UNDERSTANDING $\ell^4$-BASED DICTIONARY LEARNING: INTERPRETATION, STABILITY, AND ROBUSTNESS

Yuexiang Zhai$^{1,2,*}$ Hermish Mehta$^1$ Zhengyuan Zhou$^3$ Yi Ma$^1$

$^1$Department of EECS, UC Berkeley $^2$ByteDance Inc. $^3$Stern School of Business, NYU

ABSTRACT

Recently, the $\ell^4$-norm maximization has been proposed to solve the sparse dictionary learning (SDL) problem. The simple MSP (matching, stretching, and projection) algorithm proposed by Zhai et al. (2019a) has shown to be surprisingly efficient and effective. This paper aims to better understand this algorithm from its strong geometric and statistical connections with the classic PCA and ICA, as well as their associated fixed-point style algorithms. Such connections provide a unified way of viewing problems that pursue principal, independent, or sparse components of high-dimensional data. Our studies reveal additional good properties of $\ell^4$-maximization: not only is the MSP algorithm for sparse coding insensitive to small noise, but it is also robust to outliers and resilient to sparse corruptions. We provide statistical justification for such inherently nice properties. To corroborate the theoretical analysis, we also provide extensive and compelling experimental evidence with both synthetic data and real images.

1 INTRODUCTION

The explosion of massive amounts of high-dimensional data has become the modern-day norm for a large number of scientific and engineering disciplines and hence presents a daunting challenge for both computation and learning. Rising to this challenge, sparse dictionary learning (SDL) provides a potent framework in representation learning that exploits the blessing of dimensionality: real data tends to lie in or near some low-dimensional subspaces or manifolds, even though the ambient dimension is often extremely large (e.g. the number of raw pixels in an image). More specifically, SDL (Olshausen & Field (1997); Mairal et al. (2008; 2012; 2014); Spielman et al. (2012); Sun et al. (2015); Bai et al. (2018); Qu et al. (2019)) concerns the problem of learning a compact, sparse representation from raw, unlabelled data: given a data matrix $Y = [y_1, y_2, \ldots, y_p] \in \mathbb{R}^{n \times p}$ that contains $p$ $n$-dimensional samples, one aims to find a linear transformation (i.e. a dictionary) $D \in \mathbb{R}^{n \times m}$ and an associated maximally sparse representation $X = [x_1, x_2, \ldots, x_p] \in \mathbb{R}^{m \times p}$ that satisfies

$$Y = DX. \tag{1}$$

As the data matrix $Y$ can represent a variety of signals (e.g. images, audios, languages, and genetics etc) in practical applications, SDL provides a versatile structure-seeking formulation that has found widespread applications in computational neuroscience, image processing, computer vision, and machine learning at large (Olshausen & Field, 1996; 1997; Argyriou et al., 2008; Ranzato, 2007; Elad & Aharon, 2006; Wright et al., 2008; Yang et al., 2010; Zhang et al., 2013; Mairal et al., 2014; Zhang et al., 2014; 2019).

Related Works. Motivated by this practical significance, there has been a growing surge of interest recently (e.g. Rambhatla et al. (2019); Bai et al. (2018); Gilboa et al. (2018); Nguyen et al. (2018); Chatterji & Bartlett (2017); Mensch et al. (2016)) that aims to tackle SDL. In attempts to recover the sparse signals $X$, these existing work adopt an $\ell^0$- or $\ell^1$-penalty function to promote the underlying sparsity and give various optimization algorithms for the resulting objectives (some of those are heuristics while a few others have theoretical convergence guarantees). Although these penalty

*$^*$Work done when interning at ByteDance. YZ would like to thank Chong You at UC Berkeley for meaningful discussions. Correspondence to: Yuexiang Zhai, ysz@berkeley.edu or Yi Ma, yima@eecs.berkeley.edu. Codes are available at https://github.com/hermish/ZMZM-ICLR-2020.
functions are indeed sparsity-promoting, the resulting optimization problems must be solved one row at a time, hence resulting as many optimization problems as the ambient dimension $n$. Consequently, $l^0$ or $l^1$-based objectives result only in local methods (i.e. cannot yield the entire solution at once) and hence entail prohibitive computational burden. Another prominent approach in SDL is Sum-of-Squares (SOS), proposed by and articulated in a series of recent work (Barak et al. (2015); Ma et al. (2016); Schramm & Steurer (2017)). The key idea there is to utilize the properties of higher order SOS polynomials to correctly recover one column of the dictionary at a time, for which there are $m$ columns in total. However, the computational complexity of these recovery methods are quasi-polynomial, thus again resulting in large computational expense.

Very recently, in the complete dictionary learning\footnote{Complete dictionary learning requires the learned dictionary $D$ in equation 1 to be square and invertible.} setting, a novel global approach has been suggested in Zhai et al. (2019a,b) that presents a formulation which can efficiently recover the sparse signal matrix $X$ once and for all. In particular, Zhai et al. (2019b) has shown that if the generative model for $Y = D_o X_o \in \mathbb{R}^{n \times m}$ satisfies that $D_o \in \text{O}(n; \mathbb{R})$ is orthonormal and $X_o \in \mathbb{R}^{n \times m}$ is Bernoulli-Gaussian sparse,\footnote{Each entry $x_{i,j}$ of $X$ can be represented as the product of a Bernoulli variable and a normal Gaussian variable: $x_{i,j} = \Omega_{i,j} V_{i,j}$, where $\Omega_{i,j} \sim_{\text{iid}} \text{Ber}(\theta)$ and $V_{i,j} \sim_{\text{iid}} \mathcal{N}(0, 1)$, similar for vectors or scalars. This is the standard setting adopted in Spielman et al. (2012); Sun et al. (2015); Bai et al. (2018).} then maximizing the $\ell^4$-norm\footnote{We abuse the notation a bit, by denoting $\| \cdot \|_4$ as the sum of element-wise $4^{th}$ power of all entries of a vector and matrix, that is, for any $a \in \mathbb{R}^n$, $\|a\|_4 = \sum_{i=1}^n a_i^4$ and for any $A \in \mathbb{R}^{n \times m}$, $\|A\|_4 = \sum_{i,j} a_{i,j}^4$.} of $AY$ over $\text{O}(n; \mathbb{R})$:  

$$\max_A \frac{1}{4} \|AY\|_4^4 \text{ subject to } A \in \text{O}(n; \mathbb{R}) \quad \text{(or } AA^* = I),$$

is able to find the ground truth dictionary $D_o$ up to an arbitrary signed permutation. Moreover, Zhai et al. (2019b) has proposed the simple “Matching, Stretching, and Projection” (MSP) algorithm, which is shown to be experimentally efficient and effective, for solving the program in equation 2:

$$\text{MSP: } A_{t+1} = \mathcal{P}_{\text{O}(n; \mathbb{R})}[((A_t Y)^{\odot 3} Y^*)]^t = U_t V_t^*,$$

where $U_t V_t^*$ are from the singular value decomposition: $U_t \Sigma_t V_t^* = \text{SVD}[[A_t Y]^{\odot 3} Y^*]$.\footnote{For any symmetric matrix $S \in \mathbb{R}^{n \times n}$ and an orthogonal matrix $A \in \text{O}(n; \mathbb{R})$, the projection of $SA$ onto the orthogonal group is $A: \mathcal{P}_{\text{O}(n; \mathbb{R})}[SA] = A$, one may see Absil & Malick (2012) for details.}

