent is able to solve (5). But these approaches fail to overcome the main cause for the high complexity – one needs to break down the complete dictionary learning problem (3) into solving n optimization programs like (4) (or (5)).

Besides the ℓ^1-minimization based dictionary learning framework, works based on the sum-of-square (SoS) SDP hierarchy (Barak et al., 2015; Ma et al., 2016; Schramm and Steurer, 2017) also provide guarantee for exact recovery of (overcomplete) dictionary learning problem in polynomial time, under some specific statistical assumption of the data model. But one still needs to solve the SoS based SDP Problem n times to recover a dictionary with n components,[6] let alone the high computational complexity for solving a high-dimensional SDP programming each time (Qu et al., 2014; Bai et al., 2018).

1.2. Our Approach and Connection to Prior Works

In this paper, we show that one can actually efficiently learn a complete (orthogonal) dictionary holistically via solving one ℓ^4-norm optimization over the entire orthogonal group $\mathsf{O}(n;\mathbb{R})$:

$$\max_{\boldsymbol{D}} \|\boldsymbol{D}^*\boldsymbol{Y}\|_4^4, \quad \text{subject to} \quad \boldsymbol{D} \in \mathsf{O}(n;\mathbb{R}), \tag{6}$$

where the ℓ^4-norm of a matrix means the sum of 4^{th} powers of all entries: $\forall \boldsymbol{A} \in \mathbb{R}^{n\times m}$, $\|\boldsymbol{A}\|_4^4 = \sum_{i,j} a_{i,j}^4$. The intuition for (6) comes from:

$$\max_{\boldsymbol{X},\boldsymbol{D}} \|\boldsymbol{X}\|_4^4, \quad \text{subject to} \quad \boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X}, \boldsymbol{D} \in \mathsf{O}(n;\mathbb{R}), \tag{7}$$

where maximizing the ℓ^4-norm of $\boldsymbol{X}$ (over a sphere) promotes "spikiness" or "sparsity" of $\boldsymbol{X}$ (Zhang et al., 2018) . It is easy to see this as the sparsest points on a unit ℓ^2-sphere

5. Each column of $\boldsymbol{X}$ can contain $O(n)$ non zero entries.
6. See Theorem 1.1 in Schramm and Steurer (2017)

(points $(0,1)$, $(0,-1)$, $(1,0)$, and $(-1,0)$) have the smallest ℓ^1-norm and largest ℓ^4-norm, as shown in Figure 1. Since the columns of orthogonal matrices have unit norm, the constraint $\boldsymbol{D} \in \mathsf{O}(n;\mathbb{R})$ can be viewed as simultaneously enforcing orthonormal constraints on n vectors on the unit ℓ^2-sphere $\mathbb{S}^1$. Moreover, comparing to the ℓ^1-norm, the ℓ^4-norm objective is everywhere smooth, so we expect it is amenable to better optimization.

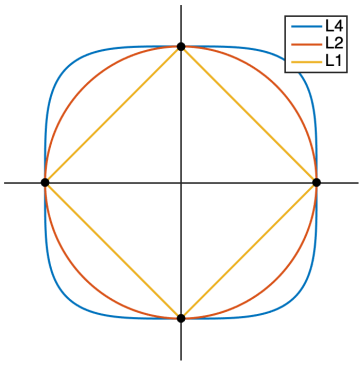


Figure 1: Unit ℓ^1-,ℓ^2-, and ℓ^4- spheres in $\mathbb{R}^2$, a similar picture can be found in Figure 1 of Li and Bresler (2018).

1.2.1. Spherical Harmonic Analysis

The property of the ℓ^4-norm has long been realized and used in seeking (orthonormal) functions with similar properties since 1970's, if not any earlier. For instance for spherical harmonics, the Stanton-Weinstein Theorem (Stanton and Weinstein, 1981; Lu, 1987) have shown that "among all the ℓ^2-normalized spherical harmonics of a given degree, the ℓ^4-norm is locally maximized by the 'highest-weight' function," which among all the eigenfunctions of the Laplacian on the sphere, is the "most concentrated" in measure (see Theorem 1 of Stanton and Weinstein (1981)). In quantum mechanics, such functions represent trajectories that are the most probable and most closely approximate classical trajectories.

1.2.2. Independent Component Analysis

We should note that 4^{th}-order statistical cumulants have been widely used in blind source separation or independent component analysis (ICA) since the 1990's, see Hyvärinen (1997); Hyvärinen and Oja (1997) and references therein. So if $\boldsymbol{x}$ are n independent components, by finding extrema of the so-called *kurtosis*: $\mathrm{kurt}(\boldsymbol{d}^*\boldsymbol{y}) \doteq \mathbb{E}[(\boldsymbol{d}^*\boldsymbol{y})^4] - 3\mathbb{E}[(\boldsymbol{d}^*\boldsymbol{y})^2]^2$, one can identify one independent (non-Gaussian) component x_i at a time. Algorithm wise, this is similar to using the ℓ^1-minimization (5) to identify one column $\boldsymbol{d}_i$ at a time for $\boldsymbol{D}$. Fast fixed-point like algorithms have been developed for this purpose (Hyvärinen, 1997; Hyvärinen and Oja, 1997). If $\boldsymbol{x}$ are indeed i.i.d. Bernoulli-Gaussian, with $\|\boldsymbol{d}\|_2^2 = 1$, the second term in $\mathrm{kurt}(\boldsymbol{d}^*\boldsymbol{y})$ would become a constant and the objective of ICA coincides with maximizing the sparsity-promoting ℓ^4-norm of a vector over a sphere.

1.2.3. Sum of Squares

The use of ℓ^4-norm can also be justified from the perspective of sum of squares (SoS). The works of Barak et al. (2015); Ma et al. (2016); Schramm and Steurer (2017) show that in theory, when $\boldsymbol{x}$ is sufficiently sparse, one can utilize properties of higher order sum of squares polynomials (such as the fourth order polynomials) to correctly recover $\boldsymbol{D}$. Although Schramm and Steurer (2017) has improved proposed faster tensor decomposition based on SoS method, again the algorithm only recovers one column $\boldsymbol{d}_i$ at a time.

1.2.4. Blind Deconvolution

Recent works in blind deconvolution (Zhang et al., 2018; Li and Bresler, 2018) have also explored the sparsity promoting property of the ℓ^4-norm in their objective function. Zhang et al. (2018); Li and Bresler (2018) have shown that, for any filter on a unit sphere, all local maxima of the ℓ^4-objective are close to the inverse of the ground truth filter (up to the intrinsic sign and shift ambiguity). Moreover, the global geometry of the ℓ^4-norm over the sphere is good – all saddle points have negative curvatures. Such nice global geometry guarantees a randomly initialized first-order Riemannian gradient descent algorithm to escape saddle points and find the ground truth, for both single channel (Zhang et al., 2018) and multi-channel (Li and Bresler, 2018; Qu et al., 2019) tasks.

1.3. Main Results

We shall first note that there is an intrinsic "signed permutation" ambiguity in all dictionary learning formulation (2), (3), and (7). For any input data matrix $\boldsymbol{Y}$, suppose $\boldsymbol{D}_\star \in \mathsf{O}(n;\mathbb{R})$ is the optimal orthogonal dictionary that gives the sparsest coefficient matrix $\boldsymbol{X}_\star$ satisfying $\boldsymbol{Y} = \boldsymbol{D}_\star \boldsymbol{X}_\star$. Then for any matrix $\boldsymbol{P}$ in the signed permutation group $\mathsf{SP}(n)$, the group of orthogonal matrices that only contain $0, \pm 1$, we have:

$$\boldsymbol{Y} = \boldsymbol{D}_\star \boldsymbol{X}_\star = \boldsymbol{D}_\star \boldsymbol{P}\boldsymbol{P}^* \boldsymbol{X}_\star,$$

where $\boldsymbol{P}^*\boldsymbol{X}_\star$ is equally sparse as $\boldsymbol{X}_\star$ and $\boldsymbol{D}_\star\boldsymbol{P} \in \mathsf{O}(n;\mathbb{R})$. So we can only expect to recover the correct dictionary (and sparse coefficient matrix) *up to an arbitrary signed permutation.* Therefore, we say the ground truth dictionary $\boldsymbol{D}_o$ is successfully recovered, if any *signed permuted* version $\boldsymbol{D}_o\boldsymbol{P}$ is found.[7] Unlike approaches (Spielman et al., 2012; Sun et al., 2015; Bai et al., 2018) that solve one column at a time for $\boldsymbol{D}$, we here attempt recover the entire dictionary $\boldsymbol{D}$ from solving the problem (6). The signed permutation ambiguity would create numerous equivalent global maximizers, which poses a serious challenge to analysis and optimization.

In this paper, we adopt the Bernoulli-Gaussian model as in prior works (Spielman et al., 2012; Sun et al., 2015; Bai et al., 2018). We assume our observation matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$ is produced by the product of a ground truth orthogonal dictionary $\boldsymbol{D}_o$, and a Bernoulli-Gaussian matrix $\boldsymbol{X}_o \in \mathbb{R}^{n\times p}$:

$$\boldsymbol{Y} = \boldsymbol{D}_o\boldsymbol{X}_o, \quad \boldsymbol{D}_o \in \mathsf{O}(n;\mathbb{R}),\ \{\boldsymbol{X}_o\}_{i,j} \sim_{iid} \mathrm{BG}(\theta). \tag{8}$$

7. In mathematical terms, we are looking for a solution in the quotient space between the orthogonal group and the signed permutation group: $\mathsf{O}(n;\mathbb{R})/\mathsf{SP}(n)$.

The Bernoulli-Gaussian model can be considered as a prototype for dictionary learning because one may adjust θ to control the sparsity level of the ground truth $\boldsymbol{X}_o$. With the Bernoulli-Gaussian assumption, we now state our main result.

1.3.1. Correctness of the Proposed Objective Function

Theorem 1 (Correctness of Global Optima, informal version of Theorem 7) *$\forall \theta \in (0,1)$, let $\boldsymbol{X}_o \in \mathbb{R}^{n\times p}$, $x_{i,j} \sim_{iid} BG(\theta)$, $\boldsymbol{D}_o \in \mathsf{O}(n;\mathbb{R})$ an arbitrary orthogonal matrix, and $\boldsymbol{Y} = \boldsymbol{D}_o\boldsymbol{X}_o$. Suppose $\hat{\boldsymbol{A}}_\star$ is a global maximizer of the optimization problem:*

$$\max_{\boldsymbol{A}} \|\boldsymbol{A}\boldsymbol{Y}\|_4^4, \quad \text{subject to} \quad \boldsymbol{A} \in \mathsf{O}(n;\mathbb{R}), \tag{9}$$

then for any $\varepsilon \in [0,1]$, there exists a signed permutation matrix $\boldsymbol{P} \in \mathsf{SP}(n)$, such that

$$\frac{1}{n}\left\|\hat{\boldsymbol{A}}_\star^* - \boldsymbol{D}_o\boldsymbol{P}\right\|_F^2 \le C\varepsilon, \tag{10}$$

holds with high probability, when p is large enough, and C is a constant depends on θ.

Theorem 1 is obtained through the following line of reasoning: 1) $\forall \boldsymbol{A} \in \mathsf{O}(n;\mathbb{R})$, we can view $\frac{1}{p}\|\boldsymbol{A}\boldsymbol{Y}\|_4^4 = \frac{1}{p}\|\boldsymbol{A}\boldsymbol{D}_o\boldsymbol{X}_o\|_4^4 = \frac{1}{p}\sum_{j=1}^p \|\boldsymbol{A}\boldsymbol{D}_o\boldsymbol{x}_j\|_4^4$ as the mean of p i.i.d. random variables, so it will concentrate to its expectation $\mathbb{E}_{\boldsymbol{x}_j}\|\boldsymbol{A}\boldsymbol{D}_o\boldsymbol{x}_j\|_4^4$; 2) the value of $\mathbb{E}_{\boldsymbol{x}_j}\|\boldsymbol{A}\boldsymbol{D}_o\boldsymbol{x}_j\|_4^4$ is largely characterized by $\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4$, a deterministic function (with respect to $\boldsymbol{A}$) on the orthogonal group, whose global optima $\boldsymbol{A}_\star$ satisfy $\boldsymbol{A}_\star = \boldsymbol{P}^*\boldsymbol{D}_o^*, \forall \boldsymbol{P} \in \mathsf{SP}(n)$. Therefore, when p is large enough, maximizing $\|\boldsymbol{A}\boldsymbol{Y}\|_4^4$ over the orthogonal group is equivalent to (with high probability) maximizing the deterministic objective $\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4$ over the orthogonal group, which yields the desire result. Formal statements and proofs are given in Section 2 and the Appendices.

1.3.2. A Fast Optimization Algorithm

Unlike almost all previous algorithms that find the correct dictionary one column $\boldsymbol{d}_i$ at a time, in Section 3 we introduce a novel *matching, stretching, and projection* (MSP) algorithm that solves the program (9) directly for the entire $\boldsymbol{D} \in \mathsf{O}(n;\mathbb{R})$. The MSP algorithm directly computes (transpose of) the optimal dictionary $\boldsymbol{A}_\star$ as the "fixed point" to the following iteration:

$$\boldsymbol{A}_{t+1} = \mathcal{P}_{\mathsf{O}(n;\mathbb{R})}\left[(\boldsymbol{A}_t\boldsymbol{Y})^{\circ 3}\boldsymbol{Y}^*\right], \tag{11}$$

where $\mathcal{P}_{\mathsf{O}(n;\mathbb{R})}(\cdot)$ is the projection onto the orthogonal group $\mathsf{O}(n;\mathbb{R})$, which can be easily calculated from SVD (see Lemma 9). Meanwhile, the MSP algorithm also efficiently maximizes the deterministic objective $\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4$ over $\mathsf{O}(n;\mathbb{R})$ by the following iteration:

$$\boldsymbol{A}_{t+1} = \mathcal{P}_{\mathsf{O}(n;\mathbb{R})}\left[(\boldsymbol{A}_t\boldsymbol{D}_o)^{\circ 3}\boldsymbol{D}_o^*\right]. \tag{12}$$

Statistical analysis for proving Theorem 1 suggests that estimates from random samples converge to their expectation. For the deterministic objective $\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4$, we further show that the proposed algorithm converges fast with a *cubic rate* around each global maximizer.

Essentially, the update of the MSP algorithm (11), (12) are performing projected gradient ascend with *infinite* step size with respect to objective function $\|\boldsymbol{A}\boldsymbol{Y}\|_4^4$, $\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4$, over

$\mathsf{O}(n;\mathbb{R})$ respectively. The MSP algorithm is similar to the FastICA algorithm (Hyvärinen and Oja, 1997) for independent component analysis (ICA) and such similarity is extensively discussed in a follow-up work (Zhai et al., 2019).

Theorem 2 (Cubic Convergence Rate, informal version of Theorem 16) *Given an orthogonal matrix $\boldsymbol{A} \in \mathsf{O}(n;\mathbb{R})$, if $\|\boldsymbol{A}-\boldsymbol{I}\|_F^2 = \varepsilon$ for a small $\varepsilon < 0.579$, and let $\boldsymbol{A}'$ denote the result after one iteration of our proposed MSP algorithm* (12)*, then we have $\|\boldsymbol{A}'-\boldsymbol{I}\|_F^2 \le O(\varepsilon^3)$.*

In Theorem 2, showing the local convergence of to identity matrix a general orthogonal matrix $\boldsymbol{A}$ to $\boldsymbol{I}$ suffices to characterize the local convergence of $\boldsymbol{A}\boldsymbol{D}_o$ to a signed permutation matrix $\boldsymbol{P}$, because of the signed permutation symmetry in the ℓ^4 norm and the orthogonal invariant of SVD. More details are provided in Section 2 and Section 3.

Although Theorem 1 characterizes the properties of the randomized objective $\|\boldsymbol{A}\boldsymbol{Y}\|_4^4$ while Theorem 2 describes a deterministic result, they are highly related with each other. We use iteration (11) to maximize the randomized objective $\|\boldsymbol{A}\boldsymbol{Y}\|_4^4$ of Theorem 1 and Theorem 2 characterizes the local convergence of iteration (12). $(\boldsymbol{A}\boldsymbol{Y})^{\circ 3}\boldsymbol{Y}^*$ in (11) concentrates to its expectation when there is enough samples and the expectation $\mathbb{E}[(\boldsymbol{A}\boldsymbol{Y})^{\circ 3}\boldsymbol{Y}^*]$ highly depends on $(\boldsymbol{A}\boldsymbol{D}_o)^{\circ 3}\boldsymbol{D}_o$ of (12). Such algorithmic relationship between (11) and (12) is discussed in Proposition 10 and Proposition 11 of Section 3.

As our algorithm is very efficient and scalable, we can test it over very large range of dimensions and settings. Extensive simulations suggest that the MSP algorithm converges globally to the correct solution under broad conditions. We give a global convergence proof for the case $n = 2$ (on $\mathsf{O}(2;\mathbb{R})$) and conjecture that similar results hold for arbitrary n (under mild conditions). Extensive experiments show that the algorithm is far more efficient than existing heuristic algorithms and (Riemannian) gradient or subgradient based algorithms. With this efficient algorithm, we characterize empirically the range of success for the program (9), which goes well beyond any existing theoretical guarantees (Sun et al., 2015; Bai et al., 2018) for the complete dictionary case.

1.3.3. Observations and Implications

Notice that at first sight, the optimization problem associated with dictionary learning is highly nonconvex, with nontrivial orthogonal constraints, and with numerous critical points and ambiguities. Hence, understandably, many recent approaches focus on introducing regularization to the objective function so as to relax the constraints (Aharon et al., 2006; Mairal et al., 2008, 2009, 2012) or analyzing and utilizing local information and design heuristic or gradient descent schemes for such objective functions (with or without regularization) (Agarwal et al., 2014; Arora et al., 2015; Ge et al., 2015; Ma et al., 2017; Li and Liang, 2018; Du et al., 2018; Allen-Zhu et al., 2018; Davis et al., 2018).

However, in our work, we have observed a surprising phenomenon that is rather contrary to conventional views: The discrete signed permutation symmetry $\mathsf{SP}(n)$ associated with the orthogonal group $\mathsf{O}(n;\mathbb{R})$ are beneficial. In our work, they both play an important role in making the algorithm efficient and effective, instead of being nuisances or difficulties to be dealt with. As we will see, such discrete symmetry of the orthogonal group makes the global landscape of the objective function amenable to global convergence and enables a fixed-point

type algorithm that converges at a super-linear rate to a correct solution in the quotient space $\mathsf{O}(n;\mathbb{R})/\mathsf{SP}(n)$. Similar phenomena that symmetry facilitates global convergence of non-convex programs have been also been observed and reported in recent works (Ge et al., 2015; Sun et al., 2015; Zhang et al., 2018; Li and Bresler, 2018; Chi et al., 2018; Kuo et al., 2019). This is a direction that deserves better and deeper study in the future which encourages significant confluence of geometry, algebra, and statistics.

1.4. Notations

We use a bold uppercase and a bold lowercase letter to denote a matrix and a vector, respectively: $\boldsymbol{X} \in \mathbb{R}^{n\times p}, \boldsymbol{x} \in \mathbb{R}^n$. Moreover, for a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, we use $\boldsymbol{x}_j \in \mathbb{R}^n, \forall j \in [p]$ to denote its j^{th} column vector as default. We use $\boldsymbol{X}^*$ or $\boldsymbol{x}^*$ to denoted the (conjugate) transpose of a matrix or a vector, respectively. We reserve lower-case letter for scalar: $x \in \mathbb{R}$. We use $\|\boldsymbol{X}\|_4$ to denote the element-wise ℓ^4-norm of a matrix $\boldsymbol{X}$ ($\|\boldsymbol{X}\|_4^4 = \sum_{i,j} x_{i,j}^4$). We use $\boldsymbol{D}_o$ to denote the ground truth orthogonal dictionary, and $\boldsymbol{A}$ is an estimate of $\boldsymbol{D}_o^*$ from solving (9). $\theta \in (0,1)$ to is sparsity level of the ground truth Bernoulli-Gaussian sparse coefficient $\boldsymbol{X}_o$: $\boldsymbol{X}_o \sim \mathrm{BG}(\theta)$. We use $\circ$ to denote the Hadamard product: $\forall \boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times m}$, $\{\boldsymbol{A} \circ \boldsymbol{B}\}_{i,j} = a_{i,j} b_{i,j}$, and $\{\boldsymbol{A}^{\circ r}\}_{i,j} = a_{i,j}^r$ is the element-wise r^{th} power of $\boldsymbol{A}$.

Given an input data matrix $\boldsymbol{Y}$ randomly generated from $\boldsymbol{Y} = \boldsymbol{D}_o \boldsymbol{X}_o$, $\boldsymbol{X}_o \sim_{iid} \mathrm{BG}(\theta)$, for any orthogonal matrix $\boldsymbol{A} \in \mathsf{O}(n;\mathbb{R})$, we define $\hat{f} : \mathsf{O}(n;\mathbb{R}) \times \mathbb{R}^{n\times p} \mapsto \mathbb{R}$ as the 4^{th} power of ℓ^4-norm of $\boldsymbol{A}\boldsymbol{Y}$:

$$\hat{f}(\boldsymbol{A}, \boldsymbol{Y}) \doteq \|\boldsymbol{A}\boldsymbol{Y}\|_4^4 . \tag{13}$$

We define $f : \mathsf{O}(n;\mathbb{R}) \mapsto \mathbb{R}$ as the expectation of $\hat{f}$ over $\boldsymbol{X}_o$:

$$f(\boldsymbol{A}) \doteq \mathbb{E}_{\boldsymbol{X}_o}[\hat{f}(\boldsymbol{A}, \boldsymbol{Y})] = \mathbb{E}_{\boldsymbol{X}_o}\big[\, \|\boldsymbol{A}\boldsymbol{Y}\|_4^4 \,\big]. \tag{14}$$

For any orthogonal matrix $\boldsymbol{W} \in \mathsf{O}(n;\mathbb{R})$, we define $g : \mathsf{O}(n;\mathbb{R}) \mapsto \mathbb{R}$ as 4^{th} power of its ℓ^4-norm:

$$g(\boldsymbol{W}) \doteq \|\boldsymbol{W}\|_4^4 . \tag{15}$$

1.5. Organization of this Paper

Rest of the paper is organized as follows. In Section 2, we characterize the global maximizers of (9) statistically via measure concentration. In Section 3, we describe the proposed MSP algorithm. We characterize fixed points of the algorithm and show its convergence results in Section 4, All related proofs can be found in the appendices. Finally, in Section 5, we conduct extensive experiments to show effectiveness and efficiency of our method, by comparing with the state of the art.

2. Key Analysis and Main Result

2.1. Expectation and Concentration of the ℓ^4-Objective

In this section, we statistically justify that one can recover the ground truth dictionary $\boldsymbol{D}_o$ by solving

$$\max_{\boldsymbol{A}} \hat{f}(\boldsymbol{A}, \boldsymbol{Y}) = \|\boldsymbol{A}\boldsymbol{Y}\|_4^4, \quad \text{subject to} \quad \boldsymbol{A} \in \mathsf{O}(n;\mathbb{R}). \tag{16}$$

Notice that the solution to the above ℓ^4-norm optimization problem:

$$\hat{\boldsymbol{A}}_\star = \operatorname*{arg\,max}_{\boldsymbol{A}\in\mathsf{O}(n;\mathbb{R})} \hat{f}(\boldsymbol{A},\boldsymbol{Y})$$

is a random variable that depends on the random samples $\boldsymbol{Y}$. We need to characterize how "close" an estimate $\hat{\boldsymbol{A}}_\star$ is to the ground truth $\boldsymbol{D}_o$. A key technique is to show that the random objective function actually concentrates on its expectation (a deterministic function) as the number of observations p increases. We first calculate $f(\boldsymbol{A})$, the expectation of $\hat{f}(\boldsymbol{A},\boldsymbol{Y})$ and provide its concentration bound in Lemma 3 and Lemma 4 respectively.

Lemma 3 (Properties of $f(\boldsymbol{A})$) *$\forall\theta\in(0,1)$, let $\boldsymbol{X}_o\in\mathbb{R}^{n\times p}$, $x_{i,j}\sim_{iid} BG(\theta)$, $\boldsymbol{D}_o\in\mathsf{O}(n;\mathbb{R})$ is an orthogonal matrix, and $\boldsymbol{Y}=\boldsymbol{D}_o\boldsymbol{X}_o$. Then, $\forall\boldsymbol{A}\in\mathsf{O}(n;\mathbb{R})$, we have*

$$\frac{1}{3p\theta} f(\boldsymbol{A}) = (1-\theta)g(\boldsymbol{A}\boldsymbol{D}_o) + \theta n. \tag{17}$$

Proof See A.1 ■

Lemma 4 (Concentration Bound of ℓ^4-Norm) *$\forall\theta\in(0,1)$, if $\boldsymbol{X}\in\mathbb{R}^{n\times p}$, $x_{i,j}\sim_{iid} BG(\theta)$, $\forall\delta>0$, the following inequality holds:*

$$\mathbb{P}\left(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})} \frac{1}{np}\left|\|\boldsymbol{W}\boldsymbol{X}\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}\boldsymbol{X}\|_4^4\right| \geq \delta\right) < \frac{1}{p}, \tag{18}$$

when $p=\Omega(\theta n^2\ln n/\delta^2)$.

Proof See A.2. ■

Lemma 4 implies that, for any orthogonal transformation $\boldsymbol{W}\boldsymbol{X},\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})$ of a Bernoulli-Gaussian matrix $\boldsymbol{X}\in\mathbb{R}^{n\times p}$, $\frac{1}{np}\|\boldsymbol{W}\boldsymbol{X}\|_4^4$ concentrates onto its expectation as long as the sample size p is large enough – in the order $\Omega(\theta n^2\ln n/\delta^2)$. By our definition, $\hat{f}(\boldsymbol{A},\boldsymbol{Y}) = \|\boldsymbol{A}\boldsymbol{Y}\|_4^4 = \|\boldsymbol{W}\boldsymbol{X}_o\|_4^4$ ($\boldsymbol{W}=\boldsymbol{A}\boldsymbol{D}_o$ is an orthogonal matrix) satisfies the concentration inequality in Lemma 4. Therefore, including designing optimization algorithms, $f(\boldsymbol{A})$ can be considered as a good proxy to the original objective $\hat{f}(\boldsymbol{A},\boldsymbol{Y})$ and we can consider (16) as maximizing its expectation:

$$\max_{\boldsymbol{A}} f(\boldsymbol{A}) = \mathbb{E}\,\|\boldsymbol{A}\boldsymbol{Y}\|_4^4 \quad \text{subject to}\, \boldsymbol{A}\in\mathsf{O}(n;\mathbb{R}). \tag{19}$$

The function $f(\boldsymbol{A})$ is a deterministic function and its geometric property would help us understand the landscape of the original objective. Moreover, Lemma 3 states that the global maximizers of the mean $f(\boldsymbol{A})$ are exactly the global maximizers of $g(\boldsymbol{A}\boldsymbol{D}_o)=\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4$. To understand how such a function can be effectively optimized, we need to study the extrema of ℓ^4-norm $g(\cdot)$ over the orthogonal group $\mathsf{O}(n;\mathbb{R})$.

2.2. Property of the ℓ^4-Norm over $\mathsf{O}(n;\mathbb{R})$

Lemma 5 (Extrema of ℓ^4-Norm over Orthogonal Group) *For any orthogonal matrix $\boldsymbol{A} \in \mathsf{O}(n;\mathbb{R})$, $g(\boldsymbol{A}) = \|\boldsymbol{A}\|_4^4 \in [1, n]$ and $g(\boldsymbol{A})$ reaches maximum if and only if $\boldsymbol{A} \in \mathsf{SP}(n)$.*

Proof See A.3. ■

This lemma implies that if $\boldsymbol{A}_\star$ is a global maximizer of $g(\boldsymbol{A}\boldsymbol{D}_o)$, i.e. $\|\boldsymbol{A}_\star \boldsymbol{D}_o\|_4^4 = n$, then it differs from $\boldsymbol{D}_o^*$ by a signed permutation. However, this lemma does not say the function may or may not have other local minima or maxima. Nevertheless, our experiments in Section 5.2 will show that even if such critical points exist, they are unlikely to be stable (or attractive). As a direct corollary to Lemma 3 and Lemma 5, we know that $\forall \boldsymbol{A} \in \mathsf{O}(n;\mathbb{R}), \theta \in (0,1)$, the maximum value of $f(\boldsymbol{A})$ satisfies:

$$\frac{1}{3p\theta} f(\boldsymbol{A}) \le n, \tag{20}$$

and the equality holds if and only if $\boldsymbol{A}\boldsymbol{D}_o \in \mathsf{SP}(n)$. Although we know that in the stochastic setting, (16) concentrates to the following ℓ^4-norm maximization over $\mathsf{O}(n;\mathbb{R})$:

$$\max_{\boldsymbol{A}} g(\boldsymbol{A}\boldsymbol{D}_o) = \|\boldsymbol{A}\boldsymbol{D}_o\|_4^4, \quad \text{subject to} \quad \boldsymbol{A} \in \mathsf{O}(n;\mathbb{R}), \tag{21}$$

we cannot hope that the estimate from $\hat{\boldsymbol{A}}_\star = \arg\max_{\boldsymbol{A}\in\mathsf{O}(n;\mathbb{R})} \hat{f}(\boldsymbol{A}, \boldsymbol{Y})$ would achieve the maximal value of $g(\boldsymbol{A}\boldsymbol{D}_o)$ precisely. But if the value is close to the maximal, how close would $\hat{\boldsymbol{A}}_\star$ be to a global maximizer? The following lemma shows that when the ℓ^4-norm of an orthogonal matrix $\boldsymbol{A}$ is close to the maximum value n, it is also close to a signed permutation matrix in Frobenius norm.

Lemma 6 (Approximate Maxima of ℓ^4-Norm over Orthogonal Group) *Suppose $\boldsymbol{W}$ is an orthogonal matrix: $\boldsymbol{W} \in \mathsf{O}(n;\mathbb{R})$. $\forall \varepsilon \in [0,1]$, if $\frac{1}{n}\|\boldsymbol{W}\|_4^4 \ge 1 - \varepsilon$, then $\exists \boldsymbol{P} \in \mathsf{SP}(n)$, such that*

$$\frac{1}{n}\|\boldsymbol{W} - \boldsymbol{P}\|_F^2 \le 2\varepsilon. \tag{22}$$

Proof See A.4. ■

This result is useful whenever we evaluate how close a solution given by an algorithm is to the optimal solution, in terms of value of the objective function. With all the above results, we are now ready to characterize in what sense a global maximizer of $\hat{\boldsymbol{A}}_\star = \arg\max_{\boldsymbol{A}\in\mathsf{O}(n;\mathbb{R})} \hat{f}(\boldsymbol{A}, \boldsymbol{Y})$ gives a "correct" estimate of the ground truth dictionary $\boldsymbol{D}_o$.

2.3. Main Statistical Result

Theorem 7 (Correctness of the Global Optima) *$\forall \theta \in (0,1)$, let $\boldsymbol{X}_o \in \mathbb{R}^{n\times p}$, $x_{i,j} \sim_{iid} BG(\theta)$, $\boldsymbol{D}_o \in \mathsf{O}(n;\mathbb{R})$ is any orthogonal matrix, and $\boldsymbol{Y} = \boldsymbol{D}_o\boldsymbol{X}_o$. Suppose $\hat{\boldsymbol{A}}_\star$ is a global maximizer of the optimization problem:*

$$\max_{\boldsymbol{A}} \hat{f}(\boldsymbol{A}, \boldsymbol{Y}) = \|\boldsymbol{A}\boldsymbol{Y}\|_4^4, \quad \text{subject to} \quad \boldsymbol{A} \in \mathsf{O}(n;\mathbb{R}),$$

then for any $\varepsilon \in [0,1]$, there exists a signed permutation matrix $\boldsymbol{P} \in \mathsf{SP}(n)$, such that

$$\frac{1}{n}\left\|\hat{\boldsymbol{A}}_\star^* - \boldsymbol{D}_o\boldsymbol{P}\right\|_F^2 \le C\varepsilon, \tag{23}$$

with probability at least $1 - \frac{1}{p}$*, when* $p = \Omega(\theta n^2 \ln n/\varepsilon^2)$*, for a constant* $C > \frac{4}{3\theta(1-\theta)}$.

Proof See A.5. ■

Theorem 7 states that with high probability, the global optimal solution to the ℓ^4-norm optimization (16) is close to the true solution (up to a signed permutation) as the sample size p is of the order $\Omega(\theta n^2 \ln n/\varepsilon^2)$, where ε is the desired accuracy. This result quantifies the sample size needed to achieve certain accuracy of recovery and it matches the intuition that more observations lead to better recovering result. Moreover, this result corresponds to our phase transition curves in Figure 8 of Section 5, and to the best of our knowledge, a sample complexity of $\Omega(n^2 \ln n)$ is currently the best result for dictionary learning task.

Remark 8 (Maximizing ℓ^{2k}-Norm) *If one were to choose maximizing ℓ^{2k}-norm to promoting sparsity, similar analysis of concentration bounds would reveal that for the same error bound, it requires much larger number p of samples for the (random) objective function $\hat{f}(\boldsymbol{A}, \boldsymbol{Y})$ to concentrate on its (deterministic) expectation $f(\boldsymbol{A})$. We will discuss the choice of k in more details in Section 4.4. Experiments in Figure 7 also corroborate with the findings.*

3. Algorithm: Matching, Stretching, and Projection (MSP)

In this section, we introduce an algorithm, based on a simple iterative *matching, stretching, and projection* (MSP) process, which efficiently solves the two related programs (16) and (21).

3.1. Algorithmic Challenges and Related Optimization Methods

Although (16) is everywhere smooth, the associated optimization is non-trivial in several ways. First, one needs to deal with the signed permutation ambiguity. The problem has exponentially many global maximizers. Furthermore, we are maximizing a convex function (or minimizing a concave function) over a constraint set. So conventional methods such as augmented Lagrangian (Bertsekas, 1997) barely works. This is because the Lagrangian:

$$\mathcal{L}(\boldsymbol{A}, \boldsymbol{\Lambda}) \doteq -\|\boldsymbol{A}\boldsymbol{Y}\|_4^4 + \langle \boldsymbol{A}\boldsymbol{A} - \boldsymbol{I}, \boldsymbol{\Lambda} \rangle$$

will go to negative infinity due to the concavity of the objective function $-\|\boldsymbol{A}\boldsymbol{Y}\|_4^4$. Notice that all of its global maximizers are on the constraint set, an ideal iterative algorithm should converge to a solution that *exactly lies on constraint set* $\mathsf{O}(n;\mathbb{R})$.

Another natural way to optimize (16) (or (21)) is to apply Riemannian gradient (or projected gradient) type methods (Edelman et al., 1998; Absil et al., 2009) on $\mathsf{O}(n;\mathbb{R})$. One can take small gradient steps to ensure convergence and such methods converge at best with a linear rate (say, if the objective function is strongly convex). Nevertheless, by better utilizing the special global geometry of the parameter space (the orthogonal group), we can choose an arbitrary large step size and the process converges much more rapidly (with a superlinear rate).

We next introduce a very effective and efficient algorithm to solve problems (16) and (21), that is based on simple iterative matching, stretching and projection (MSP) operators.

Mathematically, our algorithm for solving (16) or (21) are essentially doing projected gradient ascent on objective $\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4$ or $\|\boldsymbol{A}\boldsymbol{Y}\|_4^4$ respectively, but with *infinite* step size. Such infinite step size gradient method is able to find global maximum of our proposed objective (16) and (21) because it exploits the coincidence between the signed permutation symmetry in our formulation and the symmetry of $\mathsf{SP}(n)$ over $\mathsf{O}(n;\mathbb{R})$.

3.2. ℓ^4-Norm Maximization over the Orthogonal Group

Since the objective function of dictionary learning (16) concentrates to the ℓ^4-norm maximization problem (21) w.h.p., we first introduce our algorithm for the simpler (deterministic) case:

$$\max_{\boldsymbol{A}} g(\boldsymbol{A}\boldsymbol{D}_o) = \|\boldsymbol{A}\boldsymbol{D}_o\|_4^4, \quad \text{subject to} \quad \boldsymbol{A} \in \mathsf{O}(n;\mathbb{R}). \tag{24}$$

In this setting, we only have information of the values of $g(\boldsymbol{A}\boldsymbol{D}_o)$ and $\nabla_{\boldsymbol{A}} g(\boldsymbol{A}\boldsymbol{D}_o)$. We want to update $\boldsymbol{A}$ to recover $\boldsymbol{D}_o^*$ based on these information. We use the following lemma to enforce the orthogonal constraint.

Lemma 9 (Projection onto the Orthogonal Group) *$\forall \boldsymbol{A} \in \mathbb{R}^{n\times n}$, the orthogonal matrix which is closest to $\boldsymbol{A}$ in Frobenius norm is the following:*

$$\mathcal{P}_{\mathsf{O}(n;\mathbb{R})}(\boldsymbol{A}) = \operatorname*{arg\,min}_{\boldsymbol{M}\in\mathsf{O}(n;\mathbb{R})} \|\boldsymbol{A}-\boldsymbol{M}\|_F^2 = \boldsymbol{U}\boldsymbol{V}^*, \tag{25}$$

where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = SVD(\boldsymbol{A})$.

Proof See B.1. ■

The MSP algorithm that maximizing (21) is outlined as Algorithm 1.

Algorithm 1 MSP for ℓ^4-Maximization over $\mathsf{O}(n;\mathbb{R})$

1: **Initialize** $\boldsymbol{A}_0 \in \mathsf{O}(n,\mathbb{R})$ ▷ Initialize $\boldsymbol{A}_0$ for iteration
2: **for** $t = 0, 1, ..., T-1$ **do**
3: $\quad \partial\boldsymbol{A}_t \doteq 4(\boldsymbol{A}_t\boldsymbol{D}_o)^{\circ 3}\boldsymbol{D}_o^*$ ▷ $\nabla_{\boldsymbol{A}}\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4 = 4(\boldsymbol{A}\boldsymbol{D}_o)^{\circ 3}\boldsymbol{D}_o^*$
4: $\quad \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \mathrm{SVD}\big(\partial\boldsymbol{A}_t\big)$
5: $\quad \boldsymbol{A}_{t+1} = \boldsymbol{U}\boldsymbol{V}^*$ ▷ Project $\boldsymbol{A}_{t+1}$ onto orthogonal group
6: **end for**
7: **Output** $\boldsymbol{A}_T$

In each iteration of Algorithm 1, we use $\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4/n$ to evaluate how "close" $\boldsymbol{A}\boldsymbol{D}_o$ is to a signed permutation matrix, since Lemma 5 shows the global maximum of $\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4$ is n and the maximal evaluation is therefore normalized to 1. In Step 3 of the MSP algorithm, the calculation of $\partial\boldsymbol{A}_t = 4(\boldsymbol{A}_t\boldsymbol{D}_o)^{\circ 3}\boldsymbol{D}_o^*$ does not require knowledge of $\boldsymbol{D}_o$. It is merely the gradient of the objective function:

$$\nabla_{\boldsymbol{A}} g(\boldsymbol{A}\boldsymbol{D}_o) = \nabla_{\boldsymbol{A}}\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4 = 4(\boldsymbol{A}\boldsymbol{D}_o)^{\circ 3}\boldsymbol{D}_o^*.$$

As the name of the algorithm suggests, each iteration actually performs a *"matching, stretching, and projection"* operation: It first matches the current estimate $\boldsymbol{A}_t$ to the true

$\boldsymbol{D}_o$. Then the element-wise cubic function $(\cdot)^{\circ 3}$ stretches all entries of $\boldsymbol{A}_t\boldsymbol{D}_o$ by promoting the large ones and suppressing the small ones. $\partial\boldsymbol{A}_t$ is the correlation between so "sparsified" pattern and the original basis $\boldsymbol{D}_o^*$, which is then projected back onto the closest orthogonal matrix $\boldsymbol{A}_{t+1}$ in Frobenius distance.

Repeating this "matching, stretching, and projection" process, $\boldsymbol{A}_t\boldsymbol{D}_o$ is increasingly sparsified while ensuring the orthogonality of each $\boldsymbol{A}_t$. Ideally the process will stop when $\boldsymbol{A}_t\boldsymbol{D}_o$ becomes the sparsest, that is, a signed permutation matrix. The iterative MSP algorithm utilizes the global geometry of the orthogonal group and acts more like the *power iteration* method or the fixed point algorithm (Hyvärinen and Oja, 1997). It is easy to see that for any fixed $\boldsymbol{D}_o$, the optimal $\boldsymbol{A}_\star$ is the "fixed point" to the following equation:

$$\boldsymbol{A}_\star = \mathcal{P}_{\mathsf{O}(n;\mathbb{R})}\big[(\boldsymbol{A}_\star\boldsymbol{D}_o)^{\circ 3}\boldsymbol{D}_o^*\big], \tag{26}$$

where $\mathcal{P}_{\mathsf{O}(n;\mathbb{R})}$ is the projection onto the orthogonal group $\mathsf{O}(n;\mathbb{R})$. The proposed "matching, stretching, and projection" algorithm is precisely to compute the fixed point of this equation in the most natural way! Our analysis (Theorem 16) will show that this scheme converges extremely well and actually achieves a *super-linear* local convergence rate.

Example 1 (One Run of Algorithm 1) *To help better visualize how well the algorithm works, we consider a special case when $\boldsymbol{D}_o = \boldsymbol{I}$. The problem reduces to finding a matrix with the maximum ℓ^4-norm over the orthogonal group:*

$$\max_{\boldsymbol{A}} g(\boldsymbol{A}) \doteq \|\boldsymbol{A}\|_4^4, \quad \text{subject to} \quad \boldsymbol{A} \in \mathsf{O}(n;\mathbb{R}). \tag{27}$$

We randomly initialize the MSP algorithm with an orthogonal matrix $\boldsymbol{A}_0 \in \mathbb{R}^{3\times 3}$, and the sequences below show how the quickly the MSP algorithm quickly converges to a signed permutation matrix:

$$\begin{aligned}
&\boldsymbol{A}_0 = \begin{pmatrix} -0.8249 & 0.3820 & -0.4168 \\ -0.5240 & -0.2398 & 0.8173 \\ -0.2122 & -0.8925 & -0.3979 \end{pmatrix} \xrightarrow{\textit{stretching}} \boldsymbol{A}_0^{\circ 3} = \begin{pmatrix} -0.5613 & 0.0557 & -0.0724 \\ -0.1439 & -0.0138 & 0.5459 \\ -0.0096 & -0.7109 & -0.0630 \end{pmatrix} \\
\xrightarrow{\textit{projection}}\ &\boldsymbol{A}_1 = \begin{pmatrix} -0.9795 & 0.0621 & -0.1917 \\ -0.1953 & -0.0594 & 0.9789 \\ -0.0494 & -0.9963 & -0.0703 \end{pmatrix} \xrightarrow{\textit{stretching}} \boldsymbol{A}_1^{\circ 3} = \begin{pmatrix} -0.9397 & 0.0002 & -0.0070 \\ -0.0075 & -0.0002 & 0.9381 \\ -0.0001 & -0.9889 & -0.0003 \end{pmatrix} \\
\xrightarrow{\textit{projection}}\ &\boldsymbol{A}_2 = \begin{pmatrix} -1.0000 & 0.0002 & -0.0077 \\ -0.0077 & -0.0003 & 1.000 \\ -0.0002 & -1.0000 & -0.0003 \end{pmatrix} \xrightarrow{\textit{stretching}} \boldsymbol{A}_2^{\circ 3} = \begin{pmatrix} -0.9999 & 0.0000 & -0.0000 \\ -0.0000 & -0.0000 & 0.9999 \\ -0.0000 & -1.0000 & -0.0000 \end{pmatrix} \\
\xrightarrow{\textit{projection}}\ &\boldsymbol{A}_3 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} \xrightarrow{\textit{output}} \boldsymbol{A}_3^{\circ 3} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}.
\end{aligned}$$

3.3. ℓ^4-Norm Maximization for Dictionary Learning

Having understood how to optimize the deterministic case for the expectation, we now consider the original dictionary learning problem (16):

$$\max_{\boldsymbol{A}} \hat{f}(\boldsymbol{A},\boldsymbol{Y}) = \|\boldsymbol{A}\boldsymbol{Y}\|_4^4, \quad \text{subject to} \quad \boldsymbol{A} \in \mathsf{O}(n;\mathbb{R}).$$

This naturally leads to a similar MSP algorithm, outlined as Algorithm 2

Algorithm 2 MSP for ℓ^4-Maximization based Dictionary Learning

1: **Initialize** $\boldsymbol{A}_0 \in \mathsf{O}(n, \mathbb{R})$ ▷ Initialize $\boldsymbol{A}_0$ for iteration
2: **for** $t = 0, 1, ..., T-1$ **do**
3: $\quad \partial \boldsymbol{A}_t \doteq 4(\boldsymbol{A}_t \boldsymbol{Y})^{\circ 3} \boldsymbol{Y}^*$ ▷ $\nabla_{\boldsymbol{A}} \|\boldsymbol{A}\boldsymbol{Y}\|_4^4 = 4(\boldsymbol{A}\boldsymbol{Y})^{\circ 3}\boldsymbol{Y}^*$
4: $\quad \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \mathrm{SVD}\big(\partial \boldsymbol{A}_t\big)$
5: $\quad \boldsymbol{A}_{t+1} = \boldsymbol{U}\boldsymbol{V}^*$ ▷ Project $\boldsymbol{A}_{t+1}$ onto orthogonal group
6: **end for**
7: **Output** $\boldsymbol{A}_T, \|\boldsymbol{A}_T \boldsymbol{Y}\|_4^4 / 3np\theta$

Note that in the output we also normalize $\|\boldsymbol{A}\boldsymbol{Y}\|_4^4$ by dividing the maximum of its expectation: $3np\theta$ so that the maximal output value would be around 1. Similar to Algorithm 1, we also use $\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4/n$ to evaluate how "close" $\boldsymbol{A}\boldsymbol{D}_o$ is to a signed permutation matrix in each iteration.

The same intuition of "matching, stretching, and projection" for the deterministic case naturally carries over here. In Step 3, the estimate $\boldsymbol{A}_t$ is matched with the observation $\boldsymbol{Y}$. The cubic function $(\cdot)^{\circ 3}$ re-scales the results and promotes entry-wise spikiness of $\boldsymbol{X}_t = \boldsymbol{A}_t\boldsymbol{Y}$ accordingly. Again, here $\partial \boldsymbol{A}_t = 4(\boldsymbol{A}_t\boldsymbol{Y})^{\circ 3}\boldsymbol{Y}^*$ is the gradient $\nabla_{\boldsymbol{A}} \hat{f}(\boldsymbol{A}, \boldsymbol{Y})$ of the objective function. However, the algorithm is not performing gradient ascent: The matrix $(\boldsymbol{A}_t\boldsymbol{Y})^{\circ 3}\boldsymbol{Y}^*$ is actually the sample covariance of the following two random vectors: $(\boldsymbol{A}_t\boldsymbol{y})^{\circ 3}$ and $\boldsymbol{y}$. The subsequent projection of this sample covariance matrix onto the orthogonal group $\mathsf{O}(n;\mathbb{R})$ normalizes the scale of the operator $\boldsymbol{A}_{t+1}$, hence normalize the covariance of $\boldsymbol{A}_{t+1}\boldsymbol{Y}$ for the next iteration.

Similar to the deterministic case, for any given sparsely generated data matrix $\boldsymbol{Y}$, the optimal dictionary $\boldsymbol{A}_\star$ is the "fixed point" to the following equation:

$$\boldsymbol{A}_\star = \mathcal{P}_{\mathsf{O}(n;\mathbb{R})}\big[(\boldsymbol{A}_\star \boldsymbol{Y})^{\circ 3}\boldsymbol{Y}^*\big], \tag{28}$$

where $\mathcal{P}_{\mathsf{O}(n;\mathbb{R})}$ is the projection onto the orthogonal group $\mathsf{O}(n;\mathbb{R})$. The iterative "matching, stretching, and projection" scheme is precisely to compute the fixed point of this equation in the most natural way!

Although the data and the objective function are random here, Proposition 10 below clarifies the relationship between this expectation and the deterministic gradient $\nabla_{\boldsymbol{A}} g(\boldsymbol{A}\boldsymbol{D}_o)$ and Proposition 11 further shows that $\nabla_{\boldsymbol{A}} \hat{f}(\boldsymbol{A}, \boldsymbol{Y})$ concentrates to its expectation when p increases.

Proposition 10 (Expectation of $\nabla_{\boldsymbol{A}} \hat{f}(\boldsymbol{A}, \boldsymbol{Y})$) *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $x_{i,j} \sim_{iid} BG(\theta)$, $\boldsymbol{D}_o \in \mathsf{O}(n;\mathbb{R})$ is any orthogonal matrix, and $\boldsymbol{Y} = \boldsymbol{D}_o\boldsymbol{X}_o$. The expectation of $\nabla_{\boldsymbol{A}} \hat{f}(\boldsymbol{A}, \boldsymbol{Y})$ satisfies this property:*

$$\mathbb{E}_{\boldsymbol{X}_o}\big[\nabla_{\boldsymbol{A}} \hat{f}(\boldsymbol{A}, \boldsymbol{Y})\big] = 3p\theta(1-\theta)\nabla_{\boldsymbol{A}} g(\boldsymbol{A}\boldsymbol{D}_o) + 12p\theta^2\boldsymbol{A}. \tag{29}$$

Proof See B.2. ■

This proposition indicates that the expected stochastic gradient $\nabla_{\boldsymbol{A}} \hat{f}(\boldsymbol{A}, \boldsymbol{Y})$ agrees well with the gradient of the deterministic objective $g(\boldsymbol{A}\boldsymbol{D}_o)$, expect for a bias that is linear in the current estimate $\boldsymbol{A}$. This suggests a possible improvement for the MSP algorithm in the stochastic case: If we have knowledge about the θ (or can estimate it online), we could

subtract a bias term $\alpha \boldsymbol{A}$ ($\alpha \in (0, 12p\theta^2]$) from $\partial \boldsymbol{A}_t$ in Step 3 of Algorithm 2. One can verify experimentally that this indeed helps further accelerate the convergence of the algorithm.

Proposition 11 (Concentration Bound of $\frac{1}{np}\nabla \hat{f}(\cdot,\cdot)$) *If $\boldsymbol{X}_o \in \mathbb{R}^{n\times p}, x_{i,j} \sim_{iid} BG(\theta)$, for any $\boldsymbol{A}, \boldsymbol{D}_o \in \mathsf{O}(n;\mathbb{R})$, and $\boldsymbol{Y} = \boldsymbol{D}_o\boldsymbol{X}_o$, the following inequality holds*

$$\mathbb{P}\left(\sup_{\boldsymbol{A}\in\mathsf{O}(n;\mathbb{R})} \frac{1}{4np} \left\| \nabla_{\boldsymbol{A}} \hat{f}(\boldsymbol{A},\boldsymbol{Y}) - \mathbb{E}\big[\nabla_{\boldsymbol{A}} \hat{f}(\boldsymbol{A},\boldsymbol{Y})\big] \right\|_F \geq \delta \right) < \frac{1}{p}, \tag{30}$$

when $p = \Omega(\theta n^2 \ln n/\delta^2)$.

Proof See B.3. ■

4. Analysis of the MSP Algorithm

In this section, we provide convergence analysis of the proposed MSP Algorithm 1 that maximizes $g(\boldsymbol{A}\boldsymbol{D}_o)$ over the orthogonal group $\mathsf{O}(n;\mathbb{R})$. Each iteration of Algorithm 1 performs the following iteration:

$$\boldsymbol{A}_{t+1} = \mathcal{P}_{\mathsf{O}(n;\mathbb{R})}\big[(\boldsymbol{A}_t\boldsymbol{D}_o)^{\circ 3}\boldsymbol{D}_o^*\big], \tag{31}$$

notice that both $\boldsymbol{A}$ and $\boldsymbol{D}_o$ are orthogonal matrices, we can further reduce (31) to

$$\boldsymbol{A}_{t+1}\boldsymbol{D}_o = \mathcal{P}_{\mathsf{O}(n;\mathbb{R})}\big[(\boldsymbol{A}_t\boldsymbol{D}_o)^{\circ 3}\big]. \tag{32}$$

If we view $\boldsymbol{A}_t\boldsymbol{D}_o$ as another orthogonal matrix $\boldsymbol{W}_t$, the convergence analysis reduces to prove that the following iteration

$$\boldsymbol{W}_{t+1} = \mathcal{P}_{\mathsf{O}(n;\mathbb{R})}\big[(\boldsymbol{W}_t)^{\circ 3}\big] \tag{33}$$

will converge to a signed permutation matrix $\boldsymbol{W}_\infty$. Hence, we conclude that the MSP algorithm 1 for maximizing $g(\boldsymbol{A}\boldsymbol{D}_o)$ is invariant of orthogonal rotation. So without loss of generality, we only need to provide convergence analysis for the case $\boldsymbol{D}_o = \boldsymbol{I}$ (or slightly abuse the notation a bit by changing $\boldsymbol{W}_t$ in (33) into $\boldsymbol{A}_t$).

When $\boldsymbol{D}_o = \boldsymbol{I}$, we wish to show the MSP algorithm converges to a signed permutation matrix for the optimization problem:

$$\max_{\boldsymbol{A}} g(\boldsymbol{A}) = \|\boldsymbol{A}\|_4^4, \quad \text{subject to} \quad \boldsymbol{A} \in \mathsf{O}(n;\mathbb{R}), \tag{34}$$

starting from any randomly initialized $\boldsymbol{A}_0$ on $\mathsf{O}(n;\mathbb{R})$ with probability 1. For this purpose, we first introduce some basic properties of the space $\mathsf{O}(n;\mathbb{R})$ and our objective function $g(\cdot)$.

4.1. Properties of the Orthogonal Group

The orthogonal group is a special type of Stiefel manifold (Absil et al., 2009) with tangent space

$$T_{\boldsymbol{W}}\mathsf{O}(n;\mathbb{R}) \doteq \{\boldsymbol{Z} \mid \boldsymbol{Z}^*\boldsymbol{W} + \boldsymbol{W}^*\boldsymbol{Z} = \boldsymbol{0}\}, \tag{35}$$

and the projection operation $\mathcal{P}_{T_{\boldsymbol{W}}\mathsf{O}(n;\mathbb{R})} : \mathbb{R}^{n\times n} \to T_{\boldsymbol{W}}\mathsf{O}(n;\mathbb{R})$ onto tangent space of $\mathsf{O}(n;\mathbb{R})$ is defined as:

$$\mathcal{P}_{T_{\boldsymbol{W}}\mathsf{O}(n;\mathbb{R})}(\boldsymbol{Z}) \doteq \boldsymbol{Z} - \frac{1}{2}\boldsymbol{W}(\boldsymbol{W}^*\boldsymbol{Z} + \boldsymbol{Z}^*\boldsymbol{W}) = \frac{1}{2}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{Z}^*\boldsymbol{W}). \tag{36}$$

We use $\nabla_{\boldsymbol{W}} g(\boldsymbol{W})$ to denote the gradient of $g(\boldsymbol{W})$ w.r.t. $\boldsymbol{W}$ in $\mathbb{R}^{n\times n}$, and $\operatorname{grad} g(\boldsymbol{W})$ to denote the Riemannian gradient of $g(\boldsymbol{W})$ w.r.t. $\boldsymbol{W}$ on $T_{\boldsymbol{W}}\mathsf{O}(n;\mathbb{R})$. Thus, we can formulate the Riemannian gradient of $\boldsymbol{W}$ on $T_{\boldsymbol{W}}\mathsf{O}(n;\mathbb{R})$ as following:

$$\operatorname{grad} g(\boldsymbol{W}) = \mathcal{P}_{T_{\boldsymbol{W}}\mathsf{O}(n;\mathbb{R})}(\nabla_{\boldsymbol{W}} g(\boldsymbol{W})). \tag{37}$$

The following proposition introduces the critical points of $g(\boldsymbol{W})$ on $T_{\boldsymbol{W}}\mathsf{O}(n;\mathbb{R})$.

Proposition 12 *The critical points of $g(\boldsymbol{W})$ on manifold $\mathsf{O}(n;\mathbb{R})$ satisfies the following condition:*

$$(\boldsymbol{W}^{\circ 3})^*\boldsymbol{W} = \boldsymbol{W}^*\boldsymbol{W}^{\circ 3}. \tag{38}$$

Proof See C.1. ■

Therefore, $\forall \boldsymbol{W} \in \mathbb{R}^{n\times n}$ we can write critical points condition of ℓ^4-norm over $\mathsf{O}(n;\mathbb{R})$ as this following equations:

$$\begin{cases} (\boldsymbol{W}^{\circ 3})^*\boldsymbol{W} = \boldsymbol{W}^*\boldsymbol{W}^{\circ 3}, \\ \boldsymbol{W}^*\boldsymbol{W} = \boldsymbol{I}. \end{cases} \tag{39}$$

Since the orthogonal group $\mathsf{O}(n;\mathbb{R})$ is a continuous manifold in $\mathbb{R}^{n\times n}$ (Absil et al., 2009; Hall, 2015), this indicates that critical points of $g(\boldsymbol{W}) = \|\boldsymbol{W}\|_4^4$ over the orthogonal group $\boldsymbol{W} \in \mathsf{O}(n;\mathbb{R})$ has measure 0.

Proposition 13 *All global maximizers of ℓ^4-norm over the orthogonal group are isolated (nondegenerate) critical points.*

Proof See C.2. ■

We conjecture that the all local maximizers of the ℓ^4-norm over $\mathsf{O}(n;\mathbb{R})$ are global maximizers, as we will discuss in the next subsection.

4.2. Relation between MSP and Projected Gradient Ascent (PGA)

Although the MSP algorithm over orthogonal group (Algorithm 1) and for dictionary learning (Algorithm 2) has the same optimization procedure – they all performs *infinite* step size projected gradient ascent w.r.t. the objective function $\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4$ and $\|\boldsymbol{A}\boldsymbol{Y}\|_4^4$ respectively, we shall state their intrinsic difference clearly.

In Proposition 10 and Proposition 11, we can see that each iteration of the MSP algorithm for dictionary learning (Algorithm 2) concentrates onto one step of PGA w.r.t. the objective $\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4$ with *fixed step size* $\frac{1-\theta}{4\theta}$, while the MSP algorithm over the orthogonal group performs PGA with *infinite* step size. We test PGA Algorithm 3 with different step size α, as shown in Table 1 for detail. We observe that:

- PGA with *arbitrary* fixed step size find *global maximizers* of (34);

- PGA converges *faster* with *larger* step size.

This experimental phenomenon supports the efficiency of the MSP algorithm – it is indeed faster than any first order gradient method.

Algorithm 3 PGA for ℓ^4-Maximization over $\mathsf{O}(n;\mathbb{R})$

Initialize $\boldsymbol{A}_0 \in \mathsf{O}(n;\mathbb{R})$, step size $\alpha > 0$ ▷ Initialize $\boldsymbol{A}_0$, and α for iteration
for $t = 0, 1, ..., T-1$ **do**
  $\partial \boldsymbol{A}_t \doteq 4\boldsymbol{A}^{\circ 3}$ ▷ $\nabla_{\boldsymbol{A}} \|\boldsymbol{A}\|_4^4 = 4(\boldsymbol{A})^{\circ 3}$
  $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \mathrm{SVD}\big(\boldsymbol{A} + \alpha \partial \boldsymbol{A}_t\big)$
  $\boldsymbol{A}_{t+1} = \boldsymbol{U}\boldsymbol{V}^*$ ▷ Project $\boldsymbol{A}_{t+1}$ onto $\mathsf{O}(n;\mathbb{R})$
end for
Output $\boldsymbol{A}_T$

	Iterations			
	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	$\alpha = +\infty$
$\mathsf{O}(5;\mathbb{R})$	13	5	4	4
$\mathsf{O}(25;\mathbb{R})$	23	7	5	5
$\mathsf{O}(50;\mathbb{R})$	35	10	8	6
$\mathsf{O}(100;\mathbb{R})$	63	12	9	9
$\mathsf{O}(200;\mathbb{R})$	70	14	11	9

Table 1: Number of iteration for PGA (Algorithm 3) on orthogonal group of different dimension n to reach global maximizers, with the same initialization for each dimension n and different step size α. We directly apply the MSP Algorithm 1 when $\alpha = +\infty$.

4.3. Convergence Analysis

We first introduce the properties of the critical points of the ℓ^4-objective (34) in Proposition 14 and Proposition 15 will show that PGA with *any fixed step size* α (even $\alpha = +\infty$) finds a critical points of (34).

Proposition 14 (Fixed Point of the MSP Algorithm) *Given $\boldsymbol{W} \in \mathsf{O}(n;\mathbb{R})$, $\boldsymbol{W}$ is a fix point of the MSP Algorithm 1 if and only if $\boldsymbol{W}$ is a critical point of the ℓ^4-norm over $\mathsf{O}(n;\mathbb{R})$.*

Proof See C.3. ■

Proposition 15 (Convergence of PGA with Arbitrary Step Size) *Iterative PGA algorithm 3 with any fixed step size $\alpha > 0$ (α can be $+\infty$ and PGA is equivalent to MSP Algorithm 1 when $\alpha = +\infty$) finds a saddle point of optimization problem* (34)

$$\max_{\boldsymbol{A} \in \mathsf{O}(n;\mathbb{R})} \|\boldsymbol{A}\|_4^4 .$$

Proof See C.4. ■

In addition to Proposition 15, the intuition for *larger gradient converges faster* is due to the landscape of function $h(\boldsymbol{A})$. As we can see in (117), when α decreases, the landscape of $h(\boldsymbol{A})$ approaches becomes "flat", hence the convergence rate decreases. On the contrary, when $\alpha \to +\infty$, $h(\boldsymbol{A})$ preserves the "sharp" curvature of $\|\boldsymbol{A}\|_4^4$, which yields the optimal convergence rate.

Although the function $g(\boldsymbol{A}) = \|\boldsymbol{A}\|_4^4$ may have many critical points, the signed permutation group $\mathsf{SP}(n)$ are the only global maximizers. As recent work has shown (Sun et al., 2015), such discrete symmetry helps regulate the global landscape of the objective function and makes it amenable to global optimization. Indeed, we have observed through our extensive experiments that, under broad conditions, the proposed MSP algorithm always converges to the globally optimal solution (set), at a super-linear convergence rate.

We only give a local result on the convergence of the MSP algorithm in this paper.[8] That is, when the initial orthogonal matrix $\boldsymbol{A}$ is "close" enough to a signed permutation matrix, the MSP algorithm converges to that signed permutation at a very fast rate. It is easy to verify the algorithm is permutation invariant. Hence w.l.o.g., we may assume the target signed permutation is the identity $\boldsymbol{I}$.

Theorem 16 (Cubic Convergence Rate around Global Maximizers) *Given an orthogonal matrix $\boldsymbol{A} \in \mathsf{O}(n;\mathbb{R})$, let $\boldsymbol{A}'$ denote the output of the MSP Algorithm 1 after one iteration: $\boldsymbol{A}' = \boldsymbol{U}\boldsymbol{V}^*$, where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = SVD(\boldsymbol{A}^{\circ 3})$. If $\|\boldsymbol{A} - \boldsymbol{I}\|_F^2 = \varepsilon$, for $\varepsilon < 0.579$, then we have $\|\boldsymbol{A}' - \boldsymbol{I}\|_F^2 < \|\boldsymbol{A} - \boldsymbol{I}\|_F^2$ and $\|\boldsymbol{A}' - \boldsymbol{I}\|_F^2 < O(\varepsilon^3)$.*

Proof See C.5. ■

Theorem 16 shows that the MSP Algorithm 1 achieves cubic convergence rate locally, which is much faster than any gradient descent methods. Our experiments in Section 5 confirm this super-linear convergence rate for the MSP algorithms, at least in the deterministic case.

The above theorem only proves local convergence. As shown in section 5.2, We have observed in experiments that the algorithm actually converges *globally* under very broad conditions. We can see why this could be true in general from the special case when $n = 2$. Note that $\mathsf{O}(n,\mathbb{R})$ is a disjoint manifold with two continuous parts $\mathsf{SO}(n,\mathbb{R})$ and its reflection (Hall, 2015), for convenience, we only consider $\mathsf{SO}(n,\mathbb{R})$, which is a "half" of $\mathsf{O}(n,\mathbb{R})$, with determinant $det(\boldsymbol{A}) = 1$. The next lemma shows the global convergence of our MSP algorithm in $\mathsf{SO}(2;\mathbb{R})$.