In this paper, we here give an alternative (arguably simpler and more revealing) derivation of the MSP algorithm (3). Consider the Lagrangian formulation of the constrained optimization problem given in equation 2. The necessary condition of critical points, $\nabla_A \frac{1}{4} \|AY\|_4^4 = \nabla_A (A A^* - I)$ for some Lagrangian multipliers $\Lambda$, implies:

$$(AY)^{\odot 3} Y^* = (A + A^*) A.$$

As the optimization is over the orthogonal group $\text{O}(n; \mathbb{R})$, restricting the condition in equation 4 onto the orthogonal group yields a necessary condition for any critical point $A$:

$$\mathcal{P}_{\text{O}(n; \mathbb{R})}[(AY)^{\odot 3} Y^*] = A.$$ 

Hence the critical point $A$ can be viewed as a “fixed point” of the map: $\mathcal{P}_{\text{O}(n; \mathbb{R})}[(\cdot Y)^{\odot 3} Y^*]$ from $\text{O}(n; \mathbb{R})$ to itself. The MSP algorithm in equation 3 is to find the fixed point(s) of this map. Notice that the orthonormal constraint $A \in \text{O}(n; \mathbb{R})$ in equation 2 can be viewed as enforcing the orthogonality of $n$ unit vectors simultaneously. So, more flexibly and generally, one may choose to compute any $k$, for $1 \leq k \leq n$, leading orthonormal bases of $D_o$ by solving the program:

$$\max_W \frac{1}{4} \|W^* Y\|_4^4 \text{ subject to } W \in \text{St}(n, k; \mathbb{R}) \subset \mathbb{R}^{n \times k},$$

where $\text{St}(n, k; \mathbb{R})$ is the Stiefel manifold.\footnote{For any $1 \leq k \leq n$, $\text{St}(n, k; \mathbb{R}) \triangleq \{ W \in \mathbb{R}^{n \times k} : W^* W = I_k \}$.} The orthogonal group $\text{O}(n; \mathbb{R})$ and the unit sphere $S^{n-1}$ can be viewed as two special cases of the Stiefel manifold $\text{St}(n, k; \mathbb{R})$, with $k = n$ and $k = 1$, respectively. In some specific tasks such as dictionary learning and blind deconvolution, optimization over the unit sphere has been widely practiced, such as in Sun et al. (2015); Bai et al. (2018); Zhang et al. (2018); Kuo et al. (2019). The more general setting of maximizing a convex function over any compact set also has been studied by Journee et al. (2010) in the context of sparse PCA, which has provided convergence guarantees for this class of programs.
Our Contributions. Our contributions are twofold. First, by taking a suitable analytical angle, we reveal novel geometric and statistical connections between PCA, ICA, and the $\ell^4$-norm maximization based SDL. We then show that algorithm-wise, the fixed-point type MSP algorithm for $\ell^4$-norm maximization has the same nature as the classic power-iteration method for PCA (Jolliffe, 2011) and the FastICA algorithm for ICA (Hyvärinen & Oja, 1997). This interpretation gives a unified view for problems that pursue principal, independent, or sparse components from high-dimensional data and enriches our understanding of low-dimensional structure recovery frameworks, classical and new, at both formulation and algorithmic fronts.

Second, and more importantly from a practical perspective, we examine how MSP performs under a variety of more realistic conditions, when the measurements $Y$ could be contaminated with noise, outliers, or sparse corruptions. We show that, similar to PCA, $\ell^4$-norm maximization and the MSP algorithm are inherently stable to small noise. Somewhat surprisingly though, unlike PCA, the MSP algorithm is further robust to outliers and resilient to sparse gross errors! We provide characterizations of these desirable properties of MSP. The claims are further corroborated with extensive experiments on both synthetic data and real images. Taken as a whole, our results contribute to the broad landscape of dictionary learning by affirming that $\ell^4$-maximization based SDL and the corresponding global algorithm MSP provide a valuable toolkit to the existing literature.

2 SDL VERSUS PCA AND ICA

2.1 Pursuit of Principal, Independent, or Sparse Components

Relation with the Geometric Interpretation of PCA. For a data matrix $Y \in \mathbb{R}^{n \times p}$, Principal Component Analysis (PCA), aiming to find the top (top $k$) left singular vector (vectors) of $Y$,

$$
\max_\mathbf{W} \frac{1}{2} \| \mathbf{W}^T \mathbf{Y} \|_2^2 \quad \text{subject to } \mathbf{W} \in \text{St}(n, k; \mathbb{R}),
$$

(7)

can be understood as finding a direction (a $k$-dimensional subspace) in row($Y$) in which $Y$ has the largest $\ell^2$-norm (Frobenius norm). For instance, finding the direction with the largest $\ell^2$-norm over the unit sphere can be viewed as calculating the spectral norm (or the largest singular value), of matrix $Y$. In comparison, we may view equation 6:

$$
\max_\mathbf{W} \frac{1}{4} \| \mathbf{W}^T \mathbf{Y} \|_4^4 \quad \text{subject to } \mathbf{W} \in \text{St}(n, k; \mathbb{R})
$$

as finding a direction, or a $k$-dimensional subspace, in row($Y$) where the projection of $Y$ has the largest $\ell^4$-norm. For instance, finding the direction with the largest $\ell^4$-norm over the unit sphere (equation 6) can be viewed as calculating the induced $\| \cdot \|_{2,4}$ norm of matrix $Y$: $\| Y \|_{2,4} \equiv \max_{\mathbf{a} \in \mathbb{S}^{n-1}} \| \mathbf{a}^T \mathbf{Y} \|_4$.

Relation with the Statistical Interpretation of PCA. View each column $y_j, j \in [p]$ of data matrix $Y \in \mathbb{R}^{n \times p}$ as an $n$ dimensional random vector that is i.i.d. drawn from a distribution of random variable $y$ and let $Y_c$ denote the centered $Y$: $Y_c \equiv Y[I - 1_p11^*]$, where $1 \in \mathbb{R}^p$ is a vector of all 1’s. Then, finding the top $k$ principal components of $Y_c$: $\max_{\mathbf{W} \in \text{St}(n, k; \mathbb{R})} \frac{1}{2} \| \mathbf{W}^T \mathbf{Y_c} \|_2^2$ is to find $k$ uncorrelated projections of $y \in \mathbb{R}^n$ that has the top $k$ sample variance (i.e. 2nd order moment) (Jolliffe, 2011; Helwig, 2017). Similar to PCA, the $\ell^4$-norm maximization of centered data matrix $Y_c$: $\max_{\mathbf{W} \in \text{St}(n, k; \mathbb{R})} \frac{1}{4} \| \mathbf{W}^T \mathbf{Y_c} \|_4^4$ can be viewed as finding $k$ uncorrelated projections of $y$ that have the top $k$ sample 4th order moment, whose statistical meaning is better revealed below.