8. We leave the study of ensuring global optimality and convergence to future work.

Proposition 17 (Global Convergence of the MSP Algorithm on $\mathsf{SO}(2;\mathbb{R})$) *When $n = 2$, if we parameterize our $\boldsymbol{A}_t \in \mathsf{SO}(2,\mathbb{R})$ as the following:*[9]

$$\boldsymbol{A}_t = \begin{pmatrix} \cos\theta_t & -\sin\theta_t \\ \sin\theta_t & \cos\theta_t \end{pmatrix}, \quad \forall \theta_t \in \Big[-\frac{\pi}{2}, \frac{\pi}{2}\Big], \tag{40}$$

then θ_t and θ_{t+1} satisfies the following relation

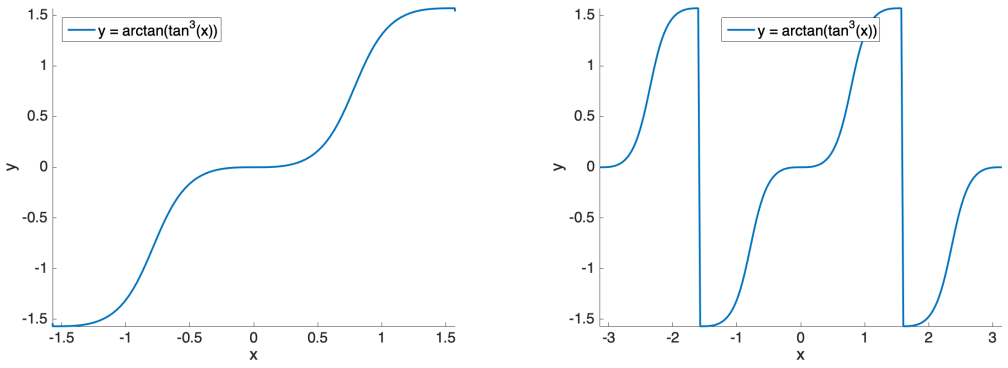
$$\theta_{t+1} = \tan^{-1}\big(\tan^3\theta_t\big). \tag{41}$$

Proof See C.6. ■

The previous proposition 17 indicates that the MSP algorithm is essentially conducting the following iteration:

$$\theta_{t+1} = \tan^{-1}\big(\tan^3\theta_t\big) \quad \text{or} \quad \tan\theta_{t+1} = \tan^3\theta_t. \tag{42}$$

See Figure 2 for a plot of this function. The iteration will converge superlinearly to the following values $\theta_\star = \pm\pi/2, 0, \pm\pi$, which all correspond to signed permutation matrices in $\mathrm{SP}(2)$. The fixed points of this iteration are $\theta = k\pi/4$, where $k \in \{-4,-3,-2,-1,0,1,2,3,4\}$. Among all these fixed points, the unstable ones are those with odd k, which correspond to Hadamard matrices. This phenomenon also justifies our experiments: Hadamard matrices are fixed point of the MSP iteration, but they are unstable and can be avoided with small random perturbation.



(a) $x \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ **(b)** $x \in [-\pi, \pi]$

Figure 2: Function $y = \tan^{-1}\big(\tan^3 x\big)$, for different domain of x

9. The result of this lemma shows an update on the $\tan x$ function, which is a periodic function with period π, so we set $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ to avoid the periodic ambiguity, one can easily generalize this result to the case where $\theta \in [-\pi, -\frac{\pi}{2}) \cup (\frac{\pi}{2}, \pi]$, since the signs of 1 won't affect our claim for pursuing a "signed permutation" matrix.

4.4. Generalized ℓ^{2k}-Norm Maximization

One may notice that we can also promote sparsity over the orthogonal group by maximizing the ℓ^{2k}-norm, $\forall k \geq 2$ on the orthogonal group:[10]

$$\max_{\boldsymbol{A}} \|\boldsymbol{A}\|_{2k}^{2k}, \quad \text{subject to} \quad \boldsymbol{A} \in \mathsf{O}(n;\mathbb{R}). \tag{43}$$

In fact, the resulting algorithm would have a higher rate of convergence for the deterministic case, as the stretching with the power $(\cdot)^{\circ 2k-1}$ sparsifies the matrix more significantly with a larger k. See Section 5.3 for experimental verification. The following example provides one run of MSP algorithm when $2k = 10$.

Example 2 (One Run of MSP Algorithm with $2k = 10$) *This example shows how a random orthogonal matrix $\boldsymbol{A}_0 \in \mathsf{O}(n;\mathbb{R})$ evolves into a signed permutation matrix, using ℓ^{10}-norm:*

$$\begin{aligned}
&\boldsymbol{A}_0 = \begin{pmatrix} -0.6142 & 0.3943 & 0.6836 \\ -0.2039 & 0.7575 & -0.6201 \\ 0.7623 & 0.5203 & 0.3849 \end{pmatrix} \xrightarrow{\textit{stretching}} \boldsymbol{A}_0^{\circ 9} = \begin{pmatrix} -0.0124 & 0.0002 & 0.0326 \\ -0.0000 & 0.0821 & -0.0136 \\ 0.0870 & 0.0028 & 0.0002 \end{pmatrix} \\
\xrightarrow{\textit{projection}} \quad &\boldsymbol{A}_1 = \begin{pmatrix} -0.2657 & 0.0908 & 0.9598 \\ -0.0150 & 0.9950 & -0.0983 \\ 0.9639 & 0.0406 & 0.2630 \end{pmatrix} \xrightarrow{\textit{stretching}} \boldsymbol{A}_1^{\circ 9} = \begin{pmatrix} -0.0000 & 0.0000 & 0.6910 \\ -0.0000 & 0.9562 & -0.0000 \\ 0.7185 & 0.0000 & 0.0000 \end{pmatrix} \\
\xrightarrow{\textit{projection}} \quad &\boldsymbol{A}_2 = \begin{pmatrix} 0.0000 & -0.0000 & 1.0000 \\ 0.0000 & 1.0000 & -0.0000 \\ 1.0000 & -0.0000 & 0.0000 \end{pmatrix} \xrightarrow{\textit{output}} \boldsymbol{A}_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.
\end{aligned}$$

Comparing with Example 1, where ℓ^4-norm MSP algorithm takes 6 iterations, the ℓ^{10}-norm MSP finds a signed permutation matrix in only 2 iterations.

In fact, one can show that if $2k \to \infty$, the corresponding MSP algorithm converges with *only one* iteration for the deterministic case!

So why don't we choose ℓ^∞-norm instead? Notice that the above behavior is for the expectation of the objective function we actually care here. For dictionary learning, the object function $\hat{f}(\boldsymbol{A}, \boldsymbol{Y})$ depends on the random samples $\boldsymbol{Y}$. Both $\hat{f}(\boldsymbol{A}, \boldsymbol{Y})$ and $\nabla_{\boldsymbol{A}} \hat{f}(\boldsymbol{A}, \boldsymbol{Y})$ concentrate on their expectations only when p is large enough. For the same error threshold, if we choose larger k, then we will need much larger sample size p for $\hat{f}(\boldsymbol{A}, \boldsymbol{Y})$, $(\boldsymbol{A}\boldsymbol{Y})^{\circ 2k-1}\boldsymbol{Y}^*$ to concentrate on their expectation $f(\boldsymbol{A})$, $\mathbb{E}[(\boldsymbol{A}\boldsymbol{Y})^{\circ 2k-1}\boldsymbol{Y}^*]$ respectively. As we will see in Section 5.3 from experiments, the choice of $2k = 4$ seems to be the best in terms of balancing these two contending factors of convergence rate and sample size.

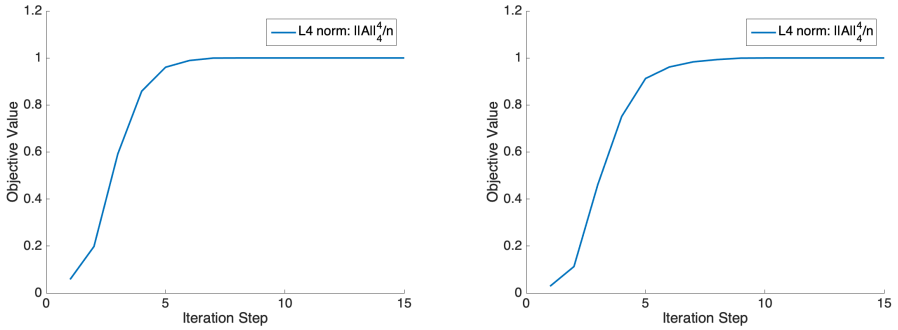
5. Experimental Verification

5.1. ℓ^4-Norm Maximization over the Orthogonal Group

First, to have some basic ideas about how fast and smoothly the proposed MSP algorithm maximizes ℓ^4-norm. Figure 3 shows one run of the MSP Algorithm 1 that maximizes $\|\boldsymbol{A}\|_4^4/n$. It reaches global maxima in less than 10 iterations for $n = 50$ and $n = 100$.

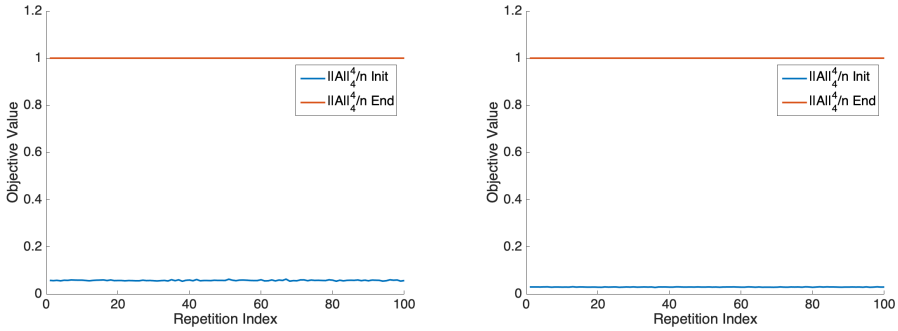
Next we run 100 trials of the algorithm to maximize $\|\boldsymbol{A}\|_4^4/n$ on $\mathsf{O}(n;\mathbb{R})$ with random initialization. Figure 4 shows that all 100 trials of the MSP algorithm reach the global maximum 1. According to Lemma 5, this indicates that all 100 trials (in different dimension $n = 50, 100$) converge to sign permutation matrices in $\mathsf{SP}(n)$.

10. Note that when $2k = 2$, $\|\boldsymbol{A}\|_2^2 = n$ is a constant, so ℓ^{2k}-norm only promotes sparsity when $2k \geq 4$.



(a) $n = 50$ **(b)** $n = 100$

Figure 3: Normalized $\|\boldsymbol{A}\|_4^4$ in one run of MSP Algorithm 1 for maximizing $\|\boldsymbol{A}\|_4^4$ on $\mathsf{O}(n;\mathbb{R})$



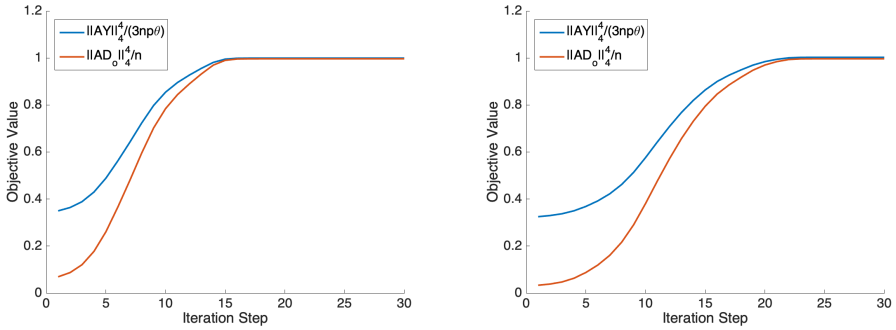
(a) $n = 50$ **(b)** $n = 100$

Figure 4: Normalized $\|\boldsymbol{A}\|_4^4/n$ for 100 trials of the MSP Algorithm 1 on $\mathsf{O}(n;\mathbb{R})$

5.2. Dictionary Learning via MSP

Figure 5(a) presents one trial of the proposed MSP Algorithm 2 for dictionary learning with $\theta = 0.3$, $n = 50$, and $p = 20,000$. The result corroborates with statements in Lemma 4 and Lemma 3: Maximizing $\hat{f}(\boldsymbol{A},\boldsymbol{Y})$ is largely equivalent to optimizing $g(\boldsymbol{A}\boldsymbol{D}_o)$, and both values reach global maximum at the same time. Meanwhile, this result also shows our MSP algorithm is able to find the global maximum at ease, since $g(\boldsymbol{A}\boldsymbol{D}_o)$ reaches its maximal value 1 (with minor errors) by maximizing $\hat{f}(\boldsymbol{A},\boldsymbol{Y})$. In Figure 5(b), we test the MSP Algorithm 2 in higher dimension $n = 100, p = 40,000, \theta = 0.3$. In both cases, our algorithm is surprisingly efficient: It only takes around 20 iterations to recover the ground truth dictionary.

In Figure 6, we run the MSP Algorithm 2 for 100 random trials with the settings $n = 50, p = 20,000, \theta = 0.3$ and $n = 100, p = 40,000, \theta = 0.3$. Among all 100 trials, $g(\boldsymbol{A}\boldsymbol{D}_o)$ achieve the global maximal value (within statistical errors) via optimizing $\hat{f}(\boldsymbol{A},\boldsymbol{Y})$ in less than 30 iterations. This experiment seems to support a conjecture: Within conditions of this experiment, the MSP algorithm recovers the globally optimal dictionary.



(a) $n = 50, p = 20,000, \theta = 0.3$ **(b)** $n = 100, p = 40,000, \theta = 0.3$

Figure 5: Normalized objective value $\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4/n$ and $\|\boldsymbol{A}\boldsymbol{Y}\|_4^4/3np\theta$ for individual trials of the MSP Algorithm 2, with different parameters n, p, θ. According to Lemma 5, $g(\boldsymbol{A}\boldsymbol{D}_o)/n$ reaches 1 indicates successful recovery for $\boldsymbol{D}_o$. This experiment shows the MSP algorithm finds global maxima of $\hat{f}(\boldsymbol{A}, \boldsymbol{Y})$ thus recovers the correct dictionary $\boldsymbol{D}_o$.

5.3. ℓ^{2k}-Norm Maximization over the Orthogonal Group

In Figure 7, we conduct experiments to support the choice of ℓ^4-norm. Figure 7(a) shows that for the deterministic case, the MSP Algorithm 1 finds signed permutation matrices faster with higher order ℓ^{2k}-norm. But Figure 7(b) indicates that as the order $2k$ increases, much more samples are needed by Algorithm 2 to achieve the same estimation error: p grows drastically as k increases. Hence, among all these sparsity-promoting norms (ℓ^{2k}-norm), the ℓ^4-norm strikes a good balance between sample size and convergence rate.[11]

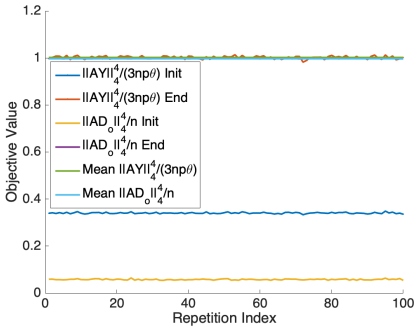
5.4. Phase Transition Plots for Working Ranges of the MSP Algorithm

Encouraged by previous experiments, we conduct more extensive experiments of the MSP Algorithm 2 in broader settings to find its working range: 1) Figure 8(a) shows the result of varying the sparsity level θ (from 0 to 1) and sample size p (from 500 to 50,000) with a fixed dimension $n = 50$; 2) Figure 8(b) and (c) show results of changing dimension n (from 10 to 1,000) and sample size p (from 1,000 to 100,000) at a fixed sparsity level $\theta = 0.5$. Notice that all figures demonstrate a clear phase transition for the working range.

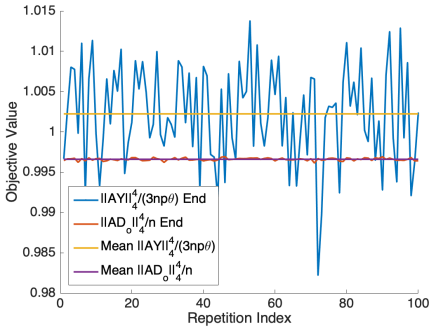
It is somewhat surprising to see in Figure 8(a) that the MSP algorithm is able to recover the dictionary correctly up to the sparsity level of $\theta \approx 0.6$ if p is large enough, which almost doubles the best existing theoretical guarantee given in Bai et al. (2018); Sun et al. (2015). We shall note that the inconsistency between the non linear phase transition curve in Figure 8(a) and the sample complexity $p = \Omega(\theta n^2 \ln n/\varepsilon^2)$ is due to the constant $C > \frac{4}{3\theta(1-\theta)}$ in Theorem 7.

Figure 8(b) and (c) show the working range for varying n, p with a fixed $\theta = 0.5$. Figure 8(b) is for a smaller range of n (from 10 to 100) and Figure 8(c) for a larger range of n (from

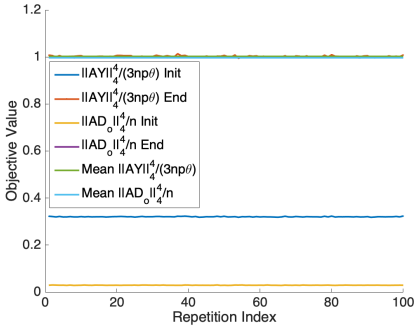
11. Later works (Shen et al., 2020; Xue et al., 2020) have shown that ℓ^3-norm maximization also has the same sparsifying effect.



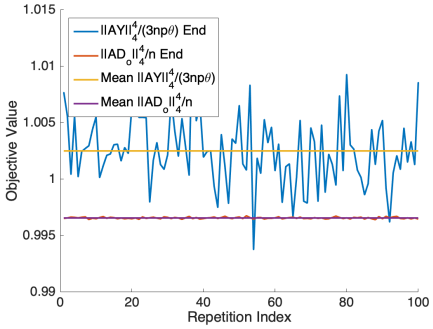
(a) $n = 50, p = 20,000, \theta = 0.3$, w/ initial values



(b) $n = 50, p = 20,000, \theta = 0.3$, w/o initial values



(c) $n = 100, p = 40,000, \theta = 0.3$, w/ initial values



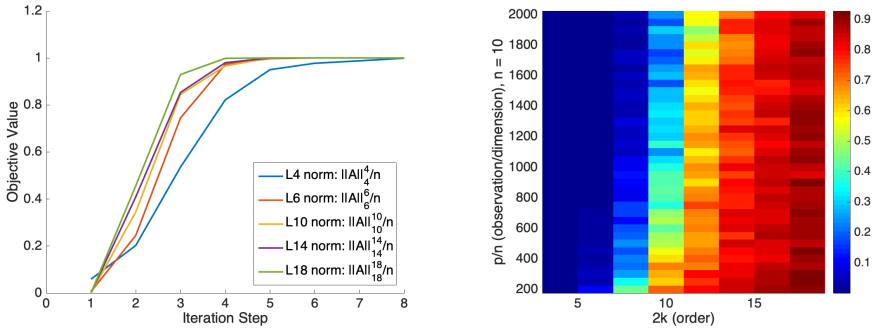
(d) $n = 100, p = 40,000, \theta = 0.3$, w/o initial values

Figure 6: Normalized initial and final objective values of $\|\boldsymbol{AD}_o\|_4^4/n$ and $\|\boldsymbol{AY}\|_4^4/3np\theta$ for 100 trials of the MSP Algorithm 2, with $n = 50$ and $n = 100$. Both objectives converge to 1 (with minor errors) for all 100 trials.

100 to 1,000). Figures (b) and (c) imply that the required sample size p for the algorithm to succeed seems to be quadratic in the dimension n: $p = \Omega(n^2)$, which corresponds to our statistical results ($p = \Omega(\theta n^2 \ln n)$) in Theorem 7 and Proposition 11. This empirical bound is significantly better than the best theoretical bounds given in Bai et al. (2018); Sun et al. (2015) for the sample size required to ensure success. Similar observations have been reported in Schramm and Steurer (2017); Bai et al. (2018).

5.5. Comparison with Prior Arts

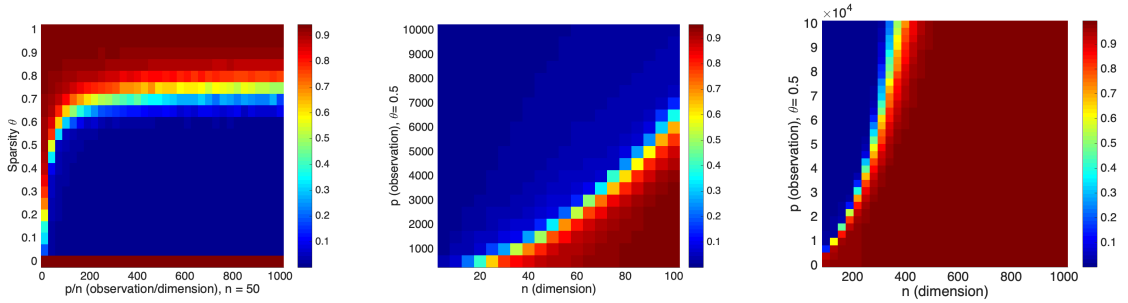
Table 2 compares the MSP method with the KSVD (Aharon et al., 2006), the SPAMS dictionary learning package Jenatton et al. (2010), and the latest subgradient method (Bai et al., 2018) for different choices of n, p under the same sparsity level $\theta = 0.3$. As one may see, our algorithm is significantly faster than the other algorithms in all trials. Further more, our algorithm has the potential for large scale experiments: It only takes 374.2 seconds to learn a 400×400 dictionaries from $160,000$ samples. While the previous algorithms either fails to



(a) Convergence plots of the MSP Algorithm 1 for the deterministic case, with the same initialization.

(b) Average normalized error of MSP Algorithm 2 among 20 trials, varying k and p, with $n = 10$ fixed.

Figure 7: Use different ℓ^{2k}-norms for Algorithm 1 and Algorithm 2.



(a) Changing θ from 0 to 1 and p from 500 to 10,000, $n = 50$.

(b) Changing n from 10 to 100 and p from 10^3 to 10^4, $\theta = 0.5$.

(c) Changing n from 100 to 1,000 and p from 10^4 to 10^5, $\theta = 0.5$.

Figure 8: Phase transition of average normalized error $|1 - \|\boldsymbol{A}\boldsymbol{D}_o\|_4^4/n|$ of 10 random trials for the MSP Algorithm 2 in different settings: **(a):** Varying θ, p with fixed n; **(b), (c):** Varying n, p with fixed θ. Red area indicates large error and blue area small error.

find the correct dictionary or is barely applicable. Within statistical errors, our algorithm gives slightly smaller values for $\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4/n$ in some trials. But the subgradient method (Bai et al., 2018) uses information of the ground truth dictionary $\boldsymbol{D}_o$ in their stopping criteria. Our MSP algorithm removes this dependency with only mild loss in accuracy. We shall note that traditional ℓ^1-minimization based dictionary learning is more accurate than the ℓ^4-maximization based method, since ℓ^1-minimization will decrease small entries to 0 while the ℓ^4-maximization tolerate small noise. Meanwhile, the numerical implementation also affects the experimental performance – the error of SPAMS is consistently better than the Subgradient method in terms of both accuracy and speed.

Although the convex relaxation based on sum-of-square SDP hierarchy (Barak et al., 2015; Ma et al., 2016; Schramm and Steurer, 2017) has theoretical guarantee for correctness under some specific statistical model, we do not include experimental comparison with the

sum-of-square method only because the SoS method incurs high computational complexity in solving large scale SDP programmings or tensor decomposition problems. In fact, on the same computing devices, sum-of-square methods cannot even solve the smallest dictionary learning problem ($n = 25$) in Table 2 due to memory and computation complexity.

			KSVD		SPAMS		Subgradient		MSP (Ours)	
n	p ($\times 10^4$)	θ	Error	Time	Error	Time	Error	Time	Error	Time
25	1	0.3	12.16%	64.5s	0.02%	5.53s	0.25%	9.2s	0.35%	**0.2s** (15 iter)
50	2	0.3	7.79%	165.4s	0.02%	23.4s	0.27%	107.0s	0.34%	**1.3s** (20 iter)
100	4	0.3	8.34%	569.4s	0.02%	365.9s	0.27%	2807.9s	0.35%	**7.2s** (25 iter)
200	8	0.3	8.46%	1064.0s	0.02%	5420.8s	N/A	>12h	0.35%	**56.0s** (40 iter)
400	16	0.3	11.94%	4646.3s	N/A	> 12h	N/A	>12h	0.35%	**494.4s** (60 iter)

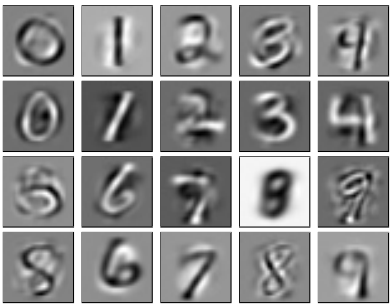
Table 2: Comparison experiments with KSVD (Aharon et al., 2006), SPAMS Jenatton et al. (2010), and Subgradient method (Bai et al., 2018) in different trials of dictionary learning: (a) $n = 25, p = 1 \times 10^4, \theta = 0.3$; (b) $n = 50, p = 2 \times 10^4, \theta = 0.3$; (c) $n = 100, p = 4 \times 10^4, \theta = 0.3$; (d) $n = 200, p = 4 \times 10^4, \theta = 0.3$; (e) $n = 400, p = 16 \times 10^4, \theta = 0.3$. Recovery error is measured as $\left|1 - \|\boldsymbol{A}\boldsymbol{D}_o\|_4^4/n\right|$, since Lemma 5 shows that a perfect recovery gives $\|\boldsymbol{A}\boldsymbol{D}_o\|_4^4/n = 1$. All experiments are averaged among 5 trials and conducted on a 2.7 GHz Intel Core i5 processor (CPU of a 13-inch Mac Pro 2015).

5.6. Learning Sparsifying Dictionaries of Real Images

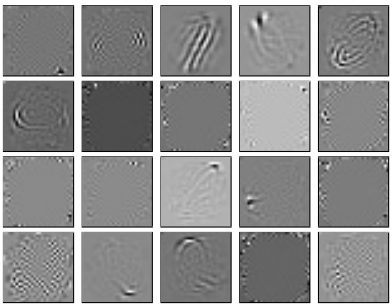
Indeed, the theoretical results of the MSP algorithm relies heavily on the Bernoulli-Gaussian assumption of the ground truth sparse code $\boldsymbol{x}_i$, that is, all $\boldsymbol{x}_i$ has the same element-wise variance. In order to demonstrate that the MSP algorithm can be applied to broader application scenarios beyond the Bernoulli-Gaussian setting, we test our algorithm on the MNIST dataset of hand-written digits (LeCun et al., 1998), whose element-wise (pixel-wise) variance are barely equal. In our implementation, we vectorize each 28×28 image j in the MNIST dataset into a vector $\boldsymbol{y}_i \in \mathbb{R}^{784}$ and directly apply our MSP algorithm 2 to the whole data set $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{50,000}]$.[12] Figure 9(a) shows some bases learned from these images which obviously capture shapes of the digits. Figure 9(b) shows some less significant base vectors in the space of $\mathbb{R}^{784}$ that are orthogonal to the above.

We also compare our results with bases learned from Principal Component Analysis (PCA). As shown in Figure 10(a), the top bases learned from PCA are blurred with shapes from different digits mixed together, unlike those bases learned from the MSP in Figure 9(a) that clearly capture features of multiple different digits.

12. The training set of MNIST contains 50,000 images of size 28×28.

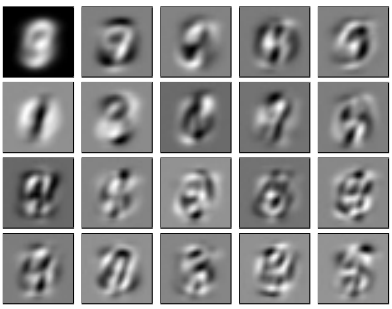


(a) Some selected "meaningful" bases learned from the MNIST dataset.

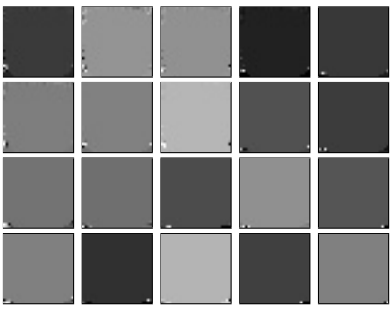


(b) Other bases learned the MNIST dataset.

Figure 9: Learned dictionary through the MSP algorithm 2 on the MNIST dataset



(a) Top 20 bases from PCA

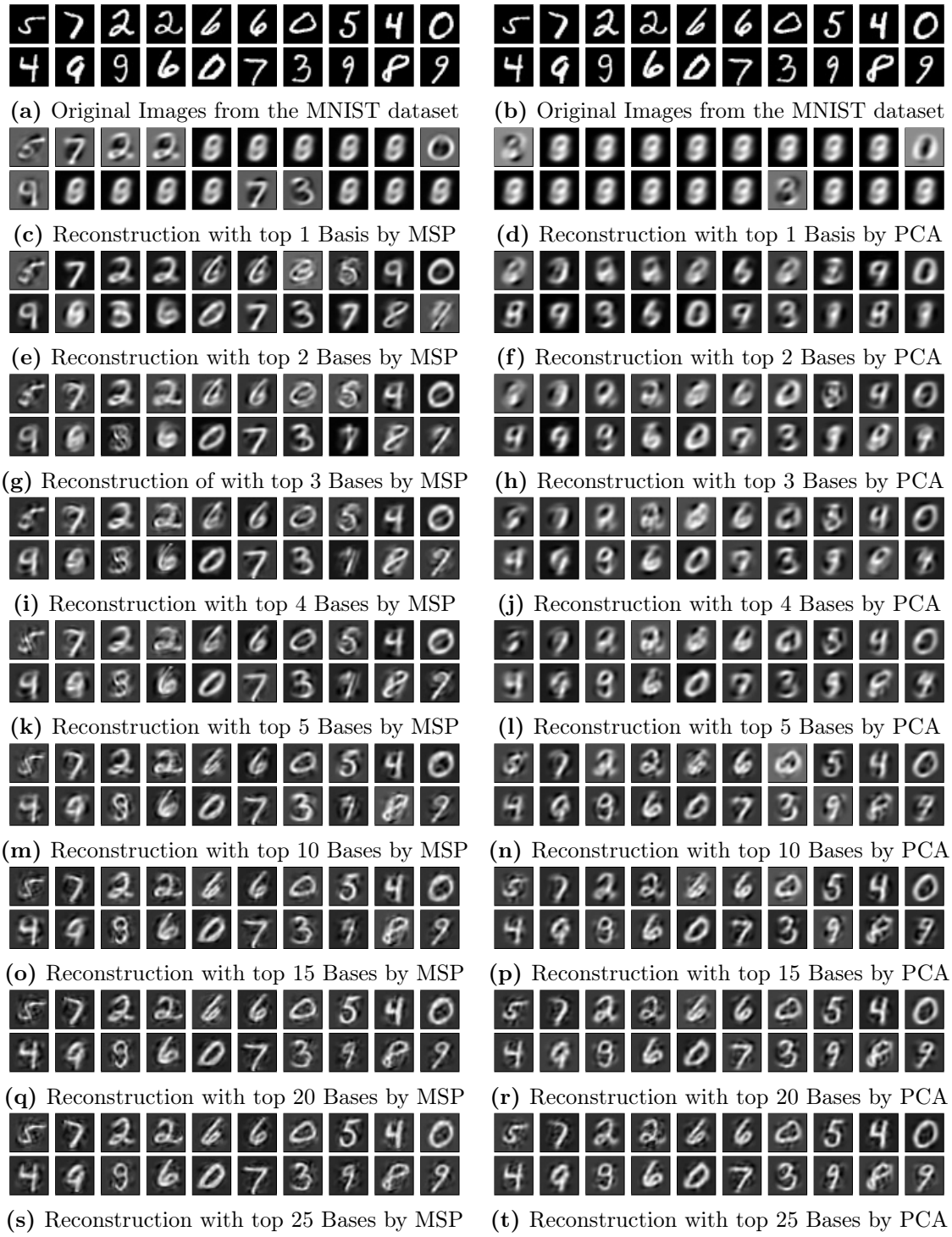


(b) Last 20 bases from PCA

Figure 10: Learned Bases by PCA on MNIST dataset

We test the reconstruction results using the learned dictionary from the MSP algorithm and compare the reconstruction results using bases learned from PCA. In Figure 11, we compare the original MNIST data (LeCun et al., 1998), the reconstruction results from the MSP algorithm and PCA, using the top 1, 2, 3, 4, 5, 10, and 25 bases, respectively. As we can see in Figure 11, the reconstruction results for both the MSP and PCA improve, when the number of used bases increases. As shown in the case when only 2 bases are used, the MSP algorithm is able to represent clear shape for most digits already (see the three top left images), while PCA only provides mostly blurred results. When top 5 bases are used, the MSP algorithm almost capture all information of the input images, while PCA still gives rather blurred reconstruction. Both methods are able to capture salient information when the number of used bases are increased above 10. This is rather reasonable as there are about 10 different digits in this dataset and our method is precisely expected to have advantages when the number of bases is below 10.