Relation with ICA and Nonnormality. The $\ell^4$-norm maximization over the Stiefel manifold is strongly related to finding the maximal or minimal kurtosis in Independent Component Analysis (ICA) (Hyvärinen & Oja, 1997; 2000). In order to identify one component of a given random vector $y \in \mathbb{R}^n$, ICA aims to find a unit vector (a direction) $w \in \mathbb{S}^{n-1}$ that maximizes or minimizes the kurtosis of $w^*y$, defined as:

$$
\text{kurt}(w^*y) = \mathbb{E} (w^*y)^4 - 3 \| w \|_4^4.
$$

(8)

Kurtosis is widely used for evaluating the nonnormality of a random variable; see DeCarlo (1997); Hyvärinen & Oja (1997; 2000). According to Huber (1985), the nonnormality of data carries “abnormal” hence interesting information in real data for many applications (e.g. Lee et al. (2003); Cain
et al. (2017)). Thus, extracting the 4th order moment helps understand such statistics of real datasets (Hyvärinen et al., 2009) and even their topology (Carlsson, 2009). One may also find that the ℓ4-maximization based dictionary learning formulation is similar to maximizing kurtosis (equation 8) with spherical constraint $w \in S^{n-1}$, in fact, these two formulations are exactly the same if one only wants to find one column of the dictionary. Intuitively, such coincidence occurs due to the fact that maximizing ℓ4-norm or kurtosis have the same effect—they both promote the “spikiness” (Zhang et al. (2018); Li & Bresler (2018)) (or “peak” (DeCarlo, 1997)) of a distribution. More rigorous analysis regarding the similarity between ℓ4-based dictionary learning and ICA is beyond the scope of this paper and we leave them for future research.

2.2 Fixed-Point Style Algorithms

In optimization, the ℓ4-norm maximization in equation 6 over the Stiefel manifold $St(k, n; \mathbb{R})$ is a special type of nonconvex optimization problem—convex maximization over a compact set. Although Journé et al. (2010); Zhai et al. (2019b) have shown that the MSP algorithm is guaranteed to find critical points, the experiments in Zhai et al. (2019b) suggest that the MSP algorithm finds global maxima of the ℓ4-norm efficiently and effectively. For better understanding, in this section we illustrate some striking similarities between the MSP algorithm and the “power-iteration” type algorithms for solving PCA as well as ICA.

**Fixed-point Perspective of Power Iteration.** For a general data matrix $Y \in \mathbb{R}^{n \times p}$, finding the top singular value of $Y$ is equivalent to solving the following optimization problem:

$$
\max_w \varphi(w) = \frac{1}{2} \|w^* Y\|_2^2 \quad \text{subject to} \quad w \in S^{n-1}.
$$

(9)

For this constrained optimization, the Lagrangian multiplier method gives the necessary condition: $\nabla w \varphi(w) = Y Y^* w = \lambda w$, similar to equation 4. If we restrict this condition onto the sphere, we obtain the fixed point condition $w = \mathcal{P}_{S^{n-1}} [\nabla w \varphi(w)]$. The classic power-iteration method

$$
w_{t+1} = \mathcal{P}_{S^{n-1}} [\nabla w \varphi(w_t)] = \frac{Y Y^* w_t}{\|Y Y^* w_t\|_2},
$$

(10)

is precisely computing this fixed point, which is arguably the most efficient and widely used algorithm to solve equation 9, for PCA (or computing SVD of $Y$).

**Fixed-point Perspective of FastICA.** In order to maximize (or minimize) the kurtosis over $S^{n-1}$,

$$
\max_w \psi(w) = \frac{1}{4} \text{kurt}(w^* y) = \frac{1}{4} E [w^* y]^4 - \frac{3}{4} \|w\|_2^4 \quad \text{subject to} \quad w \in S^{n-1},
$$

(11)

Hyvärinen & Oja (1997) has proposed the following fixed-point type iteration:

$$
w_{t+1} = \mathcal{P}_{S^{n-1}} [\nabla w \psi(w_t)] = \frac{E [y (y^* w_t)^3] - 3 \|w_t\|_2^2 w_t}{\|E [y (y^* w_t)^3] - 3 \|w_t\|_2^2 w_t\|_2},
$$

(12)

which enjoys cubic (at least quadratic) rate of convergence, under the ICA model assumption.

**Fixed-point Perspective of MSP.** For the ℓ4-norm maximization program:

$$
\max_W \phi(W) = \frac{1}{4} \|W^* Y\|_4^4 \quad \text{subject to} \quad W \in St(n, k; \mathbb{R}),
$$

through a similar derivation to that in Section 1, one can show that the MSP iteration in equation 5 for the orthogonal group generalizes to the Stiefel manifold case as:

$$
W_{t+1} = \mathcal{P}_{St(n, k; \mathbb{R})} [\nabla w \phi(W_t)] = U_t V_t^*,
$$

(13)

where $U_t, \Sigma_t, V_t^* = \text{SVD} [Y (Y^* W_t)^{<3}]$. The above iteration has the same nature as the power iteration in equation 10 and equation 12, since they all solve a fixed-point type problem, by projecting gradient of the objective function $\nabla \varphi(\cdot), \nabla \phi(\cdot), \nabla \psi(\cdot)$ onto the constraint manifold $S^{n-1}$ and $St(n, k; \mathbb{R})$, respectively. Table 1 summarizes these striking similarities.
3 STABILITY AND ROBUSTNESS OF $\ell^4$-NORM MAXIMIZATION

Even though the MSP algorithm for $\ell^4$-norm maximization is similar to power-iteration for PCA, in real applications, PCA often requires modifications to improve its robustness (Candes et al., 2011; Xu et al., 2010; 2012). In this section, we want to examine the stability and robustness of the $\ell^4$-maximization framework. Assume our clean observation matrix $Y \in \mathbb{R}^{n \times p}$ is produced by the product of a ground truth orthogonal dictionary $D_o$ and a Bernoulli-Gaussian matrix $X_o \in \mathbb{R}^{n \times p}$:

$$Y = D_oX_o, \quad D_o \in O(n; \mathbb{R}), \quad \{X_o\}_{i,j} \sim iid \mathrm{BG}(	heta).$$

(14)

Now let us assume we only observe different types of imperfect measurements of $Y$:

Noisy Measurements: $Y_N := Y + G$, where $G \in \mathbb{R}^{n \times p}$ is matrix that satisfies $g_{i,j} \sim iid \mathcal{N}(0, \eta^2)$ and $\eta > 0$ controls the variance of the noise.

Measurements with Outliers: $Y_O := [Y, G']$, where $Y_O$ contains extra columns $(G' \in \mathbb{R}^{n \times \tau p})^6$ generated from an independent Gaussian process $g'_{i,j} \sim iid \mathcal{N}(0, 1)$. Here, $\tau$ controls the portion of the outliers w.r.t. the clean data size $p$.

Measurements with Sparse Corruptions: $Y_C := Y + \sigma B \circ S$, where $\sigma > 0$ controls the scale of corrupting entries, $B \in \mathbb{R}^{n \times p}$ is a Bernoulli matrix so $b_{i,j} \sim iid \text{Ber}(\beta)$ with $\beta \in (0, 1)$ controlling the ratio of the sparse corruptions, and $S \in \mathbb{R}^{n \times p}$ has entries $s_{i,j}$ drawn i.i.d. from a Rademacher distribution:

$$s_{i,j} = \begin{cases} 1 & \text{with probability 1/2} \\ -1 & \text{with probability 1/2} \end{cases}.$$  

(15)

3.2 STATISTICAL ANALYSIS AND JUSTIFICATION

The analysis for the stability and robustness of the $\ell^4$-norm maximization follows similar statistical analysis techniques in Zhai et al. (2019b) to show that the global maximum of

$$W_* \in \arg \max_W \mathbb{E} \|W^* Y_o\|_4^4, \quad \text{subject to } W \in O(n; \mathbb{R})$$

(16)

satisfies $W_* D_o \in SP(n)$.\(^8\) Note here $Y_o$ denotes different imperfect measurements: noisy ($Y_N$), with outliers ($Y_O$), and with sparse corruptions ($Y_C$). Below we provide the expectation and concentration results for $\|W^* Y_o\|_4^4$ over the data matrix. We show that $\mathbb{E} \|W^* Y_o\|_4^4$ is largely determined by $\|W D_o\|_4^4$, a quantity that indicates a “distance” from $W^* D_o$ to $SP(n)$. As shown in Lemma 2.3 and Lemma 2.4 in Zhai et al. (2019b), the only global maximizers of $\|W^* D_o\|_4^4$ are signed permutation matrices, and $W^* D_o$ converges to a signed permutation matrix as $\|W^* D_o\|_4^4$ reaches its global maximum.

---

\(^8\)In case $\tau p$ is not an integer, we round $\tau p$ to the closest integer.

\(^7\)In our context, $\tau = 1$ is already corruption of large magnitude, since the variance of the sparse signal is 1.

\(^6\)SP$(n)$ is the signed permutation group, a group of orthogonal matrices that only contain $0, \pm 1$.  

---

<table>
<thead>
<tr>
<th>Objectives</th>
<th>Constraint Sets</th>
<th>Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power Iteration $\varphi(w) \doteq \frac{1}{2} |w^T Y|_2^2$</td>
<td>$w \in S^{n-1}$</td>
<td>$w_{t+1} = P_{S^{n-1}}[\nabla w \varphi(w_t)]$</td>
</tr>
<tr>
<td>FastICA $\psi(w) \doteq \frac{1}{2} \kurt(w^T y)$</td>
<td>$w \in S^{n-1}$</td>
<td>$w_{t+1} = P_{S^{n-1}}[\nabla w \psi(w_t)]$</td>
</tr>
<tr>
<td>MSP $\phi(W) \doteq \frac{1}{2} |W^* Y|_4^4$</td>
<td>$W \in St(n, k; \mathbb{R})$</td>
<td>$W_{t+1} = P_{St(n, k; \mathbb{R})}[\nabla w \phi(W_t)]$</td>
</tr>
</tbody>
</table>

Table 1: Similarities among fixed-point algorithms for Power Iteration, FastICA, and MSP.
Proposition 3.1 (Expectation of Objective with Small Noise) \( \forall \theta \in (0,1), \) let \( X_\circ \in \mathbb{R}^{n \times p}, \) \( x_{i,j} \sim \text{iid } \mathcal{BG}(\theta). \) Let \( D_\circ \in \mathcal{O}(n; \mathbb{R}) \) be an orthogonal matrix and assume \( Y = D_\circ X_\circ. \) For any orthogonal matrix \( W \in \mathcal{O}(n; \mathbb{R}) \) and any random Gaussian matrix \( G \in \mathbb{R}^{n \times p}, g_{i,j} \sim \text{iid } \mathcal{N}(0, \eta^2) \) independent of \( X_\circ, \) let \( Y_N = Y + G \) denote the data with noise. Then the expectation of \( \frac{1}{np} \|W^* Y_N\|^4_4 \) satisfies:

\[
\frac{1}{np} \mathbb{E}_{X_\circ} \|W^* Y_N\|^4_4 = 3\theta (1 - \theta) \frac{\|W^* D_\circ\|^4_4}{n} + C_{\theta, \eta},
\]

where \( C_{\theta, \eta} \) is a constant which depends on \( \theta \) and \( \eta. \)

Proof See Appendix A.1.

Theorem 3.2 (Concentration of Objective with Small Noise) \( \forall \theta \in (0,1), \) let \( X_\circ \in \mathbb{R}^{n \times p}, \) \( x_{i,j} \sim \text{iid } \mathcal{BG}(\theta). \) Let \( D_\circ \in \mathcal{O}(n; \mathbb{R}) \) be an orthogonal matrix and assume \( Y = D_\circ X_\circ. \) For any orthogonal matrix \( W \in \mathcal{O}(n; \mathbb{R}) \) and any random Gaussian matrix \( G \in \mathbb{R}^{n \times p}, g_{i,j} \sim \text{iid } \mathcal{N}(0, \eta^2) \) independent of \( X_\circ, \) let \( Y_N = Y + G \) denote the input with noise, then:

\[
\mathbb{P} \left( \sup_{W \in \mathcal{O}(n; \mathbb{R})} \frac{1}{np} \left\|W^* Y_N\|^4_4 - \mathbb{E} \|W^* Y_N\|^4_4 \right\| \geq \delta \right) < \frac{1}{p},
\]

when \( p = \Omega \left((1 + \eta^2)^4 n^2 \ln n / \delta^2\right). \)

Proof See Appendix A.2.

Proposition 3.3 (Expectation of Objective with Outliers) \( \forall \theta \in (0,1), \) let \( X_\circ \in \mathbb{R}^{n \times p}, \) \( x_{i,j} \sim \text{iid } \mathcal{BG}(\theta). \) Let \( D_\circ \in \mathcal{O}(n; \mathbb{R}) \) be an orthogonal matrix and assume \( Y = D_\circ X_\circ. \) For any orthogonal matrix \( W \in \mathcal{O}(n; \mathbb{R}) \) and any random Gaussian matrix \( G' \in \mathbb{R}^{n \times p}, g_{i,j} \sim \text{iid } \mathcal{N}(0,1) \) independent of \( X_\circ, \) let \( Y_O = [Y, G'] \) denote the data with outliers \( G'. \) Then the expectation of \( \frac{1}{np} \|W^* Y_O\|^4_4 \) satisfies:

\[
\frac{1}{np} \mathbb{E}_{X_\circ, G'} \|W^* Y_O\|^4_4 = 3\theta (1 - \theta) \frac{\|W^* D_\circ\|^4_4}{n} + C_{\theta, \tau},
\]

where \( C_{\theta, \tau} \) is a constant which depends on \( \theta, \tau. \)

Proof See Appendix A.3

Theorem 3.4 (Concentration of Objective with Outliers) \( \forall \theta \in (0,1), \) let \( X_\circ \in \mathbb{R}^{n \times p}, \) \( x_{i,j} \sim \text{iid } \mathcal{BG}(\theta). \) Let \( D_\circ \in \mathcal{O}(n; \mathbb{R}) \) be an orthogonal matrix and assume \( Y = D_\circ X_\circ. \) For any orthogonal matrix \( W \in \mathcal{O}(n; \mathbb{R}) \) and any random Gaussian matrix \( G' \in \mathbb{R}^{n \times p}, g_{i,j} \sim \text{iid } \mathcal{N}(0,1) \) independent of \( X_\circ, \) let \( Y_O = [Y, G'] \) denote the input with outlier \( G', \) then:

\[
\mathbb{P} \left( \sup_{W \in \mathcal{O}(n; \mathbb{R})} \frac{1}{np} \left\|W^* Y_O\|^4_4 - \mathbb{E} \|W^* Y_O\|^4_4 \right\| \geq \delta \right) < \frac{1}{p},
\]

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

Proof See Appendix A.4.