This experiment has shown that the proposed MSP algorithm can be applied to more general setting beyond the Bernoulli-Gaussian setting. Moreover, due to its efficiency, the MSP algorithm can be extend to other large scale visual dataset such as CIFAR10 (Krizhevsky et al., 2014; Zhai et al., 2019).



(a) Original Images from the MNIST dataset **(b)** Original Images from the MNIST dataset

(c) Reconstruction with top 1 Basis by MSP **(d)** Reconstruction with top 1 Basis by PCA

(e) Reconstruction with top 2 Bases by MSP **(f)** Reconstruction with top 2 Bases by PCA

(g) Reconstruction of with top 3 Bases by MSP **(h)** Reconstruction with top 3 Bases by PCA

(i) Reconstruction with top 4 Bases by MSP **(j)** Reconstruction with top 4 Bases by PCA

(k) Reconstruction with top 5 Bases by MSP **(l)** Reconstruction with top 5 Bases by PCA

(m) Reconstruction with top 10 Bases by MSP **(n)** Reconstruction with top 10 Bases by PCA

(o) Reconstruction with top 15 Bases by MSP **(p)** Reconstruction with top 15 Bases by PCA

(q) Reconstruction with top 20 Bases by MSP **(r)** Reconstruction with top 20 Bases by PCA

(s) Reconstruction with top 25 Bases by MSP **(t)** Reconstruction with top 25 Bases by PCA

Figure 11: Comparison of compact representation with learned dictionary from the MSP algorithm 2 with the PCA bases for the MNIST image dataset.

6. Conclusions and Discussions

In this paper, we see that a complete dictionary can be effectively and efficiently learned with nearly minimum sample complexity by the proposed simple MSP algorithm. The computational complexity of the algorithm is essentially a few dozens of SVDs, allowing us to learn dictionary for very high-dimensional data.

Regarding sample complexity, the two main measure concentration results of Theorem 7 and Proposition 11 both require a sample complexity of $p = \Omega(\theta n^2 \log n/\varepsilon^2)$ for the global maximizers of the ℓ^4-objective to be (close to) the correct $n \times n$ dictionary. This bound is consistent with our experiments in section 5.4. Barak et al. (2015); Ma et al. (2016); Schramm and Steurer (2017); Bai et al. (2018) have reported similar empirical evidences that at least $p = \Omega(n^2)$ samples are needed to recover an $n \times n$ dictionary.

Regarding the order of ℓ^{2k}-norm, we adopt ℓ^4-norm because $2k = 4$ is the minimum even order that promotes sparsity. However, later works (Shen et al., 2020; Xue et al., 2020) show that ℓ^3-norm maximization also promotes sparsity and the MSP algorithm (PGA with infinite step size) for ℓ^4-norm maximization naturally generalizes to ℓ^3-norm maximization.

Regarding optimality, Proposition 15 shows than random initialization of PGA algorithm with any step size finds a critical points of the ℓ^4-norm over $\mathsf{O}(n;\mathbb{R})$, Theorem 16 has shown the local convergence of the proposed MSP Algorithm 1 with a cubic rate for general n around global maximizers, and Proposition 17 has proven its global convergence for $n = 2$. It is natural to conjecture that the signed-permutations would be the only stable maximizers of the ℓ^4-norm over the entire orthogonal group, hence the proposed algorithm converges globally, just like the experiments have indicated. Nevertheless, a rigorous proof is still elusive at this point.

Although some initial experiments have already suggested that the proposed algorithm works stably with mild noise and real data – showing clear advantages of the so learned dictionary over the classic PCA bases, the proposed algorithm actually generalizes well to data with dense noise, outliers, and sparse corruptions (Zhai et al., 2019). This paper has only addressed the case with a complete (square) dictionary. But there are ample reasons to believe that similar formulations and techniques presented in this paper can be extended to the over-complete case $\boldsymbol{D} \in \mathbb{R}^{n\times m}$ with $n < m$, at least when $m = O(n)$.

7. Acknowledgement

We would like to thank Yichao Zhou and Dr. Chong You of Berkeley EECS Department for stimulating discussions during preparation of this manuscript. We would like to thank professor Ju Sun of University of Minnesota CSE and Dr. Julien Mairal of Inria for references to existing work and experimental comparison. We would also like to thank Haozhi Qi of Berkeley EECS for help with the experiments. YZ would like to thank professor Yuejie Chi of CMU ECE for meaningful comments on the local convergence result. YZ would also like to thank Ye Xue from HKUST for insightful suggestions and proofreading. Yi likes to thank professor Bernd Sturmfels of Berkeley Math Department for help with analyzing algebraic properties of critical points of the ℓ^4-norm over the orthogonal group and thank professor Alan Weinstein of Berkeley Math Department for discussions and references on the role of ℓ^4-norm in spherical harmonic analysis. Yi would like to acknowledge that

this work is partially supported by research grant N00014-20-1-2002 from Office of Naval Research (ONR) and research grant from Tsinghua-Berkeley Shenzhen Institute (TBSI). JW gratefully acknowledges support from NSF grants 1733857, 1838061, 1740833, and 1740391, and thanks Sam Buchanan for discussions related to the geometry of ℓ^4.

Appendices

A. Proofs of Section 2

A.1. Proof of Lemma 3

Claim 18 *$\forall \theta \in (0,1)$, let $\boldsymbol{X}_o \in \mathbb{R}^{n\times p}$, $x_{i,j} \sim_{iid} BG(\theta)$, $\boldsymbol{D}_o \in \mathsf{O}(n;\mathbb{R})$ is any orthogonal matrix, and $\boldsymbol{Y} = \boldsymbol{D}_o\boldsymbol{X}_o$. Then, $\forall \boldsymbol{A} \in \mathsf{O}(n;\mathbb{R})$, we have*

$$\frac{1}{3p\theta} f(\boldsymbol{A}) = (1-\theta)g(\boldsymbol{A}\boldsymbol{D}_o) + \theta n. \tag{44}$$

Proof For simplicity, since $\boldsymbol{D}_o \in \mathsf{O}(n;\mathbb{R})$, let $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{D}_o$, we know that $\boldsymbol{W} \in \mathsf{O}(n;\mathbb{R})$. By the fact that $\boldsymbol{W} \in \mathsf{O}(n;\mathbb{R})$, we have:

$$\begin{aligned} n = \sum_{i=1}^{n}\Big(\sum_{k=1}^{n} w_{i,k}^2\Big)^2 &= \underbrace{\sum_{i=1}^{n}\sum_{k=1}^{n} w_{i,k}^4}_{g(\boldsymbol{W})} + 2\sum_{i=1}^{n}\sum_{1\le k_1<k_2\le n} w_{i,k_1}^2 w_{i,k_2}^2 \\ \Longrightarrow \sum_{i=1}^{n}\sum_{1\le k_1<k_2\le n} w_{i,k_1}^2 w_{i,k_2}^2 &= \frac{n-g(\boldsymbol{W})}{2}. \end{aligned} \tag{45}$$

Next, we calculate $\frac{1}{3p\theta} f(\boldsymbol{A})$:

$$\begin{aligned} \frac{f(\boldsymbol{A})}{3p\theta} &= \frac{1}{3p\theta}\mathbb{E}_{\boldsymbol{X}_o}\hat{f}(\boldsymbol{A};\boldsymbol{Y}) = \frac{1}{3p\theta}\mathbb{E}_{\boldsymbol{X}_o}\|\boldsymbol{A}\boldsymbol{D}_o\boldsymbol{X}_o\|_4^4 \\ &= \frac{1}{3p\theta}\mathbb{E}_{\boldsymbol{X}_o}\|\boldsymbol{W}\boldsymbol{X}_o\|_4^4 = \frac{1}{3p\theta}\sum_{i=1}^{n}\sum_{j=1}^{p}\mathbb{E}_{\boldsymbol{X}_0}\Big(\sum_{k=1}^{n} w_{i,k}x_{k,j}\Big)^4 \\ &= \frac{1}{3p\theta}\sum_{i=1}^{n}\sum_{j=1}^{p}\Big[3\theta\sum_{k=1}^{n} w_{i,k}^4 + 6\theta^2\sum_{1\le k_1<k_2\le n} w_{i,k_1}^2 w_{i,k_2}^2\Big] \\ &= \underbrace{\sum_{i=1}^{n}\sum_{k=1}^{n} w_{i,k}^4}_{g(\boldsymbol{W})} + 2\theta\underbrace{\sum_{i=1}^{n}\sum_{1\le k_1<k_2\le n} w_{i,k_1}^2 w_{i,k_2}^2}_{\frac{n-g(\boldsymbol{W})}{2}} \\ &= (1-\theta)g(\boldsymbol{W}) + \theta n = (1-\theta)g(\boldsymbol{A}\boldsymbol{D}_o) + \theta n \quad \text{(By (45))}, \end{aligned} \tag{46}$$

which completes the proof. ■

A.2. Proof of Lemma 4

Claim 19 (Concentration Bound of $\frac{1}{np}\hat{f}(\cdot,\cdot)$) *$\forall\theta\in(0,1)$, if $\boldsymbol{X}\in\mathbb{R}^{n\times p}$, $x_{i,j}\sim_{iid} BG(\theta)$, for any $\delta>0$, the following inequality holds*

$$\begin{aligned}&\mathbb{P}\left(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})}\frac{1}{np}\left|\|\boldsymbol{W}\boldsymbol{X}\|_4^4-\mathbb{E}\,\|\boldsymbol{W}\boldsymbol{X}\|_4^4\right|\geq\delta\right)\\&<\exp\left(-\frac{3p\delta^2}{c_1\theta+8n(\ln p)^4\delta}+n^2\ln\Big(\frac{60np(\ln p)^4}{\delta}\Big)\right)\\&\quad+\exp\left(-\frac{p\delta^2}{c_2\theta}+n^2\ln\Big(\frac{60np(\ln p)^4}{\delta}\Big)\right)+2np\theta\exp\left(-\frac{(\ln p)^2}{2}\right),\end{aligned}\tag{47}$$

for some constants $c_1>10^4, c_2>3360$. Moreover

$$\begin{aligned}&\exp\left(-\frac{3p\delta^2}{c_1\theta+8n(\ln p)^4\delta}+n^2\ln\Big(\frac{60np(\ln p)^4}{\delta}\Big)\right)\\&+\exp\left(-\frac{p\delta^2}{c_2\theta}+n^2\ln\Big(\frac{60np(\ln p)^4}{\delta}\Big)\right)+2np\theta\exp\left(-\frac{(\ln p)^2}{2}\right)\leq\frac{1}{p},\end{aligned}\tag{48}$$

when $p=\Omega(\theta n^2\ln n/\delta^2)$.

Proof Let $\bar{\boldsymbol{X}}\in\mathbb{R}^{n\times p}$ denote the truncated $\boldsymbol{X}$ by bound B

$$\bar{x}_{i,j}=\begin{cases}x_{i,j} & \text{if}\quad |x_{i,j}|\leq B\\ 0 & \text{else}\end{cases}.\tag{49}$$

By Lemma 34, we know that $\|\boldsymbol{X}\|_\infty\leq B$ happens with probability at least $1-2np\theta\exp(-B^2/2)$ and $\bar{\boldsymbol{X}}=\boldsymbol{X}$ holds whenever $\|\boldsymbol{X}\|_\infty\leq B$. So we know that $\bar{\boldsymbol{X}}\neq\boldsymbol{X}$ holds with probability at most $2np\theta\exp(-B^2/2)$, and thus:

$$\begin{aligned}&\mathbb{P}\left(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})}\frac{1}{np}\left|\|\boldsymbol{W}\boldsymbol{X}\|_4^4-\mathbb{E}\,\|\boldsymbol{W}\boldsymbol{X}\|_4^4\right|\geq\delta\right)\\\leq&\mathbb{P}\left(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})}\frac{1}{np}\left|\|\boldsymbol{W}\boldsymbol{X}\|_4^4-\mathbb{E}\,\|\boldsymbol{W}\boldsymbol{X}\|_4^4\right|\geq\delta,\boldsymbol{X}=\bar{\boldsymbol{X}}\right)+\mathbb{P}\Big(\boldsymbol{X}\neq\bar{\boldsymbol{X}}\Big)\\\leq&\mathbb{P}\left(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})}\frac{1}{np}\left|\|\boldsymbol{W}\bar{\boldsymbol{X}}\|_4^4-\mathbb{E}\,\|\boldsymbol{W}\boldsymbol{X}\|_4^4\right|\geq\delta\right)+2np\theta e^{-\frac{B^2}{2}}.\end{aligned}\tag{50}$$

$\varepsilon-$net Covering. For any positive ε satisfying:

$$\varepsilon\leq\frac{\delta}{10npB^4},\tag{51}$$

by Lemma 35, there exists an $\varepsilon-$nets:

$$\mathcal{S}_\varepsilon=\{\boldsymbol{W}_1,\boldsymbol{W}_2,\ldots,\boldsymbol{W}_{|\mathcal{S}_\varepsilon|}\},\tag{52}$$

which covers $\mathsf{O}(n;\mathbb{R})$

$$\mathsf{O}(n;\mathbb{R}) \subset \bigcup_{l=1}^{|\mathcal{S}_\varepsilon|} \mathbb{B}(\boldsymbol{W}_l, \varepsilon), \tag{53}$$

in operator norm $\|\cdot\|_2$. Moreover, we have:

$$|\mathcal{S}_\varepsilon| \le \Big(\frac{6}{\varepsilon}\Big)^{n^2}. \tag{54}$$

So $\forall \boldsymbol{W} \in \mathsf{O}(n;\mathbb{R})$, there exists $l \in [|\mathcal{S}_\varepsilon|]$, such that $\|\boldsymbol{W} - \boldsymbol{W}_l\|_2 \le \varepsilon$. Thus, we have:

$$\begin{aligned}
&\mathbb{P}\bigg(\frac{1}{np}\Big|\big\|\boldsymbol{W}\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}\boldsymbol{X}\|_4^4\Big| \ge \delta\bigg)\\
\le&\mathbb{P}\bigg(\frac{1}{np}\Big(\Big|\big\|\boldsymbol{W}\bar{\boldsymbol{X}}\big\|_4^4 - \big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4\Big| + \Big|\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{X}\|_4^4\Big| + \Big|\mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{X}\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}\boldsymbol{X}\|_4^4\Big|\Big) \ge \delta\bigg)\\
\le&\mathbb{P}\bigg(\frac{1}{np}\Big|\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{X}\|_4^4\Big| + [4npB^4 + 12n\theta(1-\theta)]\,\|\boldsymbol{W} - \boldsymbol{W}_l\|_2 \ge \delta\bigg) \quad \text{(Lemma 38, 39)}\\
<&\mathbb{P}\bigg(\frac{1}{np}\Big|\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{X}\|_4^4\Big| + 5npB^4\varepsilon \ge \delta\bigg) \quad (p, B \text{ are large numbers, } \|\boldsymbol{W} - \boldsymbol{W}_l\|_2 \le \varepsilon)\\
\le&\mathbb{P}\bigg(\frac{1}{np}\Big|\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{X}\|_4^4\Big| + \frac{\delta}{2} \ge \delta\bigg) \quad \Big(\text{We assume } \varepsilon \le \frac{\delta}{10npB^4}\Big)\\
=&\mathbb{P}\bigg(\frac{1}{np}\Big|\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{X}\|_4^4\Big| \ge \frac{\delta}{2}\bigg).
\end{aligned} \tag{55}$$

Analysis. For random variable $\bar{\boldsymbol{X}}$, we have:

$$\Big|\mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{X}\|_4^4 - \mathbb{E}\,\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4\Big| = \Big|\mathbb{E}\big[\,\|\boldsymbol{W}_l\boldsymbol{X}\|_4^4 \cdot \mathbb{1}_{\{\|\boldsymbol{X}\|_\infty > B\}}\big]\Big| \le \sqrt{\mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{X}\|_4^8}\sqrt{\mathbb{E}\mathbb{1}_{\{\|\boldsymbol{X}\|_\infty > B\}}}, \tag{56}$$

moreover, if we view $\|\boldsymbol{W}_l\boldsymbol{X}\|_4^4 = \sum_{j=1}^p \|\boldsymbol{W}_l\boldsymbol{x}_j\|_4^4$ as sum of p independent random variables, we have:

$$\begin{aligned}
&\mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{X}\|_4^8\\
=&\mathbb{E}\bigg(\Big[\sum_{j=1}^p \|\boldsymbol{W}_l\boldsymbol{x}_j\|_4^4\Big]^2\bigg) = \mathbb{E}\bigg(\sum_{j=1}^p \|\boldsymbol{W}_l\boldsymbol{x}_j\|_4^8\bigg) + 2\mathbb{E}\bigg(\sum_{1\le j_1<j_2\le p} \|\boldsymbol{W}_l\boldsymbol{x}_{j_1}\|_4^4\,\|\boldsymbol{W}_l\boldsymbol{x}_{j_2}\|_4^4\bigg)\\
=&\mathbb{E}\bigg(\sum_{j=1}^p \|\boldsymbol{W}_l\boldsymbol{x}_j\|_4^8\bigg) + 2\sum_{1\le j_1<j_2\le p} \mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{x}_{j_1}\|_4^4\,\mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{x}_{j_2}\|_4^4 \quad (\boldsymbol{x}_j \text{ are independent})\\
=&\sum_{j=1}^p \mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{x}_j\|_4^8 + p(p-1)\Big[3\theta(1-\theta)\,\|\boldsymbol{W}_l\|_4^4 + 3\theta^2 n\Big]^2 \quad \text{(By Lemma 3)}\\
\le&Cpn^2\theta + 9p(p-1)n^2\theta^2 \quad \text{(By (194) of Lemma 42)}\\
\le&C_1p^2n^2\theta^2
\end{aligned} \tag{57}$$

for some constants $C_1 > 9$, for sufficiently large n, p. Also, by Lemma 34, we have:

$$\mathbb{E}\mathbb{1}_{\{\|\boldsymbol{X}\|_\infty > B\}} \le 2np\theta e^{-\frac{B^2}{2}}. \tag{58}$$

Substitute the results of previous two inequalities (57), (58) into (56), yields

$$\Big|\mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{X}\|_4^4 - \mathbb{E}\,\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4\Big| \le \sqrt{C_1}pn\theta\cdot\sqrt{2np\theta\exp\big(-B^2/2\big)} = C_2 n^{\frac{3}{2}}p^{\frac{3}{2}}\theta^{\frac{3}{2}}e^{-\frac{B^2}{4}}, \tag{59}$$

for some constants $C_2 > 3\sqrt{2}$. Hence, when

$$B > 2\sqrt{\ln\Big(\frac{C_3 n^{\frac{1}{2}}p^{\frac{1}{2}}\theta^{\frac{3}{2}}}{\delta}\Big)}, \tag{60}$$

for some constants $C_3 > 12\sqrt{2}$, we have:

$$\frac{1}{np}\Big|\mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{X}\|_4^4 - \mathbb{E}\,\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4\Big| \le C_2 n^{\frac{1}{2}}p^{\frac{1}{2}}\theta^{\frac{3}{2}}e^{-\frac{B^2}{4}} < \frac{\delta}{4}. \tag{61}$$

Therefore, combine (55), we have:

$$\begin{aligned}
&\mathbb{P}\bigg(\frac{1}{np}\Big|\big\|\boldsymbol{W}\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}\boldsymbol{X}\|_4^4\Big| \ge \delta\bigg) < \mathbb{P}\bigg(\frac{1}{np}\Big|\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{X}\|_4^4\Big| \ge \frac{\delta}{2}\bigg)\\
=&\mathbb{P}\bigg(\frac{1}{np}\Big|\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 + \mathbb{E}\,\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{X}\|_4^4\Big| \ge \frac{\delta}{2}\bigg)\\
\le&\mathbb{P}\bigg(\frac{1}{np}\Big|\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4\Big| + \frac{1}{np}\Big|\mathbb{E}\,\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{X}\|_4^4\Big| \ge \frac{\delta}{2}\bigg)\\
\le&\mathbb{P}\bigg(\frac{1}{np}\Big|\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4\Big| \ge \frac{\delta}{4}\bigg) \quad \text{By (61)}\\
=&\underbrace{\mathbb{P}\bigg(\frac{1}{np}\Big(\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4\Big) \ge \frac{\delta}{4}\bigg)}_{\Gamma_1} + \underbrace{\mathbb{P}\bigg(\frac{1}{np}\Big(\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4\Big) \le -\frac{\delta}{4}\bigg)}_{\Gamma_2}.
\end{aligned} \tag{62}$$

Point-wise Bernstein Inequality. Next, we will apply Bernstein's inequality on $\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4$ to bound its upper (Γ_1) and lower tail (Γ_2). Note that we can view $\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4$ as sum of p independent variables $\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 = \sum_{j=1}^p \|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_4^4$, and each of them is bounded by

$$\|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_4^4 = \|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_2^4\cdot\bigg\|\frac{\boldsymbol{W}_l\bar{\boldsymbol{x}}_j}{\|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_2}\bigg\|_4^4 \le \|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_2^4 = \|\bar{\boldsymbol{x}}_j\|_2^4 \le n^2B^4. \tag{63}$$

Also, in order to bound $\mathbb{E}\,\|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_4^8$, we consider

$$\mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{x}_j\|_4^8 - \mathbb{E}\,\|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_4^8 = \mathbb{E}\big(\,\|\boldsymbol{W}_l\boldsymbol{x}_j\|_4^8 - \|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_4^8\big) = \mathbb{E}\big(\,\|\boldsymbol{W}_l\boldsymbol{x}_j\|_4^8\cdot\mathbb{1}_{\{\|\boldsymbol{X}\|_\infty > B\}}\big) \ge 0, \tag{64}$$

along with (194) in Lemma 42, this implies

$$\mathbb{E}\,\|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_4^8 \le \mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{x}_j\|_4^8 \le Cn^2\theta, \tag{65}$$

where $C > 105$ is the same constant as (57). Now we apply Bernstein's inequality on Γ_1:

$$\begin{aligned}
\Gamma_1 =&\mathbb{P}\bigg(\frac{1}{np}\Big(\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4\Big) \ge \frac{\delta}{4}\bigg)\\
=&\mathbb{P}\bigg(\sum_{j=1}^{p}\Big(\|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_4^4\Big) \ge p\cdot\frac{n\delta}{4}\bigg)\\
\le&\exp\bigg(-\frac{pn^2\delta^2/16}{2\big(\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\,\|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_4^8 + n^2B^4\cdot n\delta/12\big)}\bigg)\\
=&\exp\bigg(-\frac{3pn^2\delta^2}{\frac{96}{p}\sum_{j=1}^{p}\mathbb{E}\,\|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_4^8 + 8n^3B^4\delta}\bigg)\\
\le&\exp\bigg(-\frac{3pn^2\delta^2}{\frac{96}{p}\sum_{j=1}^{p}\mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{x}_j\|_4^8 + 8n^3B^4\delta}\bigg)\\
\le&\exp\bigg(-\frac{3pn^2\delta^2}{c_1n^2\theta + 8n^3B^4\delta}\bigg) = \exp\bigg(-\frac{3p\delta^2}{c_1\theta + 8nB^4\delta}\bigg),
\end{aligned} \tag{66}$$

for a constant $c_1 > 10^4$. Next, we apply Bernstein's inequality on Γ_2, along with result in (57). Note that $\forall j \in [p]$, $\|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_4^4$ is lower bounded by 0, hence we have:

$$\begin{aligned}
\Gamma_2 =&\mathbb{P}\bigg(\frac{1}{np}\Big(\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\big\|\boldsymbol{W}_l\bar{\boldsymbol{X}}\big\|_4^4\Big) \le -\frac{\delta}{4}\bigg)\\
=&\mathbb{P}\bigg(\sum_{j=1}^{p}\Big(\|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_4^4\Big) \le -p\cdot\frac{n\delta}{4}\bigg)\\
\le&\exp\bigg(-\frac{pn^2\delta^2/16}{\frac{2}{p}\sum_{j=1}^{p}\mathbb{E}\,\|\boldsymbol{W}_l\bar{\boldsymbol{x}}_j\|_4^8}\bigg) \le \exp\bigg(-\frac{pn^2\delta^2/16}{\frac{2}{p}\sum_{j=1}^{p}\mathbb{E}\,\|\boldsymbol{W}_l\boldsymbol{x}_j\|_4^8}\bigg)\\
\le&\exp\bigg(-\frac{pn^2\delta^2}{32Cn^2\theta}\bigg) \le \exp\bigg(-\frac{p\delta^2}{c_2\theta}\bigg),
\end{aligned} \tag{67}$$

where $C > 105$ is the same constant as (57) and $c_2 > 3360$ is another constant. Replacing (66) and (67) into (62), yields

$$\mathbb{P}\bigg(\frac{1}{np}\Big|\big\|\boldsymbol{W}\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}\boldsymbol{X}\|_4^4\Big| \ge \delta\bigg) \le \Gamma_1 + \Gamma_2 \le \exp\bigg(-\frac{3p\delta^2}{c_1\theta + 8nB^4\delta}\bigg) + \exp\bigg(-\frac{p\delta^2}{c_2\theta}\bigg) \tag{68}$$

for some constants $c_1 > 10^4, c_2 > 3360$.

Union Bound. Now, we will give a union bound for

$$\mathbb{P}\bigg(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})}\frac{1}{np}\Big|\big\|\boldsymbol{W}\bar{\boldsymbol{X}}\big\|_4^4 - \mathbb{E}\,\|\boldsymbol{W}\boldsymbol{X}\|_4^4\Big| \ge \delta\bigg). \tag{69}$$

Notice that

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{W}\in \mathsf{O}(n;\mathbb{R})} \frac{1}{np}\left| \left\|\boldsymbol{W}\bar{\boldsymbol{X}}\right\|_4^4 - \mathbb{E}\left\|\boldsymbol{W}\boldsymbol{X}\right\|_4^4 \right| \geq \delta \right) \\
\leq &\sum_{l=1}^{|\mathcal{S}_\varepsilon|} \mathbb{P}\left(\sup_{\boldsymbol{W}\in \mathbb{B}(\boldsymbol{W}_l,\varepsilon)} \frac{1}{np}\left| \left\|\boldsymbol{W}\bar{\boldsymbol{X}}\right\|_4^4 - \mathbb{E}\left\|\boldsymbol{W}\boldsymbol{X}\right\|_4^4 \right| \geq \delta \right) \quad \text{(By } \varepsilon\text{−covering in (53))} \\
\leq &\sum_{l=1}^{|\mathcal{S}_\varepsilon|} \left[\exp\left(-\frac{3p\delta^2}{c_1\theta + 8nB^4\delta} \right) + \exp\left(-\frac{p\delta^2}{c_2\theta} \right) \right] \quad \text{(By (68) when } \varepsilon \leq \frac{\delta}{10npB^4}\text{)} \\
\leq &\left(\frac{6}{\varepsilon}\right)^{n^2} \left[\exp\left(-\frac{3p\delta^2}{c_1\theta + 8nB^4\delta} \right) + \exp\left(-\frac{p\delta^2}{c_2\theta} \right) \right] \quad \text{(By (54))} \\
= &\exp\left(n^2 \ln\left(\frac{60npB^4}{\delta}\right) \right) \left[\exp\left(-\frac{3p\delta^2}{c_1\theta + 8nB^4\delta} \right) + \exp\left(-\frac{p\delta^2}{c_2\theta} \right) \right] \quad \text{(Let } \varepsilon = \frac{\delta}{10npB^4}\text{)} \\
= &\exp\left(-\frac{3p\delta^2}{c_1\theta + 8nB^4\delta} + n^2 \ln\left(\frac{60npB^4}{\delta}\right) \right) + \exp\left(-\frac{p\delta^2}{c_2\theta} + n^2 \ln\left(\frac{60npB^4}{\delta}\right) \right).
\end{aligned}
\tag{70}
$$

Note that (60) requires a lower bound on B, here we can choose $B = \ln p$, which satisfies (60) when p is large enough (say $p = \Omega(n)$). Combine (50) and substitute $B = \ln p$ into (70), we have:

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{W}\in \mathsf{O}(n;\mathbb{R})} \frac{1}{np}\left| \left\|\boldsymbol{W}\boldsymbol{X}\right\|_4^4 - \mathbb{E}\left\|\boldsymbol{W}\boldsymbol{X}\right\|_4^4 \right| \geq \delta \right) \\
\leq &\mathbb{P}\left(\sup_{\boldsymbol{W}\in \mathsf{O}(n;\mathbb{R})} \frac{1}{np}\left| \left\|\boldsymbol{W}\bar{\boldsymbol{X}}\right\|_4^4 - \mathbb{E}\left\|\boldsymbol{W}\boldsymbol{X}\right\|_4^4 \right| \geq \delta \right) + 2np\theta e^{-\frac{B^2}{2}} \\
\leq &\exp\left(-\frac{3p\delta^2}{c_1\theta + 8n(\ln p)^4\delta} + n^2 \ln\left(\frac{60np(\ln p)^4}{\delta}\right) \right) \\
&+ \exp\left(-\frac{p\delta^2}{c_2\theta} + n^2 \ln\left(\frac{60np(\ln p)^4}{\delta}\right) \right) + 2np\theta \exp\left(-\frac{(\ln p)^2}{2} \right),
\end{aligned}
\tag{71}
$$

for some constants $c_1 > 10^4, c_2 > 3360$, which completes the proof for (47). When $p = \Omega(\theta n^2 \ln n/\delta^2)$, we have:

$$
\begin{aligned}
&\exp\left(-\frac{3p\delta^2}{c_1\theta + 8n(\ln p)^4\delta} + n^2 \ln\left(\frac{60np(\ln p)^4}{\delta}\right) \right) \\
&+ \exp\left(-\frac{p\delta^2}{c_2\theta} + n^2 \ln\left(\frac{60np(\ln p)^4}{\delta}\right) \right) + 2np\theta \exp\left(-\frac{(\ln p)^2}{2} \right) \leq \frac{1}{3p} + \frac{1}{3p} + \frac{1}{3p} = \frac{1}{p},
\end{aligned}
\tag{72}
$$

which completes the proof for (48). ■

A.3. Proof of Lemma 5

Claim 20 (Extrema of ℓ^4-Norm over Orthogonal Group) *For any orthogonal matrix $\boldsymbol{A} \in \mathsf{O}(n;\mathbb{R})$, $g(\boldsymbol{A}) = \|\boldsymbol{A}\|_4^4 \in [1,n]$, $g(\boldsymbol{A})$ reaches maximum if and only if $\boldsymbol{A} \in SP(n)$.*

Proof For the maximum of $g(\boldsymbol{A})$

$$g(\boldsymbol{A}) = \sum_{i=1}^n \sum_{j=1}^n a_{i,j}^4 \le \sum_{i=1}^n \sum_{j=1}^n a_{i,j}^4 + 2\sum_{i=1}^n \sum_{1\le j_1<j_2\le n} a_{i,j_1}^2 a_{i,j_2}^2 = \sum_{i=1}^n \Big(\sum_{j=1}^n a_{i,j}^2\Big)^2 = n. \quad (73)$$

And when equality holds, we have

$$a_{i,j_1} a_{i,j_2} = 0, \quad \forall i, j_1 \neq, j_2 \in [n], \quad (74)$$

which implies $\boldsymbol{A} \in \mathrm{SP}(n)$. ∎

A.4. Proof of Lemma 6

Claim 21 (Approximate Maxima of ℓ^4-Norm over the Orthogonal Group) *Suppose $\boldsymbol{W}$ is an orthogonal matrix: $\boldsymbol{W} \in \mathsf{O}(n;\mathbb{R})$. $\forall \varepsilon \in [0,1]$, if $\frac{1}{n}\|\boldsymbol{W}\|_4^4 \ge 1-\varepsilon$, then $\exists \boldsymbol{P} \in SP(n)$, such that*

$$\frac{1}{n}\|\boldsymbol{W} - \boldsymbol{P}\|_F^2 \le 2\varepsilon. \quad (75)$$