In the above results, Proposition 3.1 and 3.3 reveal that both normalized \( \frac{1}{np} \mathbb{E} \|W^* Y_N\|^4_4, \) \( \frac{1}{np} \mathbb{E} \|W^* Y_O\|^4_4 \) are only determined by \( \|W^* D_\circ\|^4_4. \) Moreover, as shown in Theorem 3.2 and 3.4, when \( p \) is large enough (\( p = \Omega((1 + \eta^2)^4 n^2 \ln n), \Omega(\tau^2 n^2 \ln n) \) respectively), both \( \frac{1}{np} \|W^* Y_N\|^4_4, \) \( \frac{1}{np} \|W^* Y_O\|^4_4 \) concentrate around their expectation with high probability. Therefore, the \( \ell^4 \)-norm maximization formulation \( \|W^* Y_\circ\|^4_4 \) for dictionary learning is insensitive to dense Gaussian noise and robust to Gaussian outliers.\(^9\)

\(^9\)In the outlier case, Proposition 3.3 and Theorem 3.4 can be generalized to any rotation invariant distributions, e.g., uniform distribution on the sphere. More details regarding rotation invariant distributions can be found in Chapter 4 of Bryc (2012).
Proposition 3.5 (Expectation of Objective with Sparse Corruptions) \( \forall \theta \in (0, 1), \) let \( X_o \in \mathbb{R}^{n \times p}, x_{i,j} \sim_{iid} BG(\theta). \) Let \( D_o \in \mathcal{O}(n; \mathbb{R}) \) be an orthogonal matrix and assume \( Y = D_o X_o. \) For any orthogonal matrix \( W \in \mathcal{O}(n; \mathbb{R}) \) and any random Bernoulli matrix \( B \in \mathbb{R}^{n \times p}, b_{i,j} \sim_{iid} Ber(\beta) \) independent of \( X_o, \) let \( Y_C = Y + \sigma B \circ S \) denote the data with sparse corruptions, and \( S \in \mathbb{R}^{n \times p} \) is defined in equation 15. Then the expectation of \( \frac{1}{np} \| W^* Y_C \|_4^4 \) satisfies:

\[
\frac{1}{np} E_{X_o, B, S} \| W^* Y_C \|_4^4 = 3\theta(1 - \theta)\frac{\| W^* D_o \|_4^4}{n} + \sigma^4(1 - 3\beta)\frac{\| W \|_4^4}{n} + C_{\theta, \sigma, \beta},
\]

where \( C_{\theta, \sigma, \beta} \) is a constant depending on \( \theta, \sigma \) and \( \beta. \)

**Proof** See Appendix A.5.

Theorem 3.6 (Concentration of Objective with Sparse Corruptions) \( \forall \theta \in (0, 1), \) let \( X_o \in \mathbb{R}^{n \times p}, x_{i,j} \sim_{iid} BG(\theta). \) Let \( D_o \in \mathcal{O}(n; \mathbb{R}) \) be an orthogonal matrix and assume \( Y = D_o X_o. \) For any orthogonal matrix \( W \in \mathcal{O}(n; \mathbb{R}) \) and any random Bernoulli matrix \( B \in \mathbb{R}^{n \times p}, b_{i,j} \sim_{iid} Ber(\beta) \) independent of \( X_o, \) let \( Y_C = Y + \sigma B \circ S \) denote the input with sparse corruptions, and \( S \in \mathbb{R}^{n \times p} \) is defined in equation 15, then:

\[
P \left( \sup_{W \in \mathcal{O}(n; \mathbb{R})} \frac{1}{np} \| W^* Y_C \|_4^4 - \mathbb{E} \| W^* Y_C \|_4^4 \geq \delta \right) < \frac{1}{p},
\]

when \( p = \Omega \left( \sigma^8 \beta^2 n^2 \ln n/\delta^2 \right). \)

**Proof** See Appendix A.6.

Unlike the cases with noise and outliers, Proposition 3.5 and Theorem 3.6 indicate that \( \frac{1}{np} \mathbb{E} \| W^* Y_C \|_4^4 \) depends on both \( \| W^* D_o \|_4^4 \) and \( \| W \|_4^4; \) when \( p = \Omega(\sigma^8 \beta^2 n^2 \ln n), \) the objective \( \frac{1}{np} \| W^* Y_C \|_4^4 \) concentrates around this expectation with high probability. Nevertheless, when the magnitude of \( \sigma^4(1 - 3\beta) \) is significantly smaller than \( 3\theta(1 - \theta), \) the landscape of the objective \( \| W^* Y_C \|_4^4 \) would largely be determined by \( \| W^* D_o \|_4^4 \) only. As shown in Figure 1, this is indeed the case whenever: (a) the sparsity level \( \theta \) of ground truth signal \( X_o, \) is “reasonably” small (neither diminishing to 0 nor larger than 0.5); (b) \( \beta, \) the sparsity level of the corruption, is small (smaller than 0.5); (c) \( \sigma, \) the magnitude of the sparse errors, is not significantly larger than the intrinsic variance of the sparse signal (the intrinsic variance of the model is 1).

Therefore, besides the insensitivity to small noise and robustness to outliers, the \( \ell^4 \)-maximization based dictionary learning also shows resilience to sparse corruptions under reasonable conditions.

4 SIMULATIONS AND EXPERIMENTS

4.1 QUANTITATIVE EVALUATION: SIMULATIONS ON SYNTHETIC DATA

**Single Trial of MSP.** In this simulation, we run the MSP algorithm from equation 3, using the imperfect measurements \( Y_C \) of different models \( (Y_N, Y_O, Y_C). \) As shown in Figure 2, the normalized value of \( \| W^* D_o \|_4^4 / n \) reaches global maximum with all types of inputs when varying the level of noise, outliers, and sparse corruptions. Moreover, as the scale of imperfect measurements increase, Figure 2 shows that (a) the iterations for convergence increases and (b) the final objective value \( \| W^* D_o \|_4^4 \) decreases almost negligibly. This numerical experiment suggests that the MSP algorithm is able to identify the ground truth orthogonal transformation \( D_o \) despite different types of imperfect measurements.

**Phase Transition.** Next, we conduct extensive simulations to study the relation between recovery accuracy and sample size \( p. \) We run the experiments by increasing the sample size \( p \) w.r.t. the scale of imperfect measurements \( \eta, \tau, \beta, \) respectively. As shown in Figure 3, the MSP algorithm (3)
Published as a conference paper at ICLR 2020

Figure 2: Normalized $\|W^*D_o\|_4^4/n$ of the MSP algorithm for dictionary learning, using imperfect measurements $Y_N, Y_O, Y_C$, respectively.

demonstrates a clear phase transition behavior w.r.t. noise, outliers, and sparse corruptions. Such phenomena suggest that the algorithm is inherently stable and robust to certain amounts of noise, outliers, and sparse corruptions. The results corroborate our concentration results in Theorem 3.2, 3.4 and 3.6—a larger sample size $p$ increases the accuracy and robustness of the MSP algorithm (3) for all types of nuisances.

Figure 3: Average normalized error $|1 - \|W^*D_o\|_4^4/n|$ of 10 random trials for the MSP algorithm: (a) Varying sample size $p$ and variance of noise $\eta^2$; (b) Varying sample size $p$ and Gaussian Outlier ratio $\tau$; (c) Varying sample size $p$ and sparse corruption ratio $\beta$, with fixed $\sigma = 1$.

Comparison with Prior Arts. We also compare the MSP algorithm (Zhai et al., 2019b) with previous complete dictionary learning algorithms: Subgradient (SG) (Bai et al., 2018) and Riemannian Trust Region (RTR) methods (Sun et al., 2015). While SG and RTR demonstrate slightly better accuracy in some cases (e.g. small Gaussian noise and Gaussian outliers), both algorithms appear unstable to sparse corruptions. Meanwhile, the MSP algorithm is stable to all imperfect measurement models mentioned in this paper and runs significantly faster than the others.

Table 2: Comparison of the MSP algorithm (Zhai et al., 2019b) with prior complete dictionary learning algorithms: Subgradient method (Bai et al., 2018) and Riemannian Trust Region (Sun et al., 2015) methods in different models with fixed ground truth sparsity $\theta = 0.3$. Note that SG only learns a unit vector each time and does not guarantee orthogonality; we therefore project the dictionary learned to $O(n; \mathbb{R})$ for fair comparison.
4.2 Qualitative Evaluation: Experiments on Real Images and Patches

Besides simulations, we also conduct extensive experiments to verify the stability and robustness of the MSP algorithm with real imagery data, at both image level and patch level. Throughout these experiments, rather than visualize all bases, we routinely show the top bases learned—heuristically, the top bases are those with the largest coefficients (here, in terms of $\ell^1$-norm).

**Image Level.** At the image level, we first vectorize all 60,000 images in the MNIST dataset (LeCun et al., 1998) into a single matrix $Y \in \mathbb{R}^{784 \times 60,000}$, then create imperfect measurements based on models specified in Section 3: $Y_N$ (MNIST with noise), $Y_O$ (MNIST with outliers), $Y_C$ (MNIST with sparse corruptions). We run the MSP algorithm 3 with $Y, Y_N, Y_O, Y_C$ and compare the learned bases. Figure 4(a), (c), (e), and (g) show examples of $Y, Y_N, Y_O, Y_C$, and Figure 4(b), (d), (f), and (h) show top 10 bases learned from $Y, Y_N, Y_C, Y_O$, respectively. Despite that we use different types of imperfect measurements of MNIST, the top bases learned from MSP algorithm 3 are very much the same. This result corroborates our analysis: the $\ell^1$-maximization and the MSP algorithm is inherently insensitive to noise, robust to outliers, and resilient to sparse corruptions.

![Examples of MNIST and its different imperfect measurements.](image)

**Patch Level.** A classic application of dictionary learning involves learning sparse representations of image patches (Elad & Aharon, 2006; Mairal et al., 2007). In this section, we extend the experiments of Zhai et al. (2019b) to learn patches from gray scale and color images. First, we construct a data matrix $Y$ by vectorizing each $8 \times 8$ patch from the 512 × 512 gray scale image, “Barbara” (see Figure 5). We then run the MSP algorithm with 100 iterations on both $Y$ and a noisy version $Y_N$, and the learned top bases are visualized in Figure 5.

![The top 12 bases learned from all 16 × 16 patches of Barbara, both with (b) and without (a) noise. The noisy image is produced by adding Gaussian noise to the clean image, resulting in a signal-to-noise ratio (SNR) of 5.87.](image)

We next consider the problem of learning a “global dictionary” (Mairal et al., 2007) for patches from many different images. To construct our data matrix $Y$, we randomly sample 100,000 $8 \times 8 \times 3$ patches with opposite intensity are considered as the same base, since they only differ by a sign.
Published as a conference paper at ICLR 2020

Figure 6: The top 12 bases learned from all $8 \times 8 \times 3$ color patches of the clean and noisy image, respectively. Here, the SNR of the noisy image is 6.56.

Figure 7: Top half (96) bases learned from 100,000 random $8 \times 8 \times 3$ patches sampled from CIFAR-10, before and after adding Gaussian noise, with SNR 6.23.

patches from the CIFAR-10 data-set (Krizhevsky et al., 2009). A noisy data matrix $Y_N$ is then generated by adding Gaussian noise. Again, we apply the MSP algorithm with 200 iterations to learn 192 bases and visualize the results in Figure 7. We leave the experiments of CIFAR-10 with outliers and sparse corruptions in the Appendix due to limited space.

In each of these experiments, the top bases in the learned dictionary remain relatively unchanged with the addition of noise. To quantify this similarity, we take the top bases from the noisy dictionary and find the closest top-20 clean bases for each. If the bases are nearly identical, then the inner product of each of these pairs should be close to 1. Table 3 reports the statistics.

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Lower Quartile</th>
<th>Median</th>
<th>Upper Quartile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barbara</td>
<td>0.3048</td>
<td>0.8471</td>
<td>0.9941</td>
<td>0.9993</td>
<td>1.0000</td>
</tr>
<tr>
<td>Duck</td>
<td>0.2510</td>
<td>0.9782</td>
<td>0.9891</td>
<td>0.9971</td>
<td>1.0000</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>0.5147</td>
<td>0.7203</td>
<td>0.9892</td>
<td>0.9998</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Table 3: Statistics about the inner products between the top 20 noisy bases and their corresponding closest top-20 clean bases.

5 CONCLUSION AND DISCUSSIONS

In this paper, we find the $\ell^4$-norm maximization based dictionary learning and corresponding MSP algorithm introduced by Zhai et al. (2019b) have strong geometric and statistical connections to classic data analysis methods PCA and ICA. These connections seem to be the reason why they all admit similar simple and efficient algorithms. Empirically, we have observed that $\ell^4$-norm maximization is surprisingly insensitive to noise, robust to outliers, and resilient to sparse corruptions. Moreover, such empirical observations corroborate our concentration analysis—larger data samples ($p$) improve the stability and robustness of the $\ell^4$-maximization based dictionary learning.

From experiments on real images, we observed that top bases learned are rather stable but tail bases can be less stable (see Figure 10 in Appendix C). We conjecture this phenomenon happens due to the fact that real images generally do not follow the uniformly sparse Bernoulli Gaussian model (of equation 14). Generalizing dictionary learning to non-uniformly sparsely generated data would be a good topic for future study. Finally, note that Bai et al. (2018) has established subgradient descent can find global optimal solutions for dictionary learning, an interesting theoretical guarantee since dictionary learning problems do not satisfy the standard structural (i.e. generalized convexity) assumptions that guarantee global convergence (Ben-Tal & Nemirovski, 2001; Zhou et al., 2017a; Nesterov, 2013; Zhou et al., 2017b; Bubeck et al., 2015; Zhou et al., 2017c; 2018; Ma et al., 2018; Chi et al., 2019). It would be desirable, although highly challenging, to establish similar global convergence guarantees for our MSP algorithm. We leave that for future work.
REFERENCES


Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Peter W Glynn, Yinyu Ye, Li-Jia Li, and Fei-Fei Li. Distributed asynchronous optimization with unbounded delays: How slow can you go? 2018.
A PROOF OF SECTION 3

A.1 PROOF OF PROPOSITION 3.1

Claim A.1 (Expectation of Objective with Small Noise) \( \forall \theta \in (0, 1), \) let \( X_0 \in \mathbb{R}^{n \times p}, x_{i,j} \sim_{iid} \mathcal{N}(0, \Sigma) \). Let \( D_o \in O(n; \mathbb{R}) \) be an orthogonal matrix and assume \( Y = D_o X_o \). For any orthogonal matrix \( W \in O(n; \mathbb{R}) \) and any random Gaussian matrix \( G \in \mathbb{R}^{n \times p} \), \( g_{i,j} \sim_{iid} \mathcal{N}(0, \eta^2) \) independent of \( X_o \), let \( Y_N = Y + G \) denote the data with noise. Then the expectation of \( \frac{1}{np} \|W^* Y_N\|_4^4 \) satisfies:

\[
\frac{1}{np} \mathbb{E}_{X_o, G} \|W^* Y_N\|_4^4 = 3\theta(1-\theta) \frac{\|W^* D_o\|_4^4}{n} + 3\theta^2 + 6\theta \eta^2 + 3\eta^4. \tag{23}
\]