Proof Note that the condition $\frac{1}{n}\|\boldsymbol{W}\|_4^4 > 1-\varepsilon$ can be viewed as

$$\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{w}_i\|_4^4 \ge 1-\varepsilon, \quad (76)$$

where each column vector $\boldsymbol{w}_i$ of $\boldsymbol{W}$ satisfies $\boldsymbol{w}_i \in \mathbb{S}^{n-1}$. By Lemma 37, we we know there exists $j_1, j_2, \dots j_n \in [n]$, such that

$$\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{w}_i - s_i \boldsymbol{e}_{j_i}\|_2^2 \le 2\varepsilon, \quad (77)$$

where $\boldsymbol{e}_{j_i}$ are one vector in canonical basis and $s_i \in \{1,-1\}$ indicates the sign of $\boldsymbol{e}_{j_i}$, $\forall i \in [n]$. Since $\boldsymbol{W} \in \mathsf{O}(n;\mathbb{R})$, one can easily show that $\forall i_1 \neq i_2, i_1, i_2 \in [n]$, the canonical vector they are corresponding to $\boldsymbol{e}_{j_{i_1}}, \boldsymbol{e}_{j_{i_2}}$ are different, that is, $j_{i_1} \neq j_{i_2}$ (otherwise suppose $\exists i_1 \neq i_2$, such that $\boldsymbol{w}_{i_1}$ and $\boldsymbol{w}_{i_2}$ correspond to the same canonical vector $\boldsymbol{e}_j$ in (77), one can easily show that $\boldsymbol{w}_{i_1}^* \boldsymbol{w}_{i_2} \neq 0$, which contradicts with the orthogonality of $\boldsymbol{W}$). Hence, there exists a sign permutation matrix:

$$\boldsymbol{P} = [sign(w_{j_1,1})\boldsymbol{e}_{j_1}, sign(w_{j_2,2})\boldsymbol{e}_{j_2}, \dots, sign(w_{j_n,n})\boldsymbol{e}_{j_n}], \quad (78)$$

such that

$$\|\boldsymbol{W} - \boldsymbol{P}\|_F^2 = \sum_{i=1}^n \|\boldsymbol{w}_i - sign(w_{j_i,1})\boldsymbol{e}_{j_i}\|_2^2 \le 2n\varepsilon \implies \frac{1}{n}\|\boldsymbol{W} - \boldsymbol{P}\|_F^2 \le 2\varepsilon, \quad (79)$$

which completes the proof. ∎

A.5. Proof of Theorem 7

Claim 22 (Correctness of Global Maxima) *$\forall \theta \in (0,1)$, assume $\boldsymbol{X}_o = \{x_{i,j}\} \in \mathbb{R}^{n\times p}$ is a Bernoulli-Gaussian matrix, $\boldsymbol{D}_o \in \mathsf{O}(n;\mathbb{R})$ is any orthogonal matrix, and $\boldsymbol{Y} = \boldsymbol{D}_o\boldsymbol{X}_o$. Suppose $\hat{\boldsymbol{A}}_\star$ is a global maximizer of the optimization problem*

$$\max_{\boldsymbol{A}} \hat{f}(\boldsymbol{A},\boldsymbol{Y}) = \|\boldsymbol{A}\boldsymbol{Y}\|_4^4, \quad \text{subject to} \quad \boldsymbol{A} \in \mathsf{O}(n;\mathbb{R}),$$

then for any $\varepsilon \in [0,1]$, there exists a signed permutation matrix $\boldsymbol{P} \in SP(n)$, such that

$$\frac{1}{n}\left\|\hat{\boldsymbol{A}}_\star^* - \boldsymbol{D}_o\boldsymbol{P}\right\|_F^2 \le C\varepsilon, \tag{80}$$

with probability at least

$$\begin{aligned} &1-\exp\left(-\frac{3p\varepsilon^2}{c_1\theta+8n(\ln p)^4\varepsilon}+n^2\ln\Big(\frac{60np(\ln p)^4}{\varepsilon}\Big)\right)\\ &-\exp\left(-\frac{p\varepsilon^2}{c_2\theta}+n^2\ln\Big(\frac{60np(\ln p)^4}{\varepsilon}\Big)\right)-2np\theta\exp\left(-\frac{(\ln p)^2}{2}\right), \end{aligned} \tag{81}$$

for some constants $c_1 > 10^4, c_2 > 3360, C > \frac{4}{3\theta(1-\theta)}$. Moreover

$$\begin{aligned} &1-\exp\left(-\frac{3p\varepsilon^2}{c_1\theta+8n(\ln p)^4\varepsilon}+n^2\ln\Big(\frac{60np(\ln p)^4}{\varepsilon}\Big)\right)\\ &-\exp\left(-\frac{p\varepsilon^2}{c_2\theta}+n^2\ln\Big(\frac{60np(\ln p)^4}{\varepsilon}\Big)\right)-2np\theta\exp\left(-\frac{(\ln p)^2}{2}\right)\ge 1-\frac{1}{p}, \end{aligned} \tag{82}$$

when $p = \Omega(\theta n^2 \ln n/\varepsilon^2)$.

Proof Suppose $\boldsymbol{A}_\star$ is the global maximizer of optimization program (19):

$$\max_{\boldsymbol{A}} f(\boldsymbol{A}) = \mathbb{E}\,\|\boldsymbol{A}\boldsymbol{Y}\|_4^4 \quad \text{subject to } \boldsymbol{A} \in \mathsf{O}(n;\mathbb{R}),$$

then by (47) and (48) in Lemma 4, when $p = \Omega(\theta n^2 \ln n/\varepsilon^2)$, we know that with probability at least

$$\begin{aligned} &1-\exp\left(-\frac{3p\varepsilon^2}{c_1\theta+8n(\ln p)^4\varepsilon}+n^2\ln\Big(\frac{60np(\ln p)^4}{\varepsilon}\Big)\right)\\ &-\exp\left(-\frac{p\varepsilon^2}{c_2\theta}+n^2\ln\Big(\frac{60np(\ln p)^4}{\varepsilon}\Big)\right)-2np\theta\exp\left(-\frac{(\ln p)^2}{2}\right)\ge 1-\frac{1}{p}, \end{aligned} \tag{83}$$

we have

$$\frac{1}{np}\left|\hat{f}(\hat{\boldsymbol{A}}_\star,\boldsymbol{Y}) - f(\hat{\boldsymbol{A}}_\star)\right| \le \varepsilon \quad \text{and} \quad \frac{1}{np}\left|\hat{f}(\boldsymbol{A}_\star,\boldsymbol{Y}) - f(\boldsymbol{A}_\star)\right| \le \varepsilon. \tag{84}$$

which implies

$$\frac{1}{np}f(\hat{\boldsymbol{A}}_\star) \le \frac{1}{np}f(\boldsymbol{A}_\star) < \frac{1}{np}\hat{f}(\boldsymbol{A}_\star,\boldsymbol{Y}) + \varepsilon \le \frac{1}{np}\hat{f}(\hat{\boldsymbol{A}}_\star,\boldsymbol{Y}) + \varepsilon < \frac{1}{np}f(\hat{\boldsymbol{A}}_\star) + 2\varepsilon. \tag{85}$$

In the above inequality, the first and the third $\leq$ is due to the global optimality of $f(\boldsymbol{A}_\star)$ and $\hat{f}(\hat{\boldsymbol{A}}_\star, \boldsymbol{Y})$, the second and the last $<$ is due to (84). Simplify (85), yields

$$\frac{1}{np} f(\hat{\boldsymbol{A}}_\star) \in \left(\frac{1}{np} f(\boldsymbol{A}_\star) - 2\varepsilon, \frac{1}{np} f(\boldsymbol{A}_\star) \right). \tag{86}$$

From Lemma 3, we have:

$$\frac{1}{3p\theta} f(\boldsymbol{A}) = (1-\theta) g(\boldsymbol{A}\boldsymbol{D}_o) + \theta n, \quad \forall \boldsymbol{A} \in \mathsf{O}(n;\mathbb{R}),$$

which implies

$$\begin{aligned} &\frac{3\theta(1-\theta)}{n} g(\hat{\boldsymbol{A}}_\star \boldsymbol{D}_o) \in \left(\frac{3\theta(1-\theta)}{n} g(\boldsymbol{A}_\star \boldsymbol{D}_o) - 2\varepsilon, \frac{3\theta(1-\theta)}{n} g(\boldsymbol{A}_\star \boldsymbol{D}_o) \right] \\ \Longrightarrow\ &\frac{1}{n} \left\| \hat{\boldsymbol{A}}_\star \boldsymbol{D}_o \right\|_4^4 \in \left(\frac{1}{n} \left\| \boldsymbol{A}_\star \boldsymbol{D}_o \right\|_4^4 - \frac{2\varepsilon}{3\theta(1-\theta)}, \frac{1}{n} \left\| \boldsymbol{A}_\star \boldsymbol{D}_o \right\|_4^4 \right]. \end{aligned} \tag{87}$$

Lemma 3 tells us that $\boldsymbol{A}_\star \boldsymbol{D}_o \in \mathrm{SP}(n)$, combining Lemma 5, we know that $\|\boldsymbol{A}_\star \boldsymbol{D}_o\|_4^4 = n$. Thus we can further simplify (87) as

$$\frac{1}{n} \left\| \hat{\boldsymbol{A}}_\star \boldsymbol{D}_o \right\|_4^4 \in \left(1 - \frac{2\varepsilon}{3\theta(1-\theta)}, 1 \right].$$

Applying Lemma 6 (change ε in Lemma 6 into $2\varepsilon/3\theta(1-\theta)$), we know that there exists $\boldsymbol{P} \in \mathrm{SP}(n)$, such that

$$\frac{1}{n} \left\| \hat{\boldsymbol{A}}_\star \boldsymbol{D}_o - \boldsymbol{P} \right\|_F^2 \leq \frac{4\varepsilon}{3\theta(1-\theta)}. \tag{88}$$

By the rotational invariant of Frobenius norm, we have

$$\frac{1}{n} \left\| \hat{\boldsymbol{A}}_\star^* - \boldsymbol{D}_o \boldsymbol{P}^* \right\|_F^2 \leq \frac{4\varepsilon}{3\theta(1-\theta)}, \tag{89}$$

which completes the proof. ■

B. Proofs of Section 3

B.1. Proof of Lemma 9

Claim 23 (Projection onto Orthogonal Group) *$\forall \boldsymbol{A} \in \mathbb{R}^{n\times n}$, the orthogonal matrix which has minimum Frobenius norm with $\boldsymbol{A}$ is the following*

$$\mathcal{P}_{\mathsf{O}(n;\mathbb{R})}(\boldsymbol{A}) = \operatorname*{arg\,min}_{\boldsymbol{M} \in \mathsf{O}(n;\mathbb{R})} \|\boldsymbol{M} - \boldsymbol{A}\|_F^2 = \boldsymbol{U}\boldsymbol{V}^*, \tag{90}$$

where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = SVD(\boldsymbol{A})$.

Proof Notice that

$$\begin{aligned}\|\boldsymbol{M}-\boldsymbol{A}\|_F^2 =& tr\big((\boldsymbol{M}-\boldsymbol{A})(\boldsymbol{M}-\boldsymbol{A})^*\big)\\ =& tr\big(\boldsymbol{I}-\boldsymbol{A}\boldsymbol{M}^*-\boldsymbol{M}\boldsymbol{A}^*+\boldsymbol{A}\boldsymbol{A}^*\big)=n-2tr\big(\boldsymbol{M}\boldsymbol{A}^*\big)+tr\big(\boldsymbol{A}\boldsymbol{A}^*\big).\end{aligned} \tag{91}$$

Since $tr\big(\boldsymbol{A}\boldsymbol{A}^*\big)$ is a constant, we know that

$$\operatorname*{arg\,min}_{\boldsymbol{M}\in\mathsf{O}(n;\mathbb{R})}\|\boldsymbol{M}-\boldsymbol{A}\|_F^2=\operatorname*{arg\,max}_{\boldsymbol{M}\in\mathsf{O}(n;\mathbb{R})}tr\big(\boldsymbol{A}\boldsymbol{M}^*\big). \tag{92}$$

Let $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ be the SVD of $\boldsymbol{A}$, then

$$tr\big(\boldsymbol{A}\boldsymbol{M}^*\big)=tr\big(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*\boldsymbol{M}\big)=tr\big(\boldsymbol{\Sigma}\boldsymbol{V}^*\boldsymbol{M}^*\boldsymbol{U}\big)\le\sum_{i=1}^n\sigma_i(\boldsymbol{A})\sigma_i(\boldsymbol{V}^*\boldsymbol{M}^*\boldsymbol{U})=\sum_{i=1}^n\sigma_i(\boldsymbol{A}), \tag{93}$$

where inequality is obtained through Von Neumann's trace inequality, and the equality holds if and only if $\boldsymbol{V}^*\boldsymbol{M}^*\boldsymbol{U}$ is diagonal matrix (in fact, identity matrix), which implies $\boldsymbol{V}^*\boldsymbol{M}^*\boldsymbol{U}=\boldsymbol{I}\implies\boldsymbol{M}=\boldsymbol{U}\boldsymbol{V}^*$. ■

B.2. Proof of Proposition 10

Claim 24 (Expectation of $\nabla_{\boldsymbol{A}}\hat{f}(\boldsymbol{A},\boldsymbol{Y})$) *Let $\boldsymbol{X}\in\mathbb{R}^{n\times p}$, $x_{i,j}\sim_{iid}BG(\theta)$, $\boldsymbol{D}_o\in\mathsf{O}(n;\mathbb{R})$ is any orthogonal matrix, and $\boldsymbol{Y}=\boldsymbol{D}_o\boldsymbol{X}_o$. The expectation of $\nabla_{\boldsymbol{A}}\hat{f}(\boldsymbol{A},\boldsymbol{Y})$ satisfies this property:*

$$\mathbb{E}_{\boldsymbol{X}_o}\nabla_{\boldsymbol{A}}\hat{f}(\boldsymbol{A},\boldsymbol{Y})=3p\theta(1-\theta)\nabla_{\boldsymbol{A}}g(\boldsymbol{A}\boldsymbol{D}_o)+12p\theta^2\boldsymbol{A}. \tag{94}$$

Proof For simplicity, since $\boldsymbol{D}_o\in\mathsf{O}(n;\mathbb{R})$, let $\boldsymbol{W}=\boldsymbol{A}\boldsymbol{D}_o$, we know that $\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})$. By the fact that $\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})$, we have:

$$\frac{1}{4}\nabla_{\boldsymbol{A}}\hat{f}(\boldsymbol{A},\boldsymbol{Y})=(\boldsymbol{A}\boldsymbol{Y})^{\circ 3}\boldsymbol{Y}^*=(\boldsymbol{W}\boldsymbol{X}_o)^{\circ 3}\boldsymbol{X}_o^*\boldsymbol{D}_o^*, \tag{95}$$

moreover,

$$\begin{aligned}\{(\boldsymbol{A}\boldsymbol{Y})^{\circ 3}\}_{i,j}=&\{(\boldsymbol{W}\boldsymbol{X}_o)^{\circ 3}\}_{i,j}=\Big(\sum_{k=1}^n w_{i,k}x_{k,j}\Big)^3\\ =&\sum_{k=1}^n w_{i,k}^3x_{k,j}^3+3\Big(\sum_{1\le k_1<k_2\le n}w_{i,k_1}^2x_{k_1,j}^2w_{i,k_2}x_{k_2,j}+w_{i,k_1}x_{k_1,j}w_{i,k_2}^2x_{k_2,j}^2\Big)\\ &+6\Big(\sum_{1\le k_1<k_2<k_3\le n}w_{i,k_1}x_{i,k_1}w_{i,k_2}x_{i,k_2}w_{i,k_3}x_{i,k_3}\Big).\end{aligned} \tag{96}$$

And hence, we know:

$$\begin{aligned}&\{(\boldsymbol{W}\boldsymbol{X}_o)^{\circ 3}\boldsymbol{X}_o^*\}_{i,j'}\\ =&\sum_{j=1}^p\Bigg[x_{j',j}\sum_{k=1}^n w_{i,k}^3x_{k,j}^3+3x_{j',j}\Big(\sum_{1\le k_1<k_2\le n}w_{i,k_1}^2x_{k_1,j}^2w_{i,k_2}x_{k_2,j}+w_{i,k_1}x_{k_1,j}w_{i,k_2}^2x_{k_2,j}^2\Big)\\ &+6x_{j',j}\Big(\sum_{1\le k_1<k_2<k_3\le n}w_{i,k_1}x_{k_1,j}w_{i,k_2}x_{k_2,j}w_{i,k_3}x_{k_3,j}\Big)\Bigg].\end{aligned} \tag{97}$$

Thus,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}_o}\{(\boldsymbol{W}\boldsymbol{X}_o)^{\circ 3}\boldsymbol{X}_o^*\}_{i,j'} &= 3p\theta w_{i,j'}^3 + 3p\theta^2 \sum_{\substack{1\le j\le n\\ j\neq j'}} w_{i,j}^2 w_{i,j'} \\ &= 3p\theta w_{i,j'}^3 + 3p\theta^2(1-w_{i,j'}^2)w_{i,j'} = 3p\theta(1-\theta)w_{i,j'}^3 + 3p\theta^2 w_{i,j'},\end{aligned} \tag{98}$$

which implies

$$\begin{aligned}\frac{1}{4p}\mathbb{E}_{\boldsymbol{X}_o}\nabla_{\boldsymbol{A}}\hat{f}(\boldsymbol{A},\boldsymbol{Y}) &= \frac{1}{p}\mathbb{E}_{\boldsymbol{X}_o}(\boldsymbol{W}\boldsymbol{X}_o)^{\circ 3}\boldsymbol{X}_o^*\boldsymbol{D}_o^* = 3\theta(1-\theta)\boldsymbol{W}^{\circ 3}\boldsymbol{D}_o^* + 3\theta^2\boldsymbol{W}\boldsymbol{D}_o^* \\ &= 3\theta(1-\theta)(\boldsymbol{A}\boldsymbol{D}_o)^{\circ 3}\boldsymbol{D}_o^* + 3\theta^2\boldsymbol{A} = \frac{3}{4}\theta(1-\theta)\nabla_{\boldsymbol{A}} g(\boldsymbol{A}\boldsymbol{D}_o) + 3\theta^2\boldsymbol{A}.\end{aligned} \tag{99}$$

■

B.3. Proof of Proposition 11

Claim 25 (Union Tail Concentration Bound of $\frac{1}{np}\nabla\hat{f}(\cdot,\cdot)$) *If $\boldsymbol{X}_o \in \mathbb{R}^{n\times p}, x_{i,j} \sim_{iid} BG(\theta)$, for any $\boldsymbol{A},\boldsymbol{D}_o \in \mathsf{O}(n;\mathbb{R})$, and $\boldsymbol{Y} = \boldsymbol{D}_o\boldsymbol{X}_o$, the following inequality holds*

$$\begin{aligned}&\mathbb{P}\bigg(\sup_{\boldsymbol{A}\in\mathsf{O}(n;\mathbb{R})}\frac{1}{4np}\Big\|\nabla_{\boldsymbol{A}}\hat{f}(\boldsymbol{A},\boldsymbol{Y}) - \mathbb{E}\big[\nabla_{\boldsymbol{A}}\hat{f}(\boldsymbol{A},\boldsymbol{Y})\big]\Big\|_F \ge \delta\bigg) \\ \le& 2n^2\exp\bigg(-\frac{3p\delta^2}{c_1\theta + 8n^{\frac{3}{2}}(\ln p)^4\delta} + n^2\ln\Big(\frac{48np(\ln p)^4}{\delta}\Big)\bigg) + 2np\theta\exp\bigg(-\frac{(\ln p)^2}{2}\bigg),\end{aligned} \tag{100}$$

for a constant $c_1 > 1.7\times 10^4$. Moreover

$$2n^2\exp\bigg(-\frac{3p\delta^2}{c_1\theta + 8n^{\frac{3}{2}}(\ln p)^4\delta} + n^2\ln\Big(\frac{48np(\ln p)^4}{\delta}\Big)\bigg) + 2np\theta\exp\bigg(-\frac{(\ln p)^2}{2}\bigg) \le \frac{1}{p} \tag{101}$$

when $p = \Omega(\theta n^2\ln n/\delta^2)$.

Proof By Proposition 10, we have:

$$\mathbb{E}_{\boldsymbol{X}_o}\nabla_{\boldsymbol{A}}\hat{f}(\boldsymbol{A},\boldsymbol{Y}) = 3p\theta(1-\theta)\nabla_{\boldsymbol{A}} g(\boldsymbol{A}\boldsymbol{D}_o) + 12p\theta^2\boldsymbol{A}, \tag{102}$$

so

$$\begin{aligned}
&\mathbb{P}\bigg(\sup_{\boldsymbol{A}\in\mathsf{O}(n;\mathbb{R})}\frac{1}{4np}\Big\|\nabla_{\boldsymbol{A}}\hat{f}(\boldsymbol{A},\boldsymbol{Y})-\mathbb{E}\big[\nabla_{\boldsymbol{A}}\hat{f}(\boldsymbol{A},\boldsymbol{Y})\big]\Big\|_F\geq\delta\bigg)\\
=&\mathbb{P}\bigg(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})}\frac{1}{np}\big\|(\boldsymbol{W}\boldsymbol{X}_o)^{\circ 3}\boldsymbol{X}_o^*\boldsymbol{D}_o^*-\mathbb{E}\big[(\boldsymbol{W}\boldsymbol{X}_o)^{\circ 3}\boldsymbol{X}_o^*\boldsymbol{D}_o^*\big]\big\|_F\geq\delta\bigg)\quad(\text{Assume }\boldsymbol{W}=\boldsymbol{A}\boldsymbol{D}_o)\\
=&\mathbb{P}\bigg(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})}\frac{1}{np}\big\|(\boldsymbol{W}\boldsymbol{X}_o)^{\circ 3}\boldsymbol{X}_o^*-\mathbb{E}\big[(\boldsymbol{W}\boldsymbol{X}_o)^{\circ 3}\boldsymbol{X}_o^*\big]\big\|_F\geq\delta\bigg)\\
\leq&2n^2\exp\bigg(-\frac{3p\delta^2}{c_1\theta+8n^{\frac{3}{2}}(\ln p)^4\delta}+n^2\ln\Big(\frac{48np(\ln p)^4}{\delta}\Big)\bigg)\\
&+2np\theta\exp\bigg(-\frac{(\ln p)^2}{2}\bigg)\quad(\text{By Lemma 44}),
\end{aligned}\tag{103}$$

for a constant $c_1>1.7\times10^4$, which completes the proof for (100). When $p=\Omega(\theta n^2\ln n/\delta^2)$, we have

$$\begin{aligned}
&2n^2\exp\bigg(-\frac{3p\delta^2}{c_1\theta+8n^{\frac{3}{2}}(\ln p)^4\delta}+n^2\ln\Big(\frac{48np(\ln p)^4}{\delta}\Big)\bigg)+2np\theta\exp\bigg(-\frac{(\ln p)^2}{2}\bigg)\\
\leq&\frac{1}{2p}+\frac{1}{2p}=\frac{1}{p},
\end{aligned}\tag{104}$$

which completes the proof for (101). ■

C. Proofs of Section 4

C.1. Proof of Proposition 12

Claim 26 *The critical points of $g(\boldsymbol{W})$ on manifold $\mathsf{O}(n;\mathbb{R})$ satisfies the following condition*

$$(\boldsymbol{W}^{\circ 3})^*\boldsymbol{W}=\boldsymbol{W}^*\boldsymbol{W}^{\circ 3}.\tag{105}$$

Proof Notice that $\nabla_{\boldsymbol{W}}g(\boldsymbol{W})=4\boldsymbol{W}^{\circ 3}$, and the critical points of $g(\boldsymbol{W})$ on $\mathsf{O}(n;\mathbb{R})$ satisfies

$$\operatorname{grad}g(\boldsymbol{W})=\mathcal{P}_{T_{\boldsymbol{W}}\mathsf{O}(n;\mathbb{R})}(\nabla_{\boldsymbol{A}}g(\boldsymbol{A}))=\frac{1}{2}(4\boldsymbol{W}^{\circ 3}-\boldsymbol{W}(4\boldsymbol{W}^{\circ 3})^*\boldsymbol{W})=\boldsymbol{0},\tag{106}$$

which yields

$$(\boldsymbol{W}^{\circ 3})^*\boldsymbol{W}=\boldsymbol{W}^*\boldsymbol{W}^{\circ 3}.\tag{107}$$

Therefore, $\forall\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})$, we can write critical points condition of ℓ^4-norm over $\mathsf{O}(n;\mathbb{R})$ as the following equations

$$\begin{cases}(\boldsymbol{W}^{\circ 3})^*\boldsymbol{W}=\boldsymbol{W}^*\boldsymbol{W}^{\circ 3},\\ \boldsymbol{W}^*\boldsymbol{W}=\boldsymbol{I}.\end{cases}\tag{108}$$

■

C.2. Proof of Proposition 13

Claim 27 *All global maximizers of ℓ^4-norm over the orthogonal group are isolated critical points.*

Proof As our objective function is invariant under signed permutation group, without loss of generality, we prove for the identity matrix $\boldsymbol{I}$. Suppose that the are not isolated. Then, there exists a $\boldsymbol{W}_0$ such that $(\boldsymbol{W}_0^{\circ 3})^*\boldsymbol{W}_0 = \boldsymbol{W}_0^*\boldsymbol{W}_0^{\circ 3}$ and $\boldsymbol{W}_0^*\boldsymbol{W}_0 = \boldsymbol{I}$, and in every neighborhood of $\boldsymbol{W}_0$ there exists some $\boldsymbol{W}$ that is a critical point. This implies that there exists a path around $\boldsymbol{W}_0$ such that

$$\boldsymbol{W}(\cdot) : (-\varepsilon, \varepsilon) \to \mathsf{O}(n, \mathbb{R}),\ (\boldsymbol{W}(t)^{\circ 3})^*\boldsymbol{W}(t) = \boldsymbol{W}(t)^*\boldsymbol{W}(t)^{\circ 3},\ \boldsymbol{W}(0) = \boldsymbol{W}_0. \tag{109}$$

Then, we expand $\boldsymbol{W}(t)$ around $t = 0$,

$$\boldsymbol{W}(t) = \boldsymbol{W}_0 + t\boldsymbol{W}_1 + t^2\boldsymbol{W}_2 + \dots \tag{110}$$

Constraint $(\boldsymbol{W}(t)^{\circ 3})^*\boldsymbol{W}(t) = \boldsymbol{W}(t)^*\boldsymbol{W}(t)^{\circ 3}$ implies that

$$\begin{aligned} &(\boldsymbol{W}_0 + t\boldsymbol{W}_1)^*(\boldsymbol{W}_0 + t\boldsymbol{W}_1)^{\circ 3} = [(\boldsymbol{W}_0 + t\boldsymbol{W}_1)^{\circ 3}]^*(\boldsymbol{W}_0 + t\boldsymbol{W}_1) \\ \implies &3\boldsymbol{W}_1 + \boldsymbol{W}_1^* = 3\boldsymbol{W}_1^* + \boldsymbol{W}_1 \Rightarrow \boldsymbol{W}_1 = \boldsymbol{W}_1^*. \end{aligned} \tag{111}$$

Constraint $\boldsymbol{W}(t)^*\boldsymbol{W}(t) = \boldsymbol{I}$ implies that

$$\frac{d}{dt}\big[\boldsymbol{W}(t)^*\boldsymbol{W}(t)\big]\Big|_{t=0} = \frac{d}{dt}\boldsymbol{I}\Big|_{t=0} = \boldsymbol{0} \implies \boldsymbol{W}_1 + \boldsymbol{W}_1^* = \boldsymbol{0}. \tag{112}$$

The above computation is equivalent to plugging $\boldsymbol{W}(t)$ into the constraints, taking the derivative and evaluating at $\boldsymbol{0}$. Combining (111) and (112), we know $\boldsymbol{W}_1 = \boldsymbol{0}$. Since $\boldsymbol{W}(t)$ is any arbitrary path, this shows that the variety formed by all the critical points does not have a tangent space around $\boldsymbol{W}_0$. We conclude that the $\boldsymbol{I}$ is an isolated critical points. In fact, by calculating the Hessian at the maximizers, one can further show that all global maximizers, are nondegenerate critical points. ∎

C.3. Proof of Proposition 14

Claim 28 (Fixed Point of the MSP Algorithm) *Given $\boldsymbol{W} \in \mathsf{O}(n; \mathbb{R})$, $\boldsymbol{W}$ is a fix point of the MSP algorithm if and only if $\boldsymbol{W}$ is a critical point of the ℓ^4-norm over $\mathsf{O}(n; \mathbb{R})$.*

Proof Let $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \boldsymbol{W}^{\circ 3}$ denote the the SVD of $\boldsymbol{W}^{\circ 3}$.

- When $\boldsymbol{W} \in \mathsf{O}(n; \mathbb{R})$ is a fixed point of MSP algorithm, we have:

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{V}^* \implies (\boldsymbol{W}^{\circ 3})^*\boldsymbol{W} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^*\boldsymbol{W} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^*, \tag{113}$$

  which implies $(\boldsymbol{W}^{\circ 3})^*\boldsymbol{W}$ is symmetric, hence we have:

$$(\boldsymbol{W}^{\circ 3})^*\boldsymbol{W} = \boldsymbol{W}^*\boldsymbol{W}^{\circ 3}, \tag{114}$$

  and by Proposition 12, $\boldsymbol{W}$ is a critical point of ℓ^4-norm over the orthogonal group.