Proof Let \( W^* D_o = M \in O(n; \mathbb{R}) \), notice that the orthogonal transformation \( (W^* G) \) of a Gaussian matrix \( G \) is still a Gaussian matrix and satisfies \( \{W^* G\}_{i,j} \sim \mathcal{N}(0, 1) \), and it is independent of \( Y(X_o) \). We abuse the notation a bit let \( G = W^* G \) in the following calculation, since \( W^* G \) is also a Gaussian matrix independent of \( X_o \) and it will not affect the final result.

\[
\begin{align*}
\mathbb{E}_{X_o, G} \|W^* Y_N\|_4^4 &= \mathbb{E}_{X_o, G} \|MX_o + G\|_4^4 \\
&= \sum_{j=1}^{p} \sum_{i=1}^{n} \mathbb{E}_{X_o, G} \left( \sum_{k=1}^{n} m_{i,k} x_{i,k,j} + g_{i,j} \right)^4 \\
&= \sum_{j=1}^{p} \sum_{i=1}^{n} \left\{ \mathbb{E}_{X_o, G} \left( \sum_{k=1}^{n} m_{i,k} x_{i,k,j} \right)^4 + 6n \mathbb{E}_{X_o, G} \left[ \left( \sum_{k=1}^{n} m_{i,k} x_{i,k,j} \right)^2 \right] \right\} + \mathbb{E}_{G} \|G\|_4^4 \\
&= \sum_{j=1}^{p} \sum_{i=1}^{n} \left\{ \mathbb{E}_{X_o, G} \left( \sum_{k=1}^{n} m_{i,k} x_{i,k,j} \right)^4 + 6n \eta^2 \mathbb{E}_{X_o} \left[ \left( \sum_{k=1}^{n} m_{i,k} x_{i,k,j} \right)^2 \right] \right\} + 3np \eta^4 \\
&= \mathbb{E}_{X_o} \|MX_o\|_4^4 + 6np \theta \eta^2 + 3np \eta^4 \\
&= 3p(1-\theta) \|M\|_4^4 + 3np \theta^2 + 6n \eta^2 + 3np \eta^4,
\end{align*}
\]

therefore,

\[
\frac{1}{np} \mathbb{E}_{X_o, G} \|W^* Y_N\|_4^4 = 3\theta(1-\theta) \frac{\|W^* D_o\|_4^4}{n} + 3\theta^2 + 6\theta \eta^2 + 3\eta^4, \tag{25}
\]

which completes the proof. \( \blacksquare \)

A.2 PROOF OF THEOREM 3.2

Claim A.2 (Concentration of Objective with Small Noise) \( \forall \theta \in (0, 1), \) let \( X_o \in \mathbb{R}^{n \times p}, x_{i,j} \sim_{iid} \mathcal{N}(0, \Sigma) \). Let \( D_o \in O(n; \mathbb{R}) \) be an orthogonal matrix and assume \( Y = D_o X_o \). For any orthogonal matrix \( W \in O(n; \mathbb{R}) \) and any random Gaussian matrix \( G \in \mathbb{R}^{n \times p} \), \( g_{i,j} \sim_{iid} \mathcal{N}(0, \eta^2) \) independent of \( X_o \), let \( Y_N = Y + G \) denote the input with noise, then:

\[
\mathbb{P} \left( \sup_{W \in O(n; \mathbb{R})} \left| \frac{1}{np} \|W^* Y_N\|_4^4 - \mathbb{E} \|W^* Y_N\|_4^4 \right| \geq \delta \right) < \frac{1}{p}, \tag{26}
\]

when \( p = \Omega \left((1 + \eta^2)^4 n^2 \ln n/\delta^2\right) \).

Proof According to Proposition 3.1, we know that

\[
\frac{1}{np} \mathbb{E}_{X_o, G} \|W^* Y_N\|_4^4 = 3\theta(1-\theta) \frac{\|W^* D_o\|_4^4}{n} + 3\theta^2 + 6\theta \eta^2 + 3\eta^4. \tag{27}
\]
Moreover, since $Y_N = [y_1 + g_1, y_2 + g_2, \ldots, y_p + g_p]$, whose columns are independent. Let $Z = [z_1, z_2, \ldots, z_p] = D_o^* Y_N = [x_1 + g_1, x_2 + g_2, \ldots, x_p + g_p] = X_o + G$, (28) note that we abuse the notation a bit by assuming $D_o^* G = G$, since $D_o^* G$ is also a Gaussian matrix independent of $X_o$ and it will not affect the final result. Define function $f_\varepsilon(W) : O(n; \mathbb{R}) \mapsto \mathbb{R}$ as
\[
f_\varepsilon(W) = \|W^* z\|_4^4.
\] (29)
Next, we will check assumption 1, 2, and 3 then apply Lemma B.4.

**Assumption 1:** $\mu(n, p)$. Since $z_{i,j} = x_{i,j} + g_{i,j}$, we know that with probability $\theta, z \sim \mathcal{N}(0, 1+\eta^2)$ and with probability $1-\theta, z \sim \mathcal{N}(0, 1)$. Hence,
\[
\mathbb{P} \left( \max_{i,j} |z_{i,j}| > B \right) \leq 2np\theta \exp \left( -\frac{B^2}{2 + 2\eta^2} \right) + 2np(1-\theta) \exp \left( -\frac{B^2}{2} \right) = \mu(n, p).
\] (30)
Specifically, we set $B = \ln p$, which yields
\[
\mu(n, p) = 2np\theta \exp \left( -\frac{(\ln p)^2}{2 + 2\eta^2} \right) + 2np(1-\theta) \exp \left( -\frac{(\ln p)^2}{2} \right).
\] (31)

**Assumption 2:** Lipschitz Constant $L_f$. By Proposition 3.1, $\forall W \in O(n; \mathbb{R})$, we know that
\[
\mathbb{E} f_\varepsilon(W) = 3\theta(1-\theta) \|W^* D_o\|_4^4 + 3n\theta^2 + 6n\theta\eta^2 + 3n\eta^4.
\] (32)
Hence, $\forall W_1, W_2 \in O(n; \mathbb{R})$, we have
\[
\left| \mathbb{E} \|W_1^* z\|_4^4 - \mathbb{E} \|W_2^* z\|_4^4 \right| = 3\theta(1-\theta) \left| \|W_1^* D_o\|_4^4 - \|W_2^* D_o\|_4^4 \right|,
\] (33)
moreover,
\[
\sum_{i,j} \left[ (W_1^* D_o)_{i,j}^4 - (W_2^* D_o)_{i,j}^4 \right]
\]
\[
= \sum_{i,j} \left\{ (W_1^* D_o)_{i,j}^2 - (W_2^* D_o)_{i,j}^2 \right\} \left[ (W_1^* D_o)_{i,j}^2 + (W_2^* D_o)_{i,j}^2 \right]
\]
\[
\leq \left\{ \sum_{i,j} \left[ (W_1^* D_o)_{i,j}^2 - (W_2^* D_o)_{i,j}^2 \right]^2 \right\}^{1/2} \left\{ \sum_{i,j} \left[ (W_1^* D_o)_{i,j}^2 + (W_2^* D_o)_{i,j}^2 \right]^2 \right\}^{1/2}
\]
\[
= \left\| (W_1^* D_o)^{o2} - (W_2^* D_o)^{o2} \right\|_F \left\| (W_1^* D_o)^{o2} + (W_2^* D_o)^{o2} \right\|_F
\]
\[
\leq \|W_1^* D_o - W_2^* D_o\|_F \|W_1^* D_o + W_2^* D_o\|_F \left( \|W_1^* D_o\|_F^2 + \|W_2^* D_o\|_F^2 \right)
\]
\[
\leq 4n^2 \|W_1^* D_o - W_2^* D_o\|_2 = 4n^2 \|W_1 - W_2\|_2.
\]
Hence, we know that
\[
\left| \mathbb{E} \|W_1^* z\|_4^4 - \mathbb{E} \|W_2^* z\|_4^4 \right| \leq 12n^2\theta(1-\theta),
\] (35)
which yields
\[
L_f = 12n^2\theta(1-\theta).
\] (36)