- When $\boldsymbol{W}$ is a critical point of ℓ^4-norm over orthogonal group, we have $(\boldsymbol{W}^{\circ 3})^*\boldsymbol{W} = \boldsymbol{W}^*\boldsymbol{W}^{\circ 3}$. So the SVD of $(\boldsymbol{W}^{\circ 3})^*\boldsymbol{W}$ should equal to the SVD of $\boldsymbol{W}^*\boldsymbol{W}^{\circ 3}$. Also by the rotational invariant of SVD, we know:

$$\mathrm{SVD}\big((\boldsymbol{W}^{\circ 3})^*\boldsymbol{W}\big) = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^*\boldsymbol{W} = \mathrm{SVD}\big(\boldsymbol{W}^*\boldsymbol{W}^{\circ 3}\big) = \boldsymbol{W}^*\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*, \tag{115}$$

  and by the uniqueness of SVD, we have:

$$\boldsymbol{V} = \boldsymbol{W}^*\boldsymbol{U} \implies \boldsymbol{W} = \boldsymbol{U}\boldsymbol{V}^*. \tag{116}$$

■

C.4. Proof of Proposition 15

Claim 29 (Convergence of PGA with Arbitrary Step Size) *Iterative PGA Algorithm 3 with any fixed step size $\alpha > 0$ (α can be $+\infty$ and PGA is equivalent to MSP Algorithm 1 when $\alpha = +\infty$) finds a saddle point of optimization problem* (34)

$$\max_{\boldsymbol{A}\in\mathsf{O}(n;\mathbb{R})} \|\boldsymbol{A}\|_4^4 .$$

Proof Consider the following objective function $h(\cdot) : \mathsf{O}(n;\mathbb{R}) \mapsto \mathbb{R}^+$:

$$h(\boldsymbol{A}) = \begin{cases} \frac{\alpha}{4}\|\boldsymbol{A}\|_4^4 + \frac{1}{2}\|\boldsymbol{A}\|_F^2 & \text{when } \alpha < \infty \\ \|\boldsymbol{A}\|_4^4 & \text{when } \alpha = +\infty \end{cases}. \tag{117}$$

Note that $h(\boldsymbol{A})$ is convex in both cases ($\alpha < +\infty$ or $\alpha = \infty$). Also note that the Stiefel manifold $\mathsf{O}(n;\mathbb{R})$ is a compact manifold, so by Theorem 1 in Journée et al. (2010), we know that the following iterative update

$$\boldsymbol{A}_{k+1} \in \arg\max\{h(\boldsymbol{A}_k) + \langle \partial h(\boldsymbol{A}_k), \boldsymbol{W} - \boldsymbol{A}_k\rangle \,|\boldsymbol{W} \in \mathsf{O}(n;\mathbb{R})\} \tag{118}$$

will find a saddle point of $h(\boldsymbol{A})$ with any initialization $\boldsymbol{A}_0 \in \mathsf{O}(n;\mathbb{R})$, where $\langle\cdot,\cdot\rangle : \mathsf{O}(n;\mathbb{R}) \times \mathsf{O}(n;\mathbb{R}) \mapsto \mathbb{R}$ is defined as

$$\langle \boldsymbol{W}_1, \boldsymbol{W}_2\rangle = \mathrm{tr}(\boldsymbol{W}_1^*\boldsymbol{W}_2). \tag{119}$$

Hence, by substituting (119) into (118), yields

$$\boldsymbol{A}_{k+1} = \begin{cases} \mathcal{P}_{\mathsf{O}(n;\mathbb{R})}(\alpha\boldsymbol{A}_k^{\circ 3} + \boldsymbol{A}_k) & \text{when } \alpha < \infty \\ \mathcal{P}_{\mathsf{O}(n;\mathbb{R})}(\boldsymbol{A}_k^{\circ 3}) & \text{when } \alpha = +\infty \end{cases}, \tag{120}$$

where $\mathcal{P}_{\mathsf{O}(n;\mathbb{R})}(\cdot) : \mathbb{R}^{n\times n} \mapsto \mathsf{O}(n;\mathbb{R})$ is the projection onto $\mathsf{O}(n;\mathbb{R})$ (Absil and Malick, 2012):

$$\mathcal{P}_{\mathsf{O}(n;\mathbb{R})}(\boldsymbol{A}) = \boldsymbol{U}\boldsymbol{V}^*, \quad \text{subject to} \quad \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = SVD(\boldsymbol{A}). \tag{121}$$

Moreover, notice that $\|\boldsymbol{A}\|_F^2 = n$ is a constant, so finding a critical point of $h(\boldsymbol{A})$ is equivalent to finding a critical point of $\|\boldsymbol{A}\|_4^4$ over $\mathsf{O}(n;\mathbb{R})$. Hence, we show that projected gradient ascent with any step size $\alpha > 0$ (including $\alpha = +\infty$) finds a critical point of $\|\boldsymbol{A}\|_4^4$ over $\mathsf{O}(n;\mathbb{R})$.[13] ■

13. The proof can easily be generalized to any Stiefel manifolds.

C.5. Proof of Theorem 16

Claim 30 (Local Convergence of the MSP Algorithm) *Given an orthogonal matrix $\boldsymbol{A} \in \mathsf{O}(n;\mathbb{R})$, let $\boldsymbol{A}'$ denote the output of the MSP Algorithm 1 after one iteration: $\boldsymbol{A}' = \boldsymbol{U}\boldsymbol{V}^*$, where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = SVD(\boldsymbol{A}^{\circ 3})$. If $\|\boldsymbol{A}-\boldsymbol{I}\|_F^2 = \varepsilon$, for $\varepsilon < 0.579$, then we have $\|\boldsymbol{A}'-\boldsymbol{I}\|_F^2 < \|\boldsymbol{A}-\boldsymbol{I}\|_F^2$ and $\|\boldsymbol{A}'-\boldsymbol{I}\|_F^2 < O(\varepsilon^3)$.*

Proof Let $\boldsymbol{A} = \boldsymbol{D} + \boldsymbol{N}$, where $\boldsymbol{D}$ is the diagonal part of $\boldsymbol{A}$ and $\boldsymbol{N}$ is the off-diagonal part. Therefore, we have:

$$\left\|\boldsymbol{N}^{\circ 3}\right\|_F \le \|\boldsymbol{N}\|_F^3 \le \|\boldsymbol{A}-\boldsymbol{I}\|_F^3 = \varepsilon^{3/2}, \tag{122}$$

where the first inequality is achieved through Cauchy-Schwarz inequality and the second inequality holds because $\boldsymbol{N}$ is the off-diagonal parts of $\boldsymbol{A}-\boldsymbol{I}$. We can view $\boldsymbol{A}^{\circ 3}$ as a $\boldsymbol{D}^{\circ 3}$ plus a small perturbation $\boldsymbol{N}^{\circ 3}$ with norm as most $\varepsilon^{3/2}$. By Lemma 45, we have:

$$\|\boldsymbol{Q}_{\boldsymbol{A}} - \boldsymbol{Q}_{\boldsymbol{D}}\|_F \le \frac{2\left\|\boldsymbol{A}^{\circ 3} - \boldsymbol{D}^{\circ 3}\right\|_F}{\sigma_n(\boldsymbol{D}^{\circ 3}) + \sigma_n(\boldsymbol{A}^{\circ 3})} = \frac{2\left\|\boldsymbol{N}^{\circ 3}\right\|_F}{\sigma_n(\boldsymbol{D}^{\circ 3}) + \sigma_n(\boldsymbol{A}^{\circ 3})}, \tag{123}$$

where $\boldsymbol{U}_{\boldsymbol{A}}\boldsymbol{\Sigma}_{\boldsymbol{A}}\boldsymbol{V}_{\boldsymbol{A}}^* = \mathrm{SVD}(\boldsymbol{A}^{\circ 3})$, $\boldsymbol{Q}_{\boldsymbol{A}} = \boldsymbol{U}_{\boldsymbol{A}}\boldsymbol{V}_{\boldsymbol{A}}^*$ and $\boldsymbol{U}_{\boldsymbol{D}}\boldsymbol{\Sigma}_{\boldsymbol{D}}\boldsymbol{V}_{\boldsymbol{D}}^* = \mathrm{SVD}(\boldsymbol{D}^{\circ 3})$, $\boldsymbol{Q}_{\boldsymbol{D}} = \boldsymbol{U}_{\boldsymbol{D}}\boldsymbol{V}_{\boldsymbol{D}}$. Notice that $\boldsymbol{D}^{\circ 3}$ is a diagonal matrix, so $\boldsymbol{Q}_{\boldsymbol{D}} = \boldsymbol{I}$, $\sigma_n(\boldsymbol{D}^3) = \min_i a_{i,i}^3$. Moreover

$$\varepsilon = \|\boldsymbol{A}-\boldsymbol{I}\|_F^2 = \sum_{i=1}^n (a_{i,i}-1)^2 + \sum_{i\neq j} a_{i,j}^2 = \sum_{i,j} a_{i,j}^2 - 2\sum_{i=1}^n a_{i,i} + n \iff \sum_{i=1}^n a_{i,i} = n - \frac{\varepsilon}{2}, \tag{124}$$

without loss of generality, we can assume $1 \ge a_{1,1} \ge a_{2,2} \ge \cdots \ge a_{n,n} > 0$, so we have:

$$a_{n,n} = n - \frac{\varepsilon}{2} - (a_{1,1} + a_{2,2} + \cdots + a_{n-1,n-1}) \ge n - \frac{\varepsilon}{2} - (n-1) = 1 - \frac{\varepsilon}{2}, \tag{125}$$

hence we know that

$$\sigma_n(\boldsymbol{D}^{\circ 3}) = \min_i a_{i,i}^3 \ge \left(1 - \frac{\varepsilon}{2}\right)^3. \tag{126}$$

Applying Lemma 46 by substituting

$$\boldsymbol{G} = \boldsymbol{D}^{\circ 3} = \boldsymbol{B}\boldsymbol{M}, \quad \delta\boldsymbol{G} = \boldsymbol{A}^{\circ 3} - \boldsymbol{D}^{\circ 3} = \boldsymbol{N}^{\circ 3} = \delta\boldsymbol{B}\boldsymbol{M}, \tag{127}$$

we know $\boldsymbol{B} = \boldsymbol{I}, \boldsymbol{M} = \boldsymbol{D}^{\circ 3}, \delta\boldsymbol{B} = \boldsymbol{N}^{\circ 3}(\boldsymbol{D}^{\circ 3})^{-1}$, and thus:

$$\frac{\left|\sigma_n(\boldsymbol{A}^{\circ 3}) - \sigma_n(\boldsymbol{D}^{\circ 3})\right|}{\sigma_n(\boldsymbol{D}^{\circ 3})} \le \frac{\|\delta\boldsymbol{B}\|_2}{\sigma_n(\boldsymbol{B})}, \tag{128}$$

which implies

$$\begin{aligned}\left|\sigma_n(\boldsymbol{A}^{\circ 3}) - \sigma_n(\boldsymbol{D}^{\circ 3})\right| &\le \left\|\boldsymbol{N}^{\circ 3}(\boldsymbol{D}^{\circ 3})^{-1}\right\|_2 \sigma_n(\boldsymbol{D}^{\circ 3}) \le \left\|\boldsymbol{N}^{\circ 3}\right\|_2 \left\|(\boldsymbol{D}^{\circ 3})^{-1}\right\|_2 \sigma_n(\boldsymbol{D}^{\circ 3}) \\ &\le \left\|\boldsymbol{N}^{\circ 3}\right\|_F \left(1 - \frac{\varepsilon}{2}\right)^{-3} \sigma_n(\boldsymbol{D}^{\circ 3}) \le \varepsilon^{3/2}\left(1 - \frac{\varepsilon}{2}\right)^{-3} \sigma_n(\boldsymbol{D}^{\circ 3}).\end{aligned} \tag{129}$$

Therefore, using (126), we know that:

$$\sigma_n(\boldsymbol{A}^{\circ 3}) + \sigma_n(\boldsymbol{D}^{\circ 3}) \ge \left[2 - \varepsilon^{3/2}\left(1 - \frac{\varepsilon}{2}\right)^{-3}\right] \sigma_n(\boldsymbol{D}^{\circ 3}) \ge 2\left(1 - \frac{\varepsilon}{2}\right)^3 - \varepsilon^{3/2}. \tag{130}$$

Substituting (122) and (130) into (123), yields

$$\|\boldsymbol{Q_A} - \boldsymbol{Q_D}\|_F \leq \frac{2\varepsilon^{3/2}}{2\left(1-\frac{\varepsilon}{2}\right)^3 - \varepsilon^{3/2}}, \tag{131}$$

hence, we have:

$$\left\|\boldsymbol{A}' - \boldsymbol{I}\right\|_F^2 = \|\boldsymbol{Q_A} - \boldsymbol{Q_D}\|_F^2 \leq O(\varepsilon^3). \tag{132}$$

Moreover, such operation is a contraction whenever:

$$\begin{aligned} &\frac{2\varepsilon^{3/2}}{2\left(1-\frac{\varepsilon}{2}\right)^3 - \varepsilon^{3/2}} < \|\boldsymbol{A} - \boldsymbol{D}\|_F = \varepsilon^{1/2} \\ \Longleftrightarrow\ & 2\varepsilon - 2\left(1-\frac{\varepsilon}{2}\right)^3 - \varepsilon^{3/2} < 0 \\ \Longleftarrow\ & \varepsilon < 0.579, \end{aligned} \tag{133}$$

which completes the proof. ■

C.6. Proof of Proposition 17

Claim 31 (Global Convergence of the MSP Algorithm on $\mathsf{SO}(2;\mathbb{R})$) *When $n = 2$, if we denote our $\boldsymbol{A}_t \in \mathsf{SO}(2,\mathbb{R})$ as following form:*

$$\boldsymbol{A}_t = \begin{pmatrix} \cos\theta_t & -\sin\theta_t \\ \sin\theta_t & \cos\theta_t \end{pmatrix}, \quad \forall \theta_t \in \Big[-\frac{\pi}{2}, \frac{\pi}{2}\Big], \tag{134}$$

then: 1) $\boldsymbol{A}_{t+1} \in \mathsf{SO}(n;\mathbb{R})$; 2) if we let

$$\boldsymbol{A}_{t+1} = \begin{pmatrix} \cos\theta_{t+1} & -\sin\theta_{t+1} \\ \sin\theta_{t+1} & \cos\theta_{t+1} \end{pmatrix}, \quad \forall \theta_{t+1} \in \Big[-\frac{\pi}{2}, \frac{\pi}{2}\Big], \tag{135}$$

θ_t and θ_{t+1} satisfies the following relation

$$\theta_{t+1} = \tan^{-1}\left(\tan^3\theta_t\right). \tag{136}$$

Proof By the update of the MSP algorithm, we know that

$$\boldsymbol{A}_{t+1} = \underset{\boldsymbol{M}\in\mathsf{O}(n;\mathbb{R})}{\arg\min} \left\|\boldsymbol{M} - \boldsymbol{A}_t^{\circ 3}\right\|_F^2. \tag{137}$$

$\forall \boldsymbol{M} \in O(n;\mathbb{R})$, we can denote $\boldsymbol{M}$ as

$$\boldsymbol{M} = \begin{cases} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} & \text{if } \det(\boldsymbol{M}) = 1 \\ \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix} & \text{if } \det(\boldsymbol{M}) = -1, \end{cases} , \theta \in (-\pi, \pi] \tag{138}$$

Then if $\det(\boldsymbol{M}) = 1$, we have:

$$\begin{aligned}\left\|\boldsymbol{M} - \boldsymbol{A}_t^{\circ 3}\right\|_F^2 &= 2(\cos^3\theta_t - \cos\theta)^2 + 2(\sin^3\theta_t - \sin\theta)^2 \\ &= 2\cos^6\theta_t + 2\sin^6\theta_t - 4\cos\theta\cos^3\theta_t - 4\sin\theta\sin^3\theta_t + 2.\end{aligned} \tag{139}$$

When $\det(\boldsymbol{M}) = -1$, we have:

$$\begin{aligned}\left\|\boldsymbol{M} - \boldsymbol{A}_t^{\circ 3}\right\|_F^2 &= (\cos^3\theta_t - \cos\theta)^2 + (\cos^3\theta_t + \cos\theta)^2 + (\sin^3\theta_t - \sin\theta)^2 + (\sin^3\theta_t + \sin\theta)^2 \\ &= 2\cos^6\theta_t + 2\sin^6\theta_t + 2.\end{aligned} \tag{140}$$

Notice that when $\det(\boldsymbol{A}) = -1$, $\left\|\boldsymbol{M} - \boldsymbol{A}_t^{\circ 3}\right\|_F^2$ is a constant, so as long as we pick θ in the same quadrant with θ_t, then

$$2\cos^6\theta_t + 2\sin^6\theta_t - 4\cos\theta\cos^3\theta_t - 4\sin\theta\sin^3\theta_t + 2 \le 2\cos^6\theta_t + 2\sin^6\theta_t + 2, \tag{141}$$

and therefore, we know $\boldsymbol{A}_{t+1} \in \mathsf{SO}(n;\mathbb{R})$. So, we know that θ should satisfy the first order condition of critical point condition of the following optimization problem:

$$\boldsymbol{A}_{t+1} = \operatorname*{arg\,min}_{\boldsymbol{M}\in\mathsf{SO}(n;\mathbb{R})} \left\|\boldsymbol{M} - \boldsymbol{A}_t^{\circ 3}\right\|_F^2, \tag{142}$$

which implies

$$\nabla_\theta\left[2(\cos^3\theta_t - \cos\theta)^2 + 2(\sin^3\theta_t - \sin\theta)^2\right] = 0 \implies \sin\theta\cos^3\theta_t - \cos\theta\sin^3\theta_t = 0 \tag{143}$$

which implies

$$\begin{cases}\tan\theta = \tan^3\theta_t, \text{ when } \theta_t \neq \pm\pi/2 \\ \cot\theta = \cot^3\theta_t, \text{ when } \theta_t \neq 0, \pi.\end{cases} \tag{144}$$

Note that we distinguish the different cases $\theta_t = \pm\pi/2, 0, \pi$ to avoid the division by zero in (143). One can also ignore this by taking the inverse on tangent, which yields

$$\theta_{t+1} = \tan^{-1}\left(\tan^3\theta_t\right), \tag{145}$$

as stated. ■

D. Related Lemmas and Inequalities

D.1. Some Basic Inequalities

Lemma 32 (One-sided Bernstein's Inequality) *Given n random variables $x_1, x_2, \dots x_n$, if $\forall i \in [n], x_i \le b$ almost surely, then*

$$\mathbb{P}\Big(\sum_{i=1}^n \big(x_i - \mathbb{E}[x_i]\big) \ge nt\Big) \le \exp\left(-\frac{nt^2}{2(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[x_i^2] + bt/3)}\right). \tag{146}$$

Proof See Proposition 2.14 in Wainwright (2019). ■

Lemma 33 (Some Useful Norm Matrix Norm Inequalities) *Given two matrix $\boldsymbol{A}, \boldsymbol{B}$:*

1. *if $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times m}$, then $\|\boldsymbol{A}\circ\boldsymbol{B}\|_F \le \|\boldsymbol{A}\|_F \|\boldsymbol{B}\|_F$*
2. *if $\boldsymbol{A} \in \mathbb{R}^{n\times r}, \boldsymbol{B} \in \mathbb{R}^{r\times m}$, then $\|\boldsymbol{A}\boldsymbol{B}\|_F \le \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_F$*
3. *if $\boldsymbol{A} \in \mathbb{R}^{n\times r}, \boldsymbol{B} \in \mathbb{R}^{r\times m}$, then $\|\boldsymbol{A}\boldsymbol{B}\|_F \le \|\boldsymbol{A}\|_F \|\boldsymbol{B}\|_F$*

Proof

1. $\|\boldsymbol{A}\circ\boldsymbol{B}\|_F^2 = \sum_{i,j}(a_{i,j}b_{i,j})^2 \le \Big(\sum_{i,j} a_{i,j}^2\Big)\Big(\sum_{i,j} b_{i,j}^2\Big) = \|\boldsymbol{A}\|_F^2 \|\boldsymbol{B}\|_F^2$.
2. $\|\boldsymbol{A}\boldsymbol{B}\|_F^2 = \sum_j \|\boldsymbol{A}\boldsymbol{b}_j\|_F^2 \le \|\boldsymbol{A}\|_2^2 \sum_j \|\boldsymbol{b}_j\|_2^2 = \|\boldsymbol{A}\|_2^2 \|\boldsymbol{B}\|_F^2$.
3. Let $\boldsymbol{a}_i, i \in [n]$ be the i^{th} row vector of $\boldsymbol{A}$ and $\boldsymbol{b}_j, j \in [m]$ be the j^{th} column vector of $\boldsymbol{B}$. So

$$\|\boldsymbol{A}\boldsymbol{B}\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} |\boldsymbol{a}_i^*\boldsymbol{b}_j|^2 \le \Big(\sum_{i=1}^{n}\|\boldsymbol{a}_i\|_2^2\Big)\Big(\sum_{j=1}^{m}\|\boldsymbol{b}_i\|_2^2\Big) = \|\boldsymbol{A}\|_F^2\|\boldsymbol{B}\|_F^2\,. \tag{147}$$

■

D.2. Truncation of Bernoulli-Gaussian Matrix

Lemma 34 (Entry-wise Truncation of a Bernoulli Gaussian Matrix) *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, where $x_{i,j} \sim_{iid} BG(\theta)$ and let $\|\cdot\|_\infty$ denote the maximum element (in absolute value) of a matrix, then*

$$\mathbb{P}\Big(\|\boldsymbol{X}\|_\infty \ge t\Big) \le 2np\theta \exp\left(-\frac{t^2}{2}\right). \tag{148}$$

Proof A Bernoulli Gaussian variable $x_{i,j}, \forall i \in [n], j \in [p]$ satisfies $x_{i,j} = b_{i,j}\cdot g_{i,j}$, where $b_{i,j} \sim_{iid} \text{Ber}(\theta)$, $g_{i,j} \sim_{iid} \mathcal{N}(0,1)$ and therefore

$$\mathbb{P}\big(|x_{i,j}| \ge t\big) = \theta\cdot\mathbb{P}\big(|g_{i,j}| \ge t\big) \le 2\theta\exp\left(-\frac{t^2}{2}\right). \tag{149}$$

By union bound, we have:

$$\mathbb{P}\big(\|\boldsymbol{X}\|_\infty \ge t\big) \le \sum_{i=1}^{n}\sum_{j=1}^{p}\mathbb{P}\big(|x_{i,j}| \ge t\big) \le 2np\theta\exp\left(-\frac{t^2}{2}\right). \tag{150}$$

■

D.3. $\varepsilon-$covering of Stiefel Manifolds

Lemma 35 ($\varepsilon-$Net Covering of Stiefel Manifolds) [14] *There is a covering $\varepsilon-$net $\mathcal{S}_\varepsilon$ for Stiefel manifold $\mathcal{M} = \{\boldsymbol{W} \in \mathbb{R}^{n\times r} | \boldsymbol{W}^*\boldsymbol{W} = \boldsymbol{I}\}, (n \geq r)$ in operator norm*

$$\forall \boldsymbol{W} \in \mathcal{M},\ \exists \boldsymbol{W}' \in \mathcal{S}_\varepsilon \quad \text{subject to} \quad \left\|\boldsymbol{W} - \boldsymbol{W}'\right\|_2 \leq \varepsilon, \tag{151}$$

of size $|\mathcal{S}_\varepsilon| \leq \left(\frac{6}{\varepsilon}\right)^{nr}$.

Proof Let $\mathcal{S}'_{\varepsilon/2} = \{\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_{|\mathcal{S}_{\varepsilon/2}|}\}$ be an $\varepsilon/2-$nets for the unit operator norm ball of $n \times r$ matrix $\{\boldsymbol{A} \in \mathbb{R}^{n\times r} |\, \|\boldsymbol{A}\|_2 \leq 1\}$, $\varepsilon-$net covering theorem shows that such construction of $\mathcal{S}'_{\varepsilon/2}$ exists and $|\mathcal{S}'_{\varepsilon/2}| \leq \left(\frac{6}{\varepsilon}\right)^{nr}$. Next, let $\mathcal{S}'$ be the subset of $\mathcal{S}'_{\varepsilon/2}$, which consists elements of $\mathcal{S}'_{\varepsilon/2}$ whose distance are within $\varepsilon/2$ with $\mathcal{M}$:

$$\mathcal{S}' = \{\boldsymbol{A} \in \mathcal{S}'_{\varepsilon/2} | \exists \boldsymbol{W} \in \mathcal{M}, \text{subject to } \|\boldsymbol{W} - \boldsymbol{A}\|_2 \leq \varepsilon/2\}. \tag{152}$$

Then, $\forall \boldsymbol{A} \in \mathcal{S}'$, let $\hat{\boldsymbol{W}}(\boldsymbol{A})$ be the nearest element of $\boldsymbol{A}$ in $\mathcal{M}$:

$$\hat{\boldsymbol{W}}(\boldsymbol{A}) = \mathop{\arg\min}_{\boldsymbol{W}\in\mathcal{M}} \|\boldsymbol{W} - \boldsymbol{A}\|_2 , \tag{153}$$

and let $\mathcal{S}_\varepsilon$ be the set of the nearest element of each $\boldsymbol{A}$ in $\mathcal{S}'$: $\mathcal{S}_\varepsilon = \{\hat{\boldsymbol{W}}(\boldsymbol{A}) | \boldsymbol{A} \in \mathcal{S}'\}$. Since $\mathcal{S}'_{\varepsilon/2}$ is an $\varepsilon/2$-nets for $\mathcal{M}$. So $\forall \boldsymbol{W} \in \mathcal{M}$, there exists $\boldsymbol{A}_l \in \mathcal{S}'_{\varepsilon/2}$, such that

$$\|\boldsymbol{W} - \boldsymbol{A}_l\|_2 \leq \frac{\varepsilon}{2}, \tag{154}$$

so $\boldsymbol{A}_l \in \mathcal{S}'$, and therefore there exists $\hat{\boldsymbol{W}}(\boldsymbol{A}_l) \in \mathcal{S}_\varepsilon$, such that

$$\left\|\hat{\boldsymbol{W}}(\boldsymbol{A}_l) - \boldsymbol{A}_l\right\|_2 \leq \|\boldsymbol{W} - \boldsymbol{A}_l\|_2 \leq \frac{\varepsilon}{2}, \tag{155}$$

hence by triangle inequality

$$\left\|\boldsymbol{W} - \hat{\boldsymbol{W}}(\boldsymbol{A}_l)\right\|_2 \leq \left\|\hat{\boldsymbol{W}}(\boldsymbol{A}_l) - \boldsymbol{A}_l\right\|_2 + \|\boldsymbol{W} - \boldsymbol{A}_l\|_2 \leq \varepsilon. \tag{156}$$

Thus, $\mathcal{S}_\varepsilon$ is an $\varepsilon-$net for $\mathcal{M}$ and $|\mathcal{S}_\varepsilon| = |\mathcal{S}'| \leq |\mathcal{S}'_{\varepsilon/2}| \leq \left(\frac{6}{\varepsilon}\right)^{nr}$. ■

D.4. Convergence to Maxima of ℓ^4-Norm over Unit Sphere

Lemma 36 (Single Vector Convergence over Unit Sphere) *Suppose $\boldsymbol{q}$ is a vector on the unit sphere: $\boldsymbol{q} \in \mathbb{S}^{n-1}$. $\forall \varepsilon \in [0, 1]$, if $\|\boldsymbol{q}\|_4^4 \geq 1 - \varepsilon$, then $\exists i \in [n]$, such that*

$$\|\boldsymbol{q} - \boldsymbol{e}_i\|_2^2 \leq 2\varepsilon \quad \textit{when } q_i > 0, \quad \|\boldsymbol{q} + \boldsymbol{e}_i\|_2^2 \leq 2\varepsilon \quad \textit{when } q_i < 0, \tag{157}$$

where $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n\}$ is the canonical basis of $\mathbb{R}^n$.

14. A similar result can be found in Lemma 4.5 of Recht et al. (2010).

Proof Let $\boldsymbol{q} = [q_1, q_2, \ldots, q_n]^*$, and without loss of generality, we can assume

$$1 \geq q_1^2 \geq q_2^2 \geq \cdots \geq q_n^2 \geq 0. \tag{158}$$

Also, from the assumption

$$q_1^4 + q_2^4 + \cdots + q_n^4 \geq 1 - \varepsilon, \tag{159}$$

and along with (158) and $\boldsymbol{q} \in \mathbb{S}^{n-1}$, we have:

$$q_1^4 + q_2^2(q_2^2 + q_3^2 + \cdots + q_n^2) = q_1^4 + q_2^2(1 - q_1^2) \geq 1 - \varepsilon, \tag{160}$$

which implies:

$$q_1^4 + q_1^2(1 - q_1^2) \geq q_1^4 + q_2^2(1 - q_1^2) \geq 1 - \varepsilon \implies q_1^2 \geq 1 - \varepsilon. \tag{161}$$

Hence, we know

$$q_2^2 + \cdots + q_n^2 \leq \varepsilon. \tag{162}$$

Moreover, (161) also implies

$$\varepsilon \geq (1 - q_1)(1 + q_1) \implies \begin{cases} 1 - q_1 \leq \varepsilon/(1 + q_1) \leq \varepsilon & \text{when } q_1 > 0, \\ 1 + q_1 \leq \varepsilon/(1 - q_1) \leq \varepsilon & \text{when } q_1 < 0. \end{cases} \tag{163}$$

When $q_1 > 0$, combine (161) and (162), we have

$$\|\boldsymbol{q} - \boldsymbol{e}_1\|_2^2 = (q_1 - 1)^2 + q_2^2 + \cdots + q_n^2 \leq \varepsilon^2 + \varepsilon \leq 2\varepsilon. \tag{164}$$

And similar result for $\|\boldsymbol{q} + \boldsymbol{e}_1\|_2^2 \leq 2\varepsilon$ can be obtained through the same reasoning. ■

D.5. Multiple Vectors Convergence to Maxima of ℓ^4-Norm over Unit Sphere

Lemma 37 (Multiple Vectors Convergence over Unit Sphere) *Suppose $\boldsymbol{q}_1, \boldsymbol{q}_2 \ldots, \boldsymbol{q}_k$ are k vectors on the unit sphere: $\boldsymbol{q}_i \in \mathbb{S}^{n-1}, \forall i \in [k]$. $\forall \varepsilon \in [0, 1]$, if*

$$\frac{1}{k} \sum_{i=1}^{k} \|\boldsymbol{q}_i\|_4^4 \geq 1 - \varepsilon, \tag{165}$$

then $\exists j_1, j_2, \ldots j_k \in [n]$, such that

$$\frac{1}{k} \sum_{i=1}^{k} \|\boldsymbol{q}_i - s_i \boldsymbol{e}_{j_i}\|_2^2 \leq 2\varepsilon, \tag{166}$$

where $\boldsymbol{e}_{j_i}$ are one vector in canonical basis and $s_i \in \{1, -1\}$ indicates the sign of $\boldsymbol{e}_{j_i}$, $\forall i \in [k]$.

Proof Note that we can reformulate the condition (165) as

$$\sum_{i=1}^{k} \|\boldsymbol{q}_i\|_4^4 \geq k - k\varepsilon. \tag{167}$$

Since $\forall i \in [k]$, $0 \leq \|\boldsymbol{q}_i\|_4^4 \leq \|\boldsymbol{q}_i\|_2^4 = 1$, we can assume

$$\|\boldsymbol{q}_i\|_4^4 = 1 - \alpha_i \varepsilon, \quad \forall i \in [k], \tag{168}$$

where α_i satisfies $\alpha_i \varepsilon \in [0, 1]$, for all $i \in [k]$ and

$$\sum_{i=1}^{k} \alpha_i \leq k. \tag{169}$$

By Lemma 36, we know there exists $s_i \in \{1, -1\}$ and j_i, such that

$$\|\boldsymbol{q}_i - s_i \boldsymbol{e}_{j_i}\|_2^2 \leq 2\alpha_i \varepsilon, \quad \forall i \in [k]. \tag{170}$$

Along with (169), we have:

$$\frac{1}{k} \sum_{i=1}^{k} \|\boldsymbol{q}_i - s_i \boldsymbol{e}_{j_i}\|_2^2 \leq \frac{2}{k} \varepsilon \sum_{i=1}^{k} \alpha_i \leq 2\varepsilon, \tag{171}$$

which completes the proof. ■

D.6. Related Lipschitz Constants

Lemma 38 (Lipschitz Constant of $\frac{1}{np}\hat{f}(\cdot,\cdot)$ over $\mathsf{O}(n;\mathbb{R})$) *If $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $x_{i,j} \sim_{iid} BG(\theta)$, and let $\bar{\boldsymbol{X}}$ be the truncation of $\boldsymbol{X}$ by bound B*

$$\bar{x}_{i,j} = \begin{cases} x_{i,j} & \text{if} \quad |x_{i,j}| \leq B \\ 0 & \text{else} \end{cases}, \tag{172}$$

then $\forall \boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathsf{O}(n;\mathbb{R})$, we have

$$\frac{1}{np} \left| \left\|\boldsymbol{W}_1 \bar{\boldsymbol{X}}\right\|_4^4 - \left\|\boldsymbol{W}_2 \bar{\boldsymbol{X}}\right\|_4^4 \right| \leq L_1 \left\|\boldsymbol{W}_1 - \boldsymbol{W}_2\right\|_2, \tag{173}$$

for a constant $L_1 \leq 4npB^4$.

Proof Notice that

$$\begin{aligned}
&\left| \left\|\boldsymbol{W}_1 \bar{\boldsymbol{X}}\right\|_4^4 - \left\|\boldsymbol{W}_2 \bar{\boldsymbol{X}}\right\|_4^4 \right| = \left| \sum_{i,j} \left[\left(\boldsymbol{W}_1 \bar{\boldsymbol{X}}\right)_{i,j}^4 - \left(\boldsymbol{W}_2 \bar{\boldsymbol{X}}\right)_{i,j}^4 \right] \right| \\
=& \left| \sum_{i,j} \left\{ \left[\left(\boldsymbol{W}_1 \bar{\boldsymbol{X}}\right)_{i,j}^2 - \left(\boldsymbol{W}_2 \bar{\boldsymbol{X}}\right)_{i,j}^2 \right] \left[\left(\boldsymbol{W}_1 \bar{\boldsymbol{X}}\right)_{i,j}^2 + \left(\boldsymbol{W}_2 \bar{\boldsymbol{X}}\right)_{i,j}^2 \right] \right\} \right| \\
\leq& \left\{ \sum_{i,j} \left[\left(\boldsymbol{W}_1 \bar{\boldsymbol{X}}\right)_{i,j}^2 - \left(\boldsymbol{W}_2 \bar{\boldsymbol{X}}\right)_{i,j}^2 \right]^2 \right\}^{1/2} \left\{ \sum_{i,j} \left[\left(\boldsymbol{W}_1 \bar{\boldsymbol{X}}\right)_{i,j}^2 + \left(\boldsymbol{W}_2 \bar{\boldsymbol{X}}\right)_{i,j}^2 \right]^2 \right\}^{1/2} \\
=& \underbrace{\left\| \left(\boldsymbol{W}_1 \bar{\boldsymbol{X}}\right)^{\circ 2} - \left(\boldsymbol{W}_2 \bar{\boldsymbol{X}}\right)^{\circ 2} \right\|_F}_{\Gamma_1} \underbrace{\left\| \left(\boldsymbol{W}_1 \bar{\boldsymbol{X}}\right)^{\circ 2} + \left(\boldsymbol{W}_2 \bar{\boldsymbol{X}}\right)^{\circ 2} \right\|_F}_{\Gamma_2},
\end{aligned} \tag{174}$$

the only inequality is obtained through Cauchy−Schwarz inequality. For Γ_1, we have

$$\begin{aligned}\Gamma_1 =& \left\| \left(\boldsymbol{W}_1\bar{\boldsymbol{X}}\right)^{\circ 2} - \left(\boldsymbol{W}_2\bar{\boldsymbol{X}}\right)^{\circ 2} \right\|_F = \left\| \left(\boldsymbol{W}_1\bar{\boldsymbol{X}} - \boldsymbol{W}_2\bar{\boldsymbol{X}}\right) \circ \left(\boldsymbol{W}_1\bar{\boldsymbol{X}} + \boldsymbol{W}_2\bar{\boldsymbol{X}}\right) \right\|_F \\ \leq& \left\| \left(\boldsymbol{W}_1 - \boldsymbol{W}_2\right)\bar{\boldsymbol{X}} \right\|_F \left\| \left(\boldsymbol{W}_1 + \boldsymbol{W}_2\right)\bar{\boldsymbol{X}} \right\|_F \quad \text{(By inequality 1 in Lemma 33)} \\ \leq& \left\| \boldsymbol{W}_1 - \boldsymbol{W}_2 \right\|_2 \left\| \boldsymbol{W}_1 + \boldsymbol{W}_2 \right\|_2 \left\| \bar{\boldsymbol{X}} \right\|_F^2 \quad \text{(By inequality 2 in Lemma 33)} \\ \leq& \left\| \boldsymbol{W}_1 - \boldsymbol{W}_2 \right\|_2 \left(\left\| \boldsymbol{W}_1 \right\|_2 + \left\| \boldsymbol{W}_2 \right\|_2 \right) \left\| \bar{\boldsymbol{X}} \right\|_F^2 \\ =& 2 \left\| \boldsymbol{W}_1 - \boldsymbol{W}_2 \right\|_2 \left\| \bar{\boldsymbol{X}} \right\|_F^2 \quad (\boldsymbol{W}_1, \boldsymbol{W}_2 \text{ are orthogonal, } \left\| \boldsymbol{W}_1 \right\|_2 = \left\| \boldsymbol{W}_2 \right\|_2 = 1) \\ \leq& 2np B^2 \left\| \boldsymbol{W}_1 - \boldsymbol{W}_2 \right\|_2 \quad (|\bar{x}_{i,j}| \leq B).\end{aligned} \tag{175}$$

For Γ_2, we have

$$\begin{aligned}\Gamma_2 =& \left\| \left(\boldsymbol{W}_1\bar{\boldsymbol{X}}\right)^{\circ 2} + \left(\boldsymbol{W}_2\bar{\boldsymbol{X}}\right)^{\circ 2} \right\|_F \\ \leq& \left\| \left(\boldsymbol{W}_1\bar{\boldsymbol{X}}\right)^{\circ 2} \right\|_F + \left\| \left(\boldsymbol{W}_1\bar{\boldsymbol{X}}\right)^{\circ 2} \right\|_F \\ \leq& \left\| \boldsymbol{W}_1\bar{\boldsymbol{X}} \right\|_F^2 + \left\| \boldsymbol{W}_2\bar{\boldsymbol{X}} \right\|_F^2 \quad \text{(By inequality 1 in Lemma 33)} \\ \leq& \left\| \boldsymbol{W}_1 \right\|_2^2 \left\| \bar{\boldsymbol{X}} \right\|_F^2 + \left\| \boldsymbol{W}_2 \right\|_2^2 \left\| \bar{\boldsymbol{X}} \right\|_F^2 \\ =& 2 \left\| \bar{\boldsymbol{X}} \right\|_F^2 \quad (\boldsymbol{W}_1, \boldsymbol{W}_2 \text{ are orthogonal matrices, } \left\| \boldsymbol{W}_1 \right\|_2 = \left\| \boldsymbol{W}_2 \right\|_2 = 1) \\ \leq& 2np B^2 \quad (|\bar{x}_{i,j}| \leq B).\end{aligned} \tag{176}$$

Thus, we know that

$$\frac{1}{np} \left| \left\| \boldsymbol{W}_1\bar{\boldsymbol{X}} \right\|_4^4 - \left\| \boldsymbol{W}_2\bar{\boldsymbol{X}} \right\|_4^4 \right| \leq \frac{1}{np}\Gamma_1\Gamma_2 \leq 4npB^4 \left\| \boldsymbol{W}_1 - \boldsymbol{W}_2 \right\|_2 . \tag{177}$$

■

Lemma 39 (Lipschitz Constant of $\frac{1}{np}f(\cdot)$ over $\mathsf{O}(n;\mathbb{R})$) *If $\boldsymbol{X} \in \mathbb{R}^{n\times p}, x_{i,j} \sim_{iid} BG(\theta)$, then $\forall \boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathsf{O}(n;\mathbb{R})$, we have*

$$\frac{1}{np} \left| \mathbb{E} \left\| \boldsymbol{W}_1\boldsymbol{X} \right\|_4^4 - \mathbb{E} \left\| \boldsymbol{W}_2\boldsymbol{X} \right\|_4^4 \right| \leq L_2 \left\| \boldsymbol{W}_1 - \boldsymbol{W}_2 \right\|_2 , \tag{178}$$

with a constant $L_2 \leq 12n\theta(1-\theta)$.

Proof According to Lemma 3, we have

$$\mathbb{E} \left\| \boldsymbol{W}\boldsymbol{X} \right\|_4^4 = 3p\theta(1-\theta) \left\| \boldsymbol{W} \right\|_4^4 + 3\theta^2 np, \quad \forall \boldsymbol{W} \in \mathsf{O}(n;\mathbb{R}), \tag{179}$$

so

$$\left| \mathbb{E} \left\| \boldsymbol{W}_1\boldsymbol{X} \right\|_4^4 - \mathbb{E} \left\| \boldsymbol{W}_2\boldsymbol{X} \right\|_4^4 \right| = 3p\theta(1-\theta) \left| \left\| \boldsymbol{W}_1 \right\|_4^4 - \left\| \boldsymbol{W}_2 \right\|_4^4 \right| . \tag{180}$$

Notice that

$$
\begin{aligned}
&\left|\|\boldsymbol{W}_1\|_4^4-\|\boldsymbol{W}_2\|_4^4\right|=\left|\sum_{i,j}\left[(\boldsymbol{W}_1)_{i,j}^4-(\boldsymbol{W}_2)_{i,j}^4\right]\right|\\
=&\left|\sum_{i,j}\left\{\left[(\boldsymbol{W}_1)_{i,j}^2-(\boldsymbol{W}_2)_{i,j}^2\right]\left[(\boldsymbol{W}_1)_{i,j}^2+(\boldsymbol{W}_2)_{i,j}^2\right]\right\}\right|\\
=&\left\{\sum_{i,j}\left[(\boldsymbol{W}_1)_{i,j}^2-(\boldsymbol{W}_2)_{i,j}^2\right]^2\right\}^{1/2}\left\{\sum_{i,j}\left[(\boldsymbol{W}_1)_{i,j}^2+(\boldsymbol{W}_2)_{i,j}^2\right]^2\right\}^{1/2}\quad\text{(Cauchy−Schwarz)}\\
=&\left\|\boldsymbol{W}_1^{\circ 2}-\boldsymbol{W}_2^{\circ 2}\right\|_F\left\|\boldsymbol{W}_1^{\circ 2}+\boldsymbol{W}_2^{\circ 2}\right\|_F=\left\|(\boldsymbol{W}_1-\boldsymbol{W}_2)\circ(\boldsymbol{W}_1+\boldsymbol{W}_2)\right\|_F\left\|\boldsymbol{W}_1^{\circ 2}+\boldsymbol{W}_2^{\circ 2}\right\|_F\\
\leq&\left\|\boldsymbol{W}_1-\boldsymbol{W}_2\right\|_F\left\|\boldsymbol{W}_1+\boldsymbol{W}_2\right\|_F\left(\left\|\boldsymbol{W}_1^{\circ 2}\right\|_F+\left\|\boldsymbol{W}_2^{\circ 2}\right\|_F\right)\quad\text{(By Lemma 33)}\\
\leq&n\left\|\boldsymbol{W}_1-\boldsymbol{W}_2\right\|_2\left\|\boldsymbol{W}_1+\boldsymbol{W}_2\right\|_2\left(\left\|\boldsymbol{W}_1\right\|_F^2+\left\|\boldsymbol{W}_2\right\|_F^2\right)\quad\text{(By Lemma 33, }\|\boldsymbol{W}\|_F\leq\sqrt{n}\,\|\boldsymbol{W}\|_2\text{)}\\
\leq&4n^2\left\|\boldsymbol{W}_1-\boldsymbol{W}_2\right\|_2\quad(\|\boldsymbol{W}_1+\boldsymbol{W}_2\|_2\leq\|\boldsymbol{W}_1\|_2+\|\boldsymbol{W}_2\|_2\text{ and }\|\boldsymbol{W}_1\|_2=\|\boldsymbol{W}_2\|_2=1).
\end{aligned}
\tag{181}
$$

Hence,

$$
\frac{1}{np}\left|\mathbb{E}\left\|\boldsymbol{W}_1\boldsymbol{X}\right\|_4^4-\mathbb{E}\left\|\boldsymbol{W}_2\boldsymbol{X}\right\|_4^4\right|\leq 12n\theta(1-\theta)\left\|\boldsymbol{W}_1-\boldsymbol{W}_2\right\|_2,\tag{182}
$$

which completes the proof. ■

Lemma 40 (Lipschitz Constant of $\nabla\hat{f}(\cdot,\cdot)$) *If $\boldsymbol{X}\in\mathbb{R}^{n\times p},x_{i,j}\sim_{iid}BG(\theta)$, and let $\bar{\boldsymbol{X}}$ be the truncation of $\boldsymbol{X}$ by bound B*

$$
\bar{x}_{i,j}=\begin{cases}x_{i,j} & if\quad |x_{i,j}|\leq B\\ 0 & else\end{cases},\tag{183}
$$

then $\forall\boldsymbol{W}_1,\boldsymbol{W}_2\in\mathsf{O}(n;\mathbb{R})$, we have

$$
\frac{1}{np}\left\|(\boldsymbol{W}_1\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*-(\boldsymbol{W}_2\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\right\|_F\leq L_1\left\|\boldsymbol{W}_1-\boldsymbol{W}_2\right\|_2,\tag{184}
$$

with a constant $L_1\leq 3npB^4$.

Proof Notice that

$$
\begin{aligned}
&\frac{1}{np}\left\|(\boldsymbol{W}_1\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*-(\boldsymbol{W}_2\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\right\|_F=\frac{1}{np}\left\|\left[(\boldsymbol{W}_1\bar{\boldsymbol{X}})^{\circ 3}-(\boldsymbol{W}_2\bar{\boldsymbol{X}})^{\circ 3}\right]\bar{\boldsymbol{X}}^*\right\|_F\\
\leq&\frac{1}{np}\left\|(\boldsymbol{W}_1\bar{\boldsymbol{X}}-\boldsymbol{W}_2\bar{\boldsymbol{X}})\circ\left[(\boldsymbol{W}_1\bar{\boldsymbol{X}})^{\circ 2}+(\boldsymbol{W}_2\bar{\boldsymbol{X}})^{\circ 2}+(\boldsymbol{W}_1\bar{\boldsymbol{X}})\circ(\boldsymbol{W}_2\bar{\boldsymbol{X}})\right]\right\|_F\left\|\bar{\boldsymbol{X}}^*\right\|_F\\
\leq&\frac{1}{np}\underbrace{\left\|\boldsymbol{W}_1\bar{\boldsymbol{X}}-\boldsymbol{W}_2\bar{\boldsymbol{X}}\right\|_F}_{\Gamma_1}\underbrace{\left\|(\boldsymbol{W}_1\bar{\boldsymbol{X}})^{\circ 2}+(\boldsymbol{W}_2\bar{\boldsymbol{X}})^{\circ 2}+(\boldsymbol{W}_1\bar{\boldsymbol{X}})\circ(\boldsymbol{W}_2\bar{\boldsymbol{X}})\right\|_F}_{\Gamma_2}\left\|\bar{\boldsymbol{X}}^*\right\|_F,
\end{aligned}
\tag{185}
$$

where the $\leq$ is achieved by inequality 1 in Lemma 33. For Γ_1, using inequality 2 in Lemma 33, we have

$$\Gamma_1 = \left\|(\boldsymbol{W}_1 - \boldsymbol{W}_2)\bar{\boldsymbol{X}}\right\|_F \leq \|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_2 \left\|\bar{\boldsymbol{X}}\right\|_F. \tag{186}$$

For Γ_2, we have

$$\begin{aligned}
\Gamma_2 &= \left\|(\boldsymbol{W}_1\bar{\boldsymbol{X}})^{\circ 2} + (\boldsymbol{W}_2\bar{\boldsymbol{X}})^{\circ 2} + (\boldsymbol{W}_1\bar{\boldsymbol{X}}) \circ (\boldsymbol{W}_2\bar{\boldsymbol{X}})\right\|_F \\
&\leq \left(\left\|(\boldsymbol{W}_1\bar{\boldsymbol{X}})^{\circ 2}\right\|_F + \left\|(\boldsymbol{W}_2\bar{\boldsymbol{X}})^{\circ 2}\right\|_F + \left\|(\boldsymbol{W}_1\bar{\boldsymbol{X}}) \circ (\boldsymbol{W}_2\bar{\boldsymbol{X}})\right\|_F \right) \\
&\leq \left(\left\|\boldsymbol{W}_1\bar{\boldsymbol{X}}\right\|_F^2 + \left\|\boldsymbol{W}_2\bar{\boldsymbol{X}}\right\|_F^2 + \left\|\boldsymbol{W}_1\bar{\boldsymbol{X}}\right\|_F \left\|\boldsymbol{W}_2\bar{\boldsymbol{X}}\right\|_F \right) \quad \text{(By inequality 1 in Lemma 33)} \\
&= 3\left\|\bar{\boldsymbol{X}}\right\|_F^2 \quad (\|\cdot\|_F \text{ is rotation invariant}).
\end{aligned} \tag{187}$$

Hence, we have

$$\begin{aligned}
&\frac{1}{np}\left\|(\boldsymbol{W}_1\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - (\boldsymbol{W}_2\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\right\|_F \leq \frac{1}{np}\Gamma_1\Gamma_2\left\|\bar{\boldsymbol{X}}\right\|_F \\
\leq &\frac{3}{np}\left\|\bar{\boldsymbol{X}}\right\|_F^4 \|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_2 = 3npB^4\|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_2,
\end{aligned} \tag{188}$$

which completes the proof. ■

Lemma 41 (Lipschitz Constant of $\mathbb{E}\frac{1}{p}\nabla\hat{f}(\cdot,\cdot)$) *If $\boldsymbol{X} \in \mathbb{R}^{n\times p}, x_{i,j} \sim_{iid} BG(\theta)$, then $\forall \boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathsf{O}(n;\mathbb{R})$, we have*

$$\frac{1}{np}\left\|\mathbb{E}\left[(\boldsymbol{W}_1\bar{\boldsymbol{X}}^*)^{\circ 3}\bar{\boldsymbol{X}}\right] - \mathbb{E}\left[(\boldsymbol{W}_2\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\right]\right\|_F \leq L_2\|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_2, \tag{189}$$

with a constant $L_2 \leq 9\theta(1-\theta) + \frac{3\theta^2}{\sqrt{n}}$.

Proof By (99) in proof B.2 of Proposition 10, we have

$$\mathbb{E}\left[(\boldsymbol{W}\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}\right] = 3p\theta(1-\theta)\boldsymbol{W}^{\circ 3} + 3p\theta^2\boldsymbol{W}, \quad \forall \boldsymbol{W} \in \mathsf{O}(n;\mathbb{R}). \tag{190}$$

Hence, we have:

$$\begin{aligned}
&\frac{1}{np}\left\|\mathbb{E}\left[(\boldsymbol{W}_1\bar{\boldsymbol{X}}^*)^{\circ 3}\bar{\boldsymbol{X}}\right] - \mathbb{E}\left[(\boldsymbol{W}_2\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\right]\right\|_F \\
= &\frac{1}{n}\left\|3\theta(1-\theta)(\boldsymbol{W}_1^{\circ 3} - \boldsymbol{W}_2^{\circ 3}) + 3\theta^2(\boldsymbol{W}_1 - \boldsymbol{W}_2))\right\|_F \\
\leq &\frac{3\theta(1-\theta)}{n}\left\|\boldsymbol{W}_1^{\circ 3} - \boldsymbol{W}_2^{\circ 3}\right\|_F + \frac{3\theta^2}{n}\|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_F \\
\leq &\frac{3\theta(1-\theta)}{n}\underbrace{\left\|\boldsymbol{W}_1^{\circ 3} - \boldsymbol{W}_2^{\circ 3}\right\|_F}_{\Gamma} + \frac{3\theta^2}{\sqrt{n}}\|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_2.
\end{aligned} \tag{191}$$

Note that

$$\begin{aligned}\Gamma =& \left\|\boldsymbol{W}_1^{\circ 3} - \boldsymbol{W}_2^{\circ 3}\right\|_F = \left\|(\boldsymbol{W}_1 - \boldsymbol{W}_2)(\boldsymbol{W}_1^{\circ 2} + \boldsymbol{W}_2^{\circ 2} + \boldsymbol{W}_1 \circ \boldsymbol{W}_2)\right\|_F \\ \leq& \left\|\boldsymbol{W}_1 - \boldsymbol{W}_2\right\|_2 \left\|\boldsymbol{W}_1^{\circ 2} + \boldsymbol{W}_2^{\circ 2} + \boldsymbol{W}_1 \circ \boldsymbol{W}_2\right\|_F \quad \text{(By inequality 2 in Lemma 33)} \\ \leq& (\left\|\boldsymbol{W}_1\right\|_F^2 + \left\|\boldsymbol{W}_2\right\|_F^2 + \left\|\boldsymbol{W}_1\right\|_F \left\|\boldsymbol{W}_2\right\|_F) \left\|\boldsymbol{W}_1 - \boldsymbol{W}_2\right\|_2 = 3n \left\|\boldsymbol{W}_1 - \boldsymbol{W}_2\right\|_2,\end{aligned} \tag{192}$$

therefore, we have:

$$\frac{1}{np} \left\|\mathbb{E}\big[(\boldsymbol{W}_1 \bar{\boldsymbol{X}}^*)^{\circ 3} \bar{\boldsymbol{X}}\big] - \mathbb{E}\big[(\boldsymbol{W}_2 \bar{\boldsymbol{X}})^{\circ 3} \bar{\boldsymbol{X}}^*\big]\right\|_F \leq \left(9\theta(1-\theta) + \frac{3\theta^2}{\sqrt{n}}\right) \left\|\boldsymbol{W}_1 - \boldsymbol{W}_2\right\|_2, \tag{193}$$

which completes the proof. ■

D.7. High Order Moment Bound of Bernoulli Gaussian Random Variables

Lemma 42 (Second Moment of $\|\cdot\|_4^4$) *Assume that $\boldsymbol{W} \in \mathsf{O}(n;\mathbb{R})$,$\boldsymbol{x} \in \mathbb{R}^n$, $x_i \sim_{iid} BG(\theta)$, $\forall i \in [n]$, then the second order moment of $\|\boldsymbol{W}\boldsymbol{x}\|_4^4$ satisfies*

$$\mathbb{E}\left\|\boldsymbol{W}\boldsymbol{x}\right\|_4^8 \leq Cn^2\theta, \tag{194}$$

for a constant $C > 105$.

Proof Assume $\boldsymbol{v}$ is a Gaussian vector and the support for Bernoulli-Gaussian vector $\boldsymbol{x}$ is $\mathcal{S}$, that is, $\forall i \in [n]$,

$$x_i = \begin{cases} v,\ v \sim \mathcal{N}(0,1) & \text{if } i \in \mathcal{S}, \\ 0 & \text{otherwise.} \end{cases} \tag{195}$$

Let $\boldsymbol{P}_{\mathcal{S}} : \mathbb{R}^n \mapsto \mathbb{R}^n$ be the projection onto set $\mathcal{S}$, that is, $\forall \boldsymbol{q} \in \mathbb{R}^n$

$$\big(\boldsymbol{P}_{\mathcal{S}}\boldsymbol{q}\big)_i = \begin{cases} q_i & \text{if } i \in \mathcal{S}, \\ 0 & \text{otherwise.} \end{cases} \tag{196}$$

Let $\boldsymbol{w}_i$ denote the i^{th} row vector of $\boldsymbol{W}$, so

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{W}\boldsymbol{x}\right\|_4^8 =& \mathbb{E}\Big[\big(\left\|\boldsymbol{W}\boldsymbol{x}\right\|_4^4\big)^2\Big] = \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n \left\langle \boldsymbol{w}_i, \boldsymbol{x}\right\rangle^4 \left\langle \boldsymbol{w}_j, \boldsymbol{x}\right\rangle^4\right] \\ =& \mathbb{E}_{\mathcal{S}} \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\Big[\left\langle \boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i, \boldsymbol{v}\right\rangle^4 \left\langle \boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_j, \boldsymbol{v}\right\rangle^4\Big] \\ \leq& \mathbb{E}_{\mathcal{S}} \sum_{i=1}^n \sum_{j=1}^n \Big[\mathbb{E}\left\langle \boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i, \boldsymbol{v}\right\rangle^8 \mathbb{E}\left\langle \boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_j, \boldsymbol{v}\right\rangle^8\Big]^{\frac{1}{2}} \quad \text{(Cauchy−Schwarz).}\end{aligned} \tag{197}$$

Since $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, so

$$\left\langle \boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i, \boldsymbol{v}\right\rangle = \sum_{k=1}^n \big(\boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i\big)_k v_k \sim \mathcal{N}(0, \left\|\boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i\right\|_2^2). \tag{198}$$

Therefore,

$$\mathbb{E}\left\langle \boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i, \boldsymbol{v}\right\rangle^8 = 105\left\|\boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i\right\|_2^8. \tag{199}$$

Hence, combine (197), we have

$$\begin{aligned}
\mathbb{E}\left\|\boldsymbol{W}\boldsymbol{x}\right\|_4^8 &\le \mathbb{E}_{\mathcal{S}}\sum_{i=1}^n\sum_{j=1}^n\left[\mathbb{E}\left\langle \boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i, \boldsymbol{v}\right\rangle^8\mathbb{E}\left\langle \boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_j, \boldsymbol{v}\right\rangle^8\right]^{\frac{1}{2}}\\
&= 105\mathbb{E}_{\mathcal{S}}\sum_{i=1}^n\sum_{j=1}^n\left\|\boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i\right\|_2^4\left\|\boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_j\right\|_2^4\\
&= 105\sum_{i=1}^n\sum_{j=1}^n\sum_{k_1,k_2,k_3,k_4}\mathbb{E}_{\mathcal{S}}\left[w_{i,k_1}^2\mathbb{1}_{k_1\in\mathcal{S}}w_{i,k_2}^2\mathbb{1}_{k_2\in\mathcal{S}}w_{j,k_3}^2\mathbb{1}_{k_3\in\mathcal{S}}w_{j,k_4}^2\mathbb{1}_{k_4\in\mathcal{S}}\right].
\end{aligned} \tag{200}$$

Now we discuss these four different cases separately:

- With probability $c_1\theta^4$ ($c_1 \le 1$), all $k_1, k_2, k_3, k_4 \in \mathcal{S}$, in this case, we have:

$$\begin{aligned}
&\sum_{i=1}^n\sum_{j=1}^n\sum_{k_1,k_2,k_3,k_4}\mathbb{E}_{\mathcal{S}}\left[w_{i,k_1}^2\mathbb{1}_{k_1\in\mathcal{S}}w_{i,k_2}^2\mathbb{1}_{k_2\in\mathcal{S}}w_{j,k_3}^2\mathbb{1}_{k_3\in\mathcal{S}}w_{j,k_4}^2\mathbb{1}_{k_4\in\mathcal{S}}\right]\\
=&\sum_{i=1}^n\sum_{j=1}^n\sum_{k_1,k_2,k_3,k_4}\mathbb{E}_{\mathcal{S}}\left[w_{i,k_1}^2w_{i,k_2}^2w_{j,k_3}^2w_{j,k_4}^2\right] = \sum_{i=1}^n\sum_{j=1}^n\sum_{k_1,k_2,k_3,k_4}w_{i,k_1}^2w_{i,k_2}^2w_{j,k_3}^2w_{j,k_4}^2\\
=&\sum_{i=1}^n\sum_{j=1}^n\sum_{k_1,k_2,k_3}w_{i,k_1}^2w_{i,k_2}^2w_{j,k_3}^2 = \sum_{i=1}^n\sum_{j=1}^n\sum_{k_1,k_2}w_{i,k_1}^2w_{i,k_2}^2 = n^2.
\end{aligned} \tag{201}$$

- With probability $c_2\theta^3$ ($c_2 \le 1$), only three among $k_1, k_2, k_3, k_4 \in \mathcal{S}$, in this case, we have:

$$\begin{aligned}
&\sum_{i=1}^n\sum_{j=1}^n\sum_{k_1,k_2,k_3,k_4}\mathbb{E}_{\mathcal{S}}\left[w_{i,k_1}^2\mathbb{1}_{k_1\in\mathcal{S}}w_{i,k_2}^2\mathbb{1}_{k_2\in\mathcal{S}}w_{j,k_3}^2\mathbb{1}_{k_3\in\mathcal{S}}w_{j,k_4}^2\mathbb{1}_{k_4\in\mathcal{S}}\right]\\
=&\sum_{i=1}^n\sum_{j=1}^n\sum_{k_1,k_2,k_3}\left[w_{i,k_1}^2w_{i,k_2}^2w_{j,k_3}^4\right] + \sum_{i=1}^n\sum_{j=1}^n\sum_{k_1,k_2,k_3}\left[w_{i,k_1}^2w_{i,k_2}^2w_{j,k_2}^2w_{j,k_3}^2\right] = n\left\|\boldsymbol{W}\right\|_4^4 + 1.
\end{aligned} \tag{202}$$

- With probability $c_3\theta^2$ ($c_3 \le 1$), only two among $k_1, k_2, k_3, k_4 \in \mathcal{S}$, in this case, we have:

$$\begin{aligned}
&\sum_{i=1}^n\sum_{j=1}^n\sum_{k_1,k_2,k_3,k_4}\mathbb{E}_{\mathcal{S}}\left[w_{i,k_1}^2\mathbb{1}_{k_1\in\mathcal{S}}w_{i,k_2}^2\mathbb{1}_{k_2\in\mathcal{S}}w_{j,k_3}^2\mathbb{1}_{k_3\in\mathcal{S}}w_{j,k_4}^2\mathbb{1}_{k_4\in\mathcal{S}}\right]\\
=&\sum_{i=1}^n\sum_{j=1}^n\sum_{k_1,k_2}w_{i,k_1}^4w_{i,k_2}^4 + \sum_{i=1}^n\sum_{j=1}^n\sum_{k_1,k_2}w_{i,k_1}^4w_{j,k_1}^2w_{j,k_2}^2 + \sum_{i=1}^n\sum_{j=1}^n\sum_{k_1,k_2}w_{i,k_1}^2w_{i,k_2}^2w_{j,k_1}^2w_{j,k_2}^2\\
\le&3\sum_{i=1}^n\sum_{j=1}^n\sum_{k_1,k_2}w_{i,k_1}^4w_{i,k_2}^4 = 3\left\|\boldsymbol{W}\right\|_4^8.
\end{aligned} \tag{203}$$

The only inequality above is achieved by Rearrangement inequality.

- With probability $c_4\theta$ ($c_4 \leq 1$), only one among $k_1, k_2, k_3, k_4 \in \mathcal{S}$, in this case, we have:

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k_1,k_2,k_3,k_4}\mathbb{E}_{\mathcal{S}}\left[w_{i,k_1}^2\mathbb{1}_{k_1\in\mathcal{S}}w_{i,k_2}^2\mathbb{1}_{k_2\in\mathcal{S}}w_{j,k_3}^2\mathbb{1}_{k_3\in\mathcal{S}}w_{j,k_4}^2\mathbb{1}_{k_4\in\mathcal{S}}\right]\\ =&\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k_1}w_{i,k_1}^4w_{j,k_1}^4\leq\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k_1}w_{i,k_1}^8=n\left\|\boldsymbol{W}\right\|_8^8.\end{aligned}\tag{204}$$

The only inequality above is achieved by Rearrangement inequality.

Substitute (201), (202), (203), and (204) into (200), yields

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{W}\boldsymbol{x}\right\|_4^8&\leq 105\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k_1,k_2,k_3,k_4}\mathbb{E}_{\mathcal{S}}\left[w_{i,k_1}^2\mathbb{1}_{k_1\in\mathcal{S}}w_{i,k_2}^2\mathbb{1}_{k_2\in\mathcal{S}}w_{j,k_3}^2\mathbb{1}_{k_3\in\mathcal{S}}w_{j,k_4}^2\mathbb{1}_{k_4\in\mathcal{S}}\right]\\ &\leq C(\theta^4n^2+\theta^3n\left\|\boldsymbol{W}\right\|_4^4+\theta^3+3\theta^2\left\|\boldsymbol{W}\right\|_4^8+\theta n\left\|\boldsymbol{W}\right\|_8^8)\leq Cn^2\theta,\end{aligned}\tag{205}$$

for a constant $C > 105$, which completes the proof. ■

Lemma 43 (Second Moment of $\nabla_{\boldsymbol{A}}\hat{f}(\boldsymbol{A},\boldsymbol{Y})$) *Assume that $\boldsymbol{W} \in \mathsf{O}(n;\mathbb{R})$,$\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $x_{i,j} \sim_{iid} BG(\theta)$, $\forall i \in [n], j \in [p]$, then each element of $\{(\boldsymbol{W}\boldsymbol{X}_o)^{\circ 3}\boldsymbol{X}_o^*\}_{i,j'}$ satisfies*

1. *$\{(\boldsymbol{W}\boldsymbol{X}_o)^{\circ 3}\boldsymbol{X}_o^*\}_{i,j'}$ can be represented as sum of p i.i.d random variables z_j*

$$\{(\boldsymbol{W}\boldsymbol{X}_o)^{\circ 3}\boldsymbol{X}_o^*\}_{i,j'}=\sum_{j=1}^{p}z_j,\quad\forall i,j'\in[n].\tag{206}$$

2. *The second moment of z_j is bounded by $C\theta$*

$$\mathbb{E}z_j^2\leq C,\quad\forall j\in[p].\tag{207}$$

for a constant $C > 177$.

Proof Assume $\boldsymbol{v}$ is a Gaussian vector and the support for Bernoulli-Gaussian vector $\boldsymbol{x}$ is $\mathcal{S}$, that is, $\forall i \in [n]$,

$$x_i=\begin{cases}v,\ v\sim\mathcal{N}(0,1) & \text{if } i\in\mathcal{S},\\ 0 & \text{otherwise.}\end{cases}\tag{208}$$

Let $\boldsymbol{P}_{\mathcal{S}} : \mathbb{R}^n \mapsto \mathbb{R}^n$ be the projection onto set $\mathcal{S}$, that is, $\forall \boldsymbol{q} \in \mathbb{R}^n$

$$\left(\boldsymbol{P}_{\mathcal{S}}\boldsymbol{q}\right)_i=\begin{cases}q_i & \text{if } i\in\mathcal{S},\\ 0 & \text{otherwise.}\end{cases}\tag{209}$$

By (95) and (96) in proof B.2 of Proposition 10, we know that

$$\frac{1}{4}\nabla_{\boldsymbol{A}}\hat{f}(\boldsymbol{A},\boldsymbol{Y})=(\boldsymbol{A}\boldsymbol{Y})^{\circ 3}\boldsymbol{Y}^*=(\boldsymbol{W}\boldsymbol{X}_o)^{\circ 3}\boldsymbol{X}_o^*\boldsymbol{D}_o^*,\tag{210}$$

and

$$\{(\boldsymbol{W}\boldsymbol{X}_o)^{\circ 3}\boldsymbol{X}_o^*\}_{i,j'} = \sum_{j=1}^{p}\Big[x_{j',j}\Big(\sum_{k=1}^{n} w_{i,k}x_{k,j}\Big)^3\Big] = \sum_{j=1}^{p}\Big[x_{j',j}\left\langle \boldsymbol{w}_i, \boldsymbol{x}_j\right\rangle^3\Big], \tag{211}$$

where $\boldsymbol{w}_i$ is the i^{th} row vector of $\boldsymbol{W}$. Notice that $x_{i,j}$ are independent with each other, if we define p random variables $z_1, z_2, \ldots, z_p$ as the following

$$z_j = x_{j',j}\left\langle \boldsymbol{w}_i, \boldsymbol{x}_j\right\rangle, \forall j \in [p], \tag{212}$$

then we can view $\{(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\}_{i,j'}$ as the mean of p i.i.d. random variables z_j. Notice that

$$\mathbb{E}(z_j^2) = \mathbb{E}\Big[x_{j',j}^2\Big(\sum_{k=1}^{n} w_{i,k}x_{k,j}\Big)^6\Big] \le \Big(\mathbb{E}x_{j',j}^4\Big)^{\frac{1}{2}}\Big(\mathbb{E}\left\langle \boldsymbol{w}_i, \boldsymbol{x}_j\right\rangle^{12}\Big)^{\frac{1}{2}} = \sqrt{3}\theta^{\frac{1}{2}}\Big(\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\boldsymbol{v}}\left\langle \boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i, \boldsymbol{v}\right\rangle^{12}\Big)^{\frac{1}{2}}, \tag{213}$$

and

$$\left\langle \boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i, \boldsymbol{v}\right\rangle = \sum_{k=1}^{n}\left(\boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i\right)_k v_k \sim \mathcal{N}(0, \|\boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i\|_2^2). \tag{214}$$

So we have

$$\Big(\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\boldsymbol{v}}\left\langle \boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i, \boldsymbol{v}\right\rangle^{12}\Big)^{\frac{1}{2}} = \Big(11!!\mathbb{E}_{\mathcal{S}}\|\boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i\|_2^{12}\Big)^{\frac{1}{2}} = \sqrt{11!!}\Big(\mathbb{E}_{\mathcal{S}}\|\boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i\|_2^{12}\Big)^{\frac{1}{2}}. \tag{215}$$

Follow the same pipe line (201), (202), (203), and (204) in Lemma 42, one can show that

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\|\boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i\|_2^{12} &= \sum_{k_1,k_2,\ldots,k_6} w_{i,k_1}^2\mathbb{1}_{k_1\in\mathcal{S}}w_{i,k_2}^2\mathbb{1}_{k_2\in\mathcal{S}}w_{i,k_3}^2\mathbb{1}_{k_3\in\mathcal{S}}w_{i,k_4}^2\mathbb{1}_{k_4\in\mathcal{S}}w_{i,k_5}^2\mathbb{1}_{k_5\in\mathcal{S}}w_{i,k_6}^2\mathbb{1}_{k_6\in\mathcal{S}}\\ &\le C'(\theta^6+\theta^5+\theta^4+\theta^3+\theta^2+\theta) \le C''\theta,\end{aligned} \tag{216}$$

for some constant $C', C'' > 1$. Therefore, combine (213), (215), and (216), we have

$$\mathbb{E}(z_j^2) \le \sqrt{3}\theta^{\frac{1}{2}}\Big(\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\boldsymbol{v}}\left\langle \boldsymbol{P}_{\mathcal{S}}\boldsymbol{w}_i, \boldsymbol{v}\right\rangle^{12}\Big)^{\frac{1}{2}} \le \sqrt{3\times 11!!\theta^2} \le C\theta, \tag{217}$$

for a constant $C > 177$, which completes the proof. ■

D.8. Union Tail Concentration Bound

Lemma 44 (Union Tail Concentration Bound of $(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*$) *If $\boldsymbol{X} \in \mathbb{R}^{n\times p}, x_{i,j} \sim_{iid} BG(\theta)$, the following inequality holds*

$$\begin{aligned}&\mathbb{P}\left(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})}\frac{1}{np}\left\|(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^* - \mathbb{E}\left[(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\right]\right\|_F \ge \delta\right)\\ \le& 2n^2\exp\left(-\frac{3p\delta^2}{c_1\theta + 8n^{\frac{3}{2}}(\ln p)^4\delta} + n^2\ln\Big(\frac{48np(\ln p)^4}{\delta}\Big)\right) + 2np\theta\exp\left(-\frac{(\ln p)^2}{2}\right),\end{aligned} \tag{218}$$

for a constant $c_1 > 1.7\times 10^4$.

Proof Let $\bar{\boldsymbol{X}} \in \mathbb{R}^{n\times p}$ denote the truncated $\boldsymbol{X}$ by bound B

$$\bar{x}_{i,j} = \begin{cases} x_{i,j} & \text{if} \quad |x_{i,j}| \le B, \\ 0 & \text{else.} \end{cases} \tag{219}$$

Note that $\bar{\boldsymbol{X}} = \boldsymbol{X}$ holds whenever $\|\boldsymbol{X}\|_\infty \le B$, and by Lemma 34, we know that with probability $\|\boldsymbol{X}\|_\infty \le B$ happens with probability at least $1-2np\theta\exp(-B^2/2)$. So we know that $\bar{\boldsymbol{X}} \neq \boldsymbol{X}$ holds with probability at most $2np\theta\exp(-B^2/2)$, and thus

$$\begin{aligned} &\mathbb{P}\bigg(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})} \frac{1}{np}\big\|(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^* - \mathbb{E}\big[(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big]\big\|_F > \delta\bigg) \\ \le &\mathbb{P}\bigg(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})} \frac{1}{np}\big\|(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^* - \mathbb{E}\big[(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big]\big\|_F > \delta, \boldsymbol{X} = \bar{\boldsymbol{X}}\bigg) + \mathbb{P}\Big(\boldsymbol{X} \neq \bar{\boldsymbol{X}}\Big) \\ \le &\mathbb{P}\bigg(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})} \frac{1}{np}\big\|(\boldsymbol{W}\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - \mathbb{E}\big[(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big]\big\|_F > \delta\bigg) + 2np\theta e^{-\frac{B^2}{2}} \end{aligned} \tag{220}$$

$\varepsilon-$net Covering. For any positive ε satisfy

$$\varepsilon \le \frac{\delta}{8npB^4}, \tag{221}$$

Lemma 35 shows there exists an $\varepsilon-$nets

$$\mathcal{S}_\varepsilon = \{\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_{|\mathcal{S}_\varepsilon|}\}, \tag{222}$$

which covers $\mathsf{O}(n;\mathbb{R})$

$$\mathsf{O}(n;\mathbb{R}) \subset \bigcup_{l=1}^{|\mathcal{S}_\varepsilon|} \mathbb{B}(\boldsymbol{W}_l, \varepsilon), \tag{223}$$

in operator norm $\|\cdot\|_2$. Moreover, we have

$$|\mathcal{S}_\varepsilon| \le \Big(\frac{6}{\varepsilon}\Big)^{n^2}. \tag{224}$$

So $\forall \boldsymbol{W} \in \mathsf{O}(n;\mathbb{R})$, there exists $l \in [|\mathcal{S}_\varepsilon|]$, such that $\|\boldsymbol{W}-\boldsymbol{W}_l\|_2 \le \varepsilon$. Thus, we have:

$$
\begin{aligned}
&\mathbb{P}\left(\frac{1}{np}\left\|(\boldsymbol{W}\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - \mathbb{E}[(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*]\right\|_F \le \delta\right)\\
\le&\mathbb{P}\Bigg(\frac{1}{np}\Big(\left\|(\boldsymbol{W}\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - (\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\right\|_F + \left\|(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - \mathbb{E}[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*]\right\|_F\\
&+\left\|\mathbb{E}[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*] - \mathbb{E}[(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*]\right\|_F\Big) \ge \delta\Bigg)\\
\le&\mathbb{P}\left(\frac{1}{np}\left\|(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - \mathbb{E}[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*]\right\|_F\right.\\
&\left.+\left(3npB^4 + 9\theta(1-\theta) + \frac{3\theta^2}{\sqrt{n}}\right)\|\boldsymbol{W}-\boldsymbol{W}_l\|_2 \ge \delta\right) \quad \text{(By Lemma 40, 41)}\\
\le&\mathbb{P}\left(\frac{1}{np}\left\|(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - \mathbb{E}[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*]\right\|_F + 4npB^4\varepsilon \ge \delta\right) \quad (p, B \text{ are large, } \|\boldsymbol{W}-\boldsymbol{W}_l\|_2 \le \varepsilon)\\
\le&\mathbb{P}\left(\frac{1}{np}\left\|(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - \mathbb{E}[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*]\right\|_F + \frac{\delta}{2} \ge \delta\right) \quad \left(\text{We assume } \varepsilon \le \frac{\delta}{8npB^4}\right)\\
=&P\left(\frac{1}{np}\left\|(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}} - \mathbb{E}[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*]\right\|_F \ge \frac{\delta}{2}\right).
\end{aligned}
\tag{225}
$$

Analysis. For random variable $\bar{\boldsymbol{X}}$, we have

$$
\begin{aligned}
&\left\|\mathbb{E}\big[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big] - \mathbb{E}\big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\big]\right\|_F = \left\|\mathbb{E}\big[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big] - \mathbb{E}\big[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\cdot \mathbb{1}_{\{\|\boldsymbol{X}\|_\infty \le B\}}\big]\right\|_F\\
=&\left\|\mathbb{E}\big[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\cdot \mathbb{1}_{\{\|\boldsymbol{X}\|_\infty > B\}}\big]\right\|_F = \sqrt{\sum_{\substack{1\le i\le n\\ 1\le j'\le n}}\left(\mathbb{E}\Big\{[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*]_{i,j'}\cdot \mathbb{1}_{\{\|\boldsymbol{X}\|_\infty > B\}}\Big\}\right)^2}\\
\le&\sqrt{\sum_{\substack{1\le i\le n\\ 1\le j'\le n}}\left(\Big\{\mathbb{E}\big[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big]^2_{i,j'}\Big\}\Big\{\mathbb{E}\mathbb{1}_{\{\|\boldsymbol{X}\|_\infty > B\}}\Big\}\right)} = \left\|\mathbb{E}\big[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big]\right\|_F\cdot\sqrt{\mathbb{E}\mathbb{1}_{\{\|\boldsymbol{X}\|_\infty > B\}}}.
\end{aligned}
\tag{226}
$$

Note that in (99) of Lemma B.2, we know that

$$
\mathbb{E}\big[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big] = 3p\theta(1-\theta)\boldsymbol{W}_l^{\circ 3} + 3p\theta^2\boldsymbol{W}_l,
\tag{227}
$$

hence

$$
\begin{aligned}
\left\|\mathbb{E}\big[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big]\right\|_F =& \left\|3p\theta(1-\theta)\boldsymbol{W}_l^{\circ 3} + 3p\theta^2\boldsymbol{W}_l\right\|_F \le 3p\theta(1-\theta)\left\|\boldsymbol{W}_l^{\circ 3}\right\|_F + 3p\theta^2\|\boldsymbol{W}_l\|_F\\
\le&3p\theta(1-\theta)\|\boldsymbol{W}_l\|_F^3 + 3p\theta^2\|\boldsymbol{W}_l\|_F \quad \text{(By inequality 1 in Lemma 33)}\\
=&3p\theta(1-\theta)n^{\frac{3}{2}} + 3p\theta^2 n^{\frac{1}{2}} < 4n^{\frac{3}{2}}p\theta \quad (n \text{ is a large number}).
\end{aligned}
\tag{228}
$$

Moreover, Lemma 34 shows that

$$\mathbb{E}\mathbb{1}_{\{\|\boldsymbol{X}\|_\infty > B\}} \le 2np\theta^{-\frac{B^2}{2}}. \tag{229}$$

Substitute (229) and (228) into (226), yield

$$\frac{1}{np}\left\|\mathbb{E}\big[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big] - \mathbb{E}\big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\big]\right\|_F \le 4\sqrt{2}np^{\frac{1}{2}}\theta^{\frac{3}{2}}e^{-\frac{B^2}{4}}. \tag{230}$$

Hence, when

$$B \ge 2\sqrt{\ln\left(\frac{16\sqrt{2}np^{\frac{1}{2}}\theta^{\frac{3}{2}}}{\delta}\right)}, \tag{231}$$

we have

$$\frac{1}{np}\left\|\mathbb{E}\big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\big] - \mathbb{E}\big[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big]\right\|_F \le 4\sqrt{2}np^{\frac{1}{2}}\theta^{\frac{3}{2}}e^{-\frac{B^2}{4}} \le \frac{\delta}{4}. \tag{232}$$

Therefore, combine (225), we have

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{np}\left\|(\boldsymbol{W}\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - \mathbb{E}[(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*]\right\|_F \le \delta\right)\\
\le&\mathbb{P}\left(\frac{1}{np}\left\|(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - \mathbb{E}[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*]\right\|_F \ge \frac{\delta}{2}\right)\\
\le&\mathbb{P}\left(\frac{1}{np}\left\|(\boldsymbol{W}\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - \mathbb{E}\big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\big]\right\|_F + \frac{1}{np}\left\|\mathbb{E}\big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\big] - \mathbb{E}\big[(\boldsymbol{W}_l\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big]\right\|_F \ge \frac{\delta}{2}\right)\\
\le&\mathbb{P}\left(\frac{1}{np}\left\|(\boldsymbol{W}\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - \mathbb{E}\big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\big]\right\|_F \ge \frac{\delta}{4}\right) \quad \text{(By (232))}\\
\le&n^2\mathbb{P}\left(\left|\Big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\Big]_{i,j'} - \mathbb{E}\Big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\Big]_{i,j'}\right| \ge p\cdot\frac{\delta}{4}\right) \quad \text{(By union bound, } \forall i,j' \in [n]\text{).}
\end{aligned} \tag{233}$$

Point-wise Bernstein's Inequality. Next, we apply Bernstein's inequality on $\big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\big]_{i,j'}$. Note that for each $i,j' \in [n]$

$$\big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\big]_{i,j'} = \sum_{j=1}^{p}\left[\bar{x}_{j',j}\Big(\sum_{k=1}^{n} w_{i,k}\bar{x}_{k,j}\Big)^3\right], \tag{234}$$

can be viewed as sum of p independent variables

$$\bar{z}_j = \bar{x}_{j',j}\Big(\sum_{k=1}^{n} w_{i,j}\bar{x}_{k,j}\Big)^3 = \bar{x}_{j',j}\left\langle \boldsymbol{w}_i, \bar{\boldsymbol{x}}_j\right\rangle^3, \quad \forall j \in [p], \tag{235}$$

where $\boldsymbol{w}_i$ is the i^{th} row vector of $\boldsymbol{W}$. Note that each $\bar{z}_j$ are bounded by

$$\left|\bar{x}_{j',j}\left\langle \boldsymbol{w}_i, \bar{\boldsymbol{x}}_j\right\rangle^3\right| \le B\left|\left\langle \boldsymbol{w}_i, \bar{\boldsymbol{x}}_j\right\rangle^3\right| = B\left\|\bar{\boldsymbol{x}}_j\right\|_2^3\left|\left\langle \boldsymbol{w}_i, \frac{\bar{\boldsymbol{x}}_j}{\|\bar{\boldsymbol{x}}_j\|_2}\right\rangle\right| \le B\cdot(nB^2)^{\frac{3}{2}} = n^{\frac{3}{2}}B^4, \quad (236)$$

also for each $\bar{x}_{j',j}\Big(\sum_{k=1}^n w_{i,j}\bar{x}_{k,j}\Big)^3$, we have

$$\begin{aligned}
&\mathbb{E}\Big[x_{j',j}^2\Big(\sum_{k=1}^n w_{i,j}x_{k,j}\Big)^6\Big] - \mathbb{E}\Big[\bar{x}_{j',j}^2\Big(\sum_{k=1}^n w_{i,j}\bar{x}_{k,j}\Big)^6\Big]\\
=&\mathbb{E}\Big[x_{j',j}^2\Big(\sum_{k=1}^n w_{i,j}x_{k,j}\Big)^6 - \bar{x}_{j',j}^2\Big(\sum_{k=1}^n w_{i,j}\bar{x}_{k,j}\Big)^6\Big] \qquad (237)\\
=&\mathbb{E}\Big[x_{j',j}^2\Big(\sum_{k=1}^n w_{i,j}x_{k,j}\Big)^6\cdot\mathbb{1}_{\{\|\boldsymbol{X}\|_\infty>B\}}\Big] \ge 0,
\end{aligned}$$

and (217) in Lemma 43 shows that

$$\mathbb{E}\Big[\bar{x}_{j',j}^2\Big(\sum_{k=1}^n w_{i,j}\bar{x}_{k,j}\Big)^6\Big] \le \mathbb{E}\Big[x_{j',j}^2\Big(\sum_{k=1}^n w_{i,j}x_{k,j}\Big)^6\Big] \le C, \quad (238)$$

for a constant $C > 177$. Hence, by Bernstein's inequality, $\forall i, j' \in [n]$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\Big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\Big]_{i,j'} - \mathbb{E}\Big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\Big]_{i,j'}\right| \ge p\cdot\frac{\delta}{4}\right)\\
=&\mathbb{P}\left(\left|\sum_{j=1}^p\Big[\bar{x}_{j',j}(\sum_{k=1}^n w_{i,k}\bar{x}_{k,j})^3\Big] - \sum_{j=1}^p\mathbb{E}\Big[\bar{x}_{j',j}(\sum_{k=1}^n w_{i,k}\bar{x}_{k,j})^3\Big]\right| \ge p\cdot\frac{\delta}{4}\right)\\
=&\mathbb{P}\left(\sum_{j=1}^p\Big[\bar{x}_{j',j}(\sum_{k=1}^n w_{i,k}\bar{x}_{k,j})^3\Big] - \sum_{j=1}^p\mathbb{E}\Big[\bar{x}_{j',j}(\sum_{k=1}^n w_{i,k}\bar{x}_{k,j})^3\Big] \ge p\cdot\frac{\delta}{4}\right)\\
&+\mathbb{P}\left(\sum_{j=1}^p\Big[\bar{x}_{j',j}(\sum_{k=1}^n w_{i,k}\bar{x}_{k,j})^3\Big] - \sum_{j=1}^p\mathbb{E}\Big[\bar{x}_{j',j}(\sum_{k=1}^n w_{i,k}\bar{x}_{k,j})^3\Big] \le -p\cdot\frac{\delta}{4}\right) \qquad (239)\\
\le&2\exp\left(-\frac{p\delta^2/16}{2\big[\frac{1}{p}\sum_{j=1}^p\bar{x}_{j',j}^2(\sum_{k=1}^n w_{i,j}\bar{x}_{k,j})^6 + \frac{n^{\frac{3}{2}}B^4\delta}{12}\big]}\right) \quad \text{(By (236))}\\
\le&2\exp\left(-\frac{p\delta^2}{\frac{96}{p}\sum_{j=1}^p C\theta + 8\delta n^{\frac{3}{2}}B^4\delta}\right) \quad \text{(By (237))}\\
=&2\exp\left(-\frac{3p\delta^2}{c_1\theta + 8n^{\frac{3}{2}}B^4\delta}\right),
\end{aligned}$$

for a constant $c_1 > 1.7 \times 10^4$. Combine (233), we have

$$\begin{aligned}&\mathbb{P}\bigg(\frac{1}{np}\big\|(\boldsymbol{W}\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - \mathbb{E}[(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*]\big\|_F \le \delta\bigg)\\ \le& n^2\mathbb{P}\bigg(\bigg|\Big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\Big]_{i,j'} - \mathbb{E}\Big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\Big]_{i,j'}\bigg| \ge p\cdot\frac{\delta}{4}\bigg).\end{aligned} \tag{240}$$

Union Bound. Now, we will give a union bound for

$$\mathbb{P}\bigg(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})}\frac{1}{np}\big\|(\boldsymbol{W}\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - \mathbb{E}\big[(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big]\big\|_F > \delta\bigg). \tag{241}$$

Notice that

$$\begin{aligned}&\mathbb{P}\bigg(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})}\frac{1}{np}\big\|(\boldsymbol{W}\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - \mathbb{E}\big[(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big]\big\|_F > \delta\bigg)\\ \le&\sum_{l=1}^{|\mathcal{S}_\varepsilon|}\mathbb{P}\bigg(\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}_l,\varepsilon)}\frac{1}{np}\big\|(\boldsymbol{W}\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - \mathbb{E}\big[(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big]\big\|_F > \delta\bigg)\quad(\text{By }\varepsilon\text{−covering in (223)})\\ \le&\sum_{l=1}^{|\mathcal{S}_\varepsilon|}n^2\mathbb{P}\bigg(\bigg|\Big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\Big]_{i,j'} - \mathbb{E}\Big[(\boldsymbol{W}_l\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^*\Big]_{i,j'}\bigg| \ge p\cdot\frac{\delta}{4}\bigg)\quad(\text{By (240)})\\ \le&\sum_{l=1}^{|\mathcal{S}_\varepsilon|}\bigg[2n^2\exp\bigg(-\frac{3p\delta^2}{c_1\theta + 8n^{\frac{3}{2}}B^4\delta}\bigg)\bigg]\quad(\text{By (239)})\\ \le&\Big(\frac{6}{\varepsilon}\Big)^{n^2}\bigg[2n^2\exp\bigg(-\frac{3p\delta^2}{c_1\theta + 8n^{\frac{3}{2}}B^4\delta}\bigg)\bigg]\quad(\text{By (224)})\\ =&\exp\bigg(n^2\ln\Big(\frac{48npB^4}{\delta}\Big)\bigg)\bigg[2n^2\exp\bigg(-\frac{3p\delta^2}{c_1\theta + 8n^{\frac{3}{2}}B^4\delta}\bigg)\bigg]\quad(\text{Let }\varepsilon = \frac{\delta}{8npB^4})\\ =&2n^2\exp\bigg(-\frac{3p\delta^2}{c_1\theta + 8n^{\frac{3}{2}}B^4\delta} + n^2\ln\Big(\frac{48npB^4}{\delta}\Big)\bigg),\end{aligned} \tag{242}$$

for a constant $c_1 > 1.4 \times 10^4$. Note that (231) requires a lower bound on B, here we can choose $B = \ln p$, which satisfies (231) when p is large enough (say $p = \Omega(n)$). Combine (220) and substitute $B = \ln p$, we have

$$\begin{aligned}&\mathbb{P}\bigg(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})}\frac{1}{np}\big\|(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^* - \mathbb{E}\big[(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big]\big\|_F > \delta\bigg)\\ \le&\mathbb{P}\bigg(\sup_{\boldsymbol{W}\in\mathsf{O}(n;\mathbb{R})}\frac{1}{np}\big\|(\boldsymbol{W}\bar{\boldsymbol{X}})^{\circ 3}\bar{\boldsymbol{X}}^* - \mathbb{E}\big[(\boldsymbol{W}\boldsymbol{X})^{\circ 3}\boldsymbol{X}^*\big]\big\|_F > \delta\bigg) + 2np\theta e^{-\frac{B^2}{2}}\\ \le&2n^2\exp\bigg(-\frac{3p\delta^2}{c_1\theta + 8n^{\frac{3}{2}}(\ln p)^4\delta} + n^2\ln\Big(\frac{48np(\ln p)^4}{\delta}\Big)\bigg) + 2np\theta\exp\bigg(-\frac{(\ln p)^2}{2}\bigg),\end{aligned} \tag{243}$$

for a constant $c_1 > 1.7 \times 10^4$, which completes the proof. ■

D.9. Perturbation Bound for Unitary Polar Factor

Lemma 45 (Perturbation Bound for Unitary Polar Factor) *Let $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathbb{R}^{n\times n}$ be two nonsigular matrices, and let $\boldsymbol{Q}_1 = \boldsymbol{U}_1\boldsymbol{V}_1^*$, $\boldsymbol{Q}_2 = \boldsymbol{U}_2\boldsymbol{V}_2^*$, where*

$$\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^* = SVD(\boldsymbol{A}_1), \quad \boldsymbol{U}_2\boldsymbol{\Sigma}_2\boldsymbol{V}_2^* = SVD(\boldsymbol{A}_2).$$

Let $\sigma_n(\boldsymbol{A}_1)$, $\sigma_n(\boldsymbol{A}_2)$ denote the smallest singular value of $\boldsymbol{A}_1$,$\boldsymbol{A}_2$ respectively. Then, for any unitary invariant norm $\|\cdot\|_\diamond$ we have

$$\|\boldsymbol{Q}_1 - \boldsymbol{Q}_2\|_\diamond \leq \frac{2}{\sigma_n(\boldsymbol{A}_1) + \sigma_n(\boldsymbol{A}_2)} \|\boldsymbol{A}_1 - \boldsymbol{A}_2\|_\diamond . \tag{244}$$

Proof See theorem 1 in Li (1995). ■

D.10. Perturbation Bound for Singular Values

Lemma 46 (Singular Value Perturbation) *Let $\boldsymbol{G} = \boldsymbol{B}\boldsymbol{M}$ be a general full rank matrix, where $\boldsymbol{M}$ ($M_{i,i}$ equals to the ℓ^2-norm of the i^{th} column of $\boldsymbol{G}$) is a chosen diagonal matrix so $\boldsymbol{B}$ has unit matrix 2 norm ($\|\boldsymbol{B}\|_2 = 1$). Let $\delta\boldsymbol{G} = \delta\boldsymbol{B}\boldsymbol{M}$ be a perturbation of $\boldsymbol{G}$ such that $\|\delta\boldsymbol{B}\|_2 \leq \sigma_{\min}(\boldsymbol{B})$ and σ_i and σ_i' be the i^{th} singular value of $\boldsymbol{G}$ and $\boldsymbol{G}+\delta\boldsymbol{G}$ respectively. Then*

$$\frac{|\sigma_i - \sigma_i'|}{\sigma_i} \leq \frac{\|\delta\boldsymbol{B}\|_2}{\sigma_{\min}(\boldsymbol{B})}. \tag{245}$$

Proof See theorem 2.17 in Demmel and Veselić (1992). ■

References

P.-A. Absil and J. Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

A. Agarwal, A. Anandkumar, and P. Netrapalli. Exact recovery of sparsely used overcomplete dictionaries. *stat*, 1050:8–39, 2013.

A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used overcomplete dictionaries. In *Conference on Learning Theory*, pages 123–137, 2014.

M. Aharon, M. Elad, A. Bruckstein, et al. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11): 4311, 2006.

Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

S. Arora, R. Ge, and A. Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Conference on Learning Theory*, pages 779–806, 2014.

S. Arora, R. Ge, T. Ma, and A. Moitra. Simple, efficient, and neural algorithms for sparse coding. 2015.

Y. Bai, Q. Jiang, and J. Sun. Subgradient descent learns orthogonal dictionaries. *arXiv preprint arXiv:1810.10702*, 2018.

B. Barak, J. Kelner, and D. Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *STOC*, 2015.

D. P. Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48 (3):334–334, 1997.

E. Candes and T. Tao. Decoding by linear programming. *arXiv preprint math/0502327*, 2005.

E. J. Candès. Mathematics of sparsity (and a few other things). Citeseer, 2014.

Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *arXiv preprint arXiv:1809.09573*, 2018.

D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. Stochastic subgradient method converges on tame functions. *arXiv preprint arXiv:1804.07795*, 2018.

J. Demmel and K. Veselić. Jacobi's method is more accurate than qr. *SIAM Journal on Matrix Analysis and Applications*, 13(4):1204–1245, 1992.

D. L. Donoho. 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: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(6): 797–829, 2006.

S. S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018.

A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

Q. Geng and J. Wright. On the local correctness of $\ell 1$-minimization for dictionary learning. In *2014 IEEE International Symposium on Information Theory*, pages 3180–3184. IEEE, 2014.

D. Gilboa, S. Buchanan, and J. Wright. Efficient dictionary learning with gradient descent. *arXiv preprint arXiv:1809.10313*, 2018.

B. C. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*. Springer, 2 edition, 2015.

A. Hyvärinen. A family of fixed-point algorithms for independent component analysis. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3917–3920, 1997.

A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9:1483–1492, 1997.

R. Jenatton, J. Mairal, G. Obozinski, and F. R. Bach. Proximal methods for sparse hierarchical dictionary learning. In *ICML*, 2010.

M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(Feb): 517–553, 2010.

Y. Katznelson. *An introduction to harmonic analysis*. Cambridge University Press, 2004.

E. Kreyszig. *Introductory functional analysis with applications*, volume 1. wiley New York, 1978.

A. Krizhevsky, V. Nair, and G. Hinton. The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55, 2014.

H.-W. Kuo, Y. Lau, Y. Zhang, and J. Wright. Geometry and symmetry in short-and-sparse deconvolution. *arXiv preprint arXiv:1901.00256*, 2019.

Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

R.-C. Li. New perturbation bounds for the unitary polar factor. *SIAM Journal on Matrix Analysis and Applications*, 16(1):327–332, 1995.

Y. Li and Y. Bresler. Global geometry of multichannel sparse blind deconvolution on the sphere. In *Advances in Neural Information Processing Systems*, pages 1132–1143, 2018.

Y. Li and Y. Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

J.-H. Lu. A note on a theorem of Stanton-Weinstein on the L4-norm of spherical harmonics. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 102, page 561, 1987.

C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467*, 2017.

T. Ma, J. Shi, and D. Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 438–446. IEEE, 2016.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. Technical report, MINNESOTA UNIV MINNEAPOLIS INST FOR MATHEMATICS AND ITS APPLICATIONS, 2008.

J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach. Supervised dictionary learning. In *Advances in neural information processing systems*, pages 1033–1040, 2009.

J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):791–804, 2012.

J. Mairal, F. Bach, and J. Ponce. Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3):85–283, 2014. ISSN 1572-2740. doi: 10.1561/0600000058. URL `http://dx.doi.org/10.1561/0600000058`.

K. G. Murty and S. N. Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

A. V. Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.

Q. Qu, J. Sun, and J. Wright. Finding a sparse vector in a subspace: Linear sparsity using alternating directions. In *Advances in Neural Information Processing Systems*, pages 3401–3409, 2014.

Q. Qu, X. Li, and Z. Zhu. A nonconvex approach for exact and efficient multichannel sparse blind deconvolution. *arXiv preprint arXiv:1908.10776*, 2019.

M. Ranzato, Y. Boureau, and Y. LeCun. Sparse feature learning for deep belief networks. In *Advances in Neural Information Processing Systems (NIPS 2007)*, volume 20, 2007.

S. Ravishankar and Y. Bresler. l0 sparsifying transform learning with efficient optimal updates and convergence guarantees. 2015.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

R. Rubinstein, A. M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

T. Schramm and D. Steurer. Fast and robust tensor decomposition with applications to dictionary learning. *arXiv preprint arXiv:1706.08672*, 2017.

Y. Shen, Y. Xue, J. Zhang, K. B. Letaief, and V. Lau. Complete dictionary learning via *ell_p*-norm maximization. *arXiv preprint arXiv:2002.10043*, 2020.

D. A. Spielman, H. Wang, and J. Wright. Exact recovery of sparsely-used dictionaries. In *Conference on Learning Theory*, pages 37–1, 2012.

R. J. Stanton and A. Weinstein. On the L4 norm of spherical harmonics. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 89, page 343, 1981.

J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere. *arXiv preprint arXiv:1504.06785*, 2015.

M. Vetterli and J. Kovacevic. Wavelets and subband coding. Technical report, Prentice-Hall, 1995.

M. Vetterli, J. Kovačević, and V. K. Goyal. *Foundations of signal processing*. Cambridge University Press, 2014.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Y. Xue, Y. Shen, V. Lau, J. Zhang, and K. B. Letaief. Blind data detection in massive mimo via ℓ_3-norm maximization over the stiefel manifold. *arXiv preprint arXiv:2004.12301*, 2020.

J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

Y. Zhai, H. Mehta, Z. Zhou, and Y. Ma. Understanding l4-based dictionary learning: Interpretation, stability, and robustness. In *International Conference on Learning Representations*, 2019.

Y. Zhang, H.-W. Kuo, and J. Wright. Structured local optima in sparse blind deconvolution. *arXiv preprint arXiv:1806.00338*, 2018.