**Assumption 2:** Lipschitz Constant $\bar{L}_f$. Since $f_\varepsilon(W) = \|W^* \tilde{z}\|_4^4$, we have
\[
\left| \|W_1^* \tilde{z}\|_4^4 - \|W_2^* \tilde{z}\|_4^4 \right| = \left| \sum_i \left\{ (W_1^* \tilde{z})_{i}^4 - (W_2^* \tilde{z})_{i}^4 \right\} \left[ (W_1^* \tilde{z})_{i}^2 + (W_2^* \tilde{z})_{i}^2 \right] \right|
\]
\[
\leq \left\{ \sum_i \left[ (W_1^* \tilde{z})_{i}^2 - (W_2^* \tilde{z})_{i}^2 \right]^2 \right\}^{1/2} \left\{ \sum_i \left[ (W_1^* \tilde{z})_{i}^2 + (W_2^* \tilde{z})_{i}^2 \right]^2 \right\}^{1/2}
\]
\[
= \left( \|W_1^* \tilde{z} - W_2^* \tilde{z}\|^2 \right)_{\Gamma_1} \left( \|W_1^* \tilde{z} + W_2^* \tilde{z}\|^2 \right)_{\Gamma_2}.
\] (37)
For $\Gamma_1$, we have
\[
\Gamma_1 = \| (W_1^* \bar{z})^2 - (W_2^* \bar{z})^2 \|_2 = \| (W_1^* \bar{z}) - (W_2^* \bar{z}) \circ (W_1^* \bar{z}) + (W_2^* \bar{z}) \|_2 \\
\leq \| W_1 - W_2 \|_2 \| W_1 + W_2 \|_2 \| \bar{z} \|_2^2 \\
\leq 2n B^2 \| W_1 - W_2 \|_2. 
\]
(38)

For $\Gamma_2$, we have
\[
\Gamma_2 = \| (W_1^* \bar{z})^2 + (W_2^* \bar{z})^2 \|_2 \leq \| (W_1^* \bar{z})^2 \|_2 + \| (W_2^* \bar{z})^2 \|_2 \\
\leq \| W_1^* \bar{z} \|_2^2 + \| W_2^* \bar{z} \|_2^2 \leq 2n B^2. 
\]
(39)

Hence, we know that
\[
\| W_1^* \bar{z} \|_4^4 - \| W_2^* \bar{z} \|_4^4 \leq 4n^2 B^4 \| W_1 - W_2 \|_2, 
\]
(40)

which yields
\[
\bar{L}_f = 4n^2 B^4. 
\]
(41)

**Assumption 3: Upper Bound $R_1$.** Notice that $\forall W \in O(n; \mathbb{R})$,
\[
\bar{f}_z(W) = \| W^* \bar{z} \|_4^4 = \| W^* \bar{z} \|_2^4 \left\| \frac{W^* \bar{z}}{\| W^* \bar{z} \|_2} \right\|_4^4 \leq \| W^* \bar{z} \|_2 = \| \bar{z} \|_2 \leq n^2 B^4 = R_1. 
\]
(42)

**Assumption 3: Upper Bound $R_2$.** Assume the support for $x$ is $\mathcal{S}$, so we know that know that that
\[
z_i = \begin{cases} 
\sqrt{1 + \eta^2 v_i}, & \text{if } i \in \mathcal{S}, \\
v_i, & \text{otherwise},
\end{cases} 
\]
(43)

where $v \sim \mathcal{N}(0, I)$ is a Gaussian vector. Let $P_{\mathcal{S}} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a projection that satisfies $\forall q \in \mathbb{R}^n$
\[
(P_{\mathcal{S}} q)_i = \begin{cases} 
\sqrt{1 + \eta^2 q_i}, & \text{if } i \in \mathcal{S}, \\
q_i, & \text{otherwise},
\end{cases} 
\]
(44)

Moreover, $\forall W \in O(n; \mathbb{R})$, let $w_i$ denote the $i^{th}$ column vector of $W$, hence
\[
\mathbb{E}[\bar{f}_z(W)] = \mathbb{E} \left[ \left( \| W^* \bar{z} \|_4^4 \right)^2 \right] = \mathbb{E} \left[ \sum_{i=1}^n \sum_{j=1}^n \langle w_i, z \rangle^4 \langle w_j, z \rangle^4 \right] \\
\leq \mathbb{E}_S \sum_{i=1}^n \sum_{j=1}^n \mathbb{E} \left[ (P_{\mathcal{S}} w_i, v)^4 (P_{\mathcal{S}} w_j, v)^4 \right] \leq \mathbb{E}_S \sum_{i=1}^n \sum_{j=1}^n \left[ \mathbb{E} (P_{\mathcal{S}} w_i, v)^8 \mathbb{E} (P_{\mathcal{S}} w_j, v)^8 \right]^\frac{1}{2}. 
\]
(45)

Notice that $v \sim \mathcal{N}(0, I)$, therefore $\forall i \in [n]$, we have
\[
\mathbb{E} (P_{\mathcal{S}} w_i, v)^8 = 105 \| P_{\mathcal{S}} w_i \|_2^8, 
\]
(46)

hence
\[
\mathbb{E}_S \sum_{i=1}^n \sum_{j=1}^n \left[ \mathbb{E} (P_{\mathcal{S}} w_i, v)^8 \mathbb{E} (P_{\mathcal{S}} w_j, v)^8 \right]^\frac{1}{2} = 105 \mathbb{E}_S \sum_{i=1}^n \sum_{j=1}^n \| P_{\mathcal{S}} w_i \|_2^4 \| P_{\mathcal{S}} w_j \|_2^4 \\
= 105 \mathbb{E}_S \sum_{i=1}^n \sum_{j=1}^n \sum_{k_1, k_2, k_3, k_4} \mathbb{E}_S \left[ \prod_{i=1}^4 (1 + \eta^2 \mathbb{1}_{k_i \in \mathcal{S}}) w_{i,k_1}^2 w_{i,k_2}^2 w_{j,k_3}^2 w_{j,k_4}^2 \right] \\
\leq 105 (1 + \eta^2)^4 \sum_{i=1}^n \sum_{j=1}^n \sum_{k_1, k_2, k_3, k_4} \left[ w_{i,k_1}^2 w_{i,k_2}^2 w_{j,k_3}^2 w_{j,k_4}^2 \right] = 105 (1 + \eta^2)^4 n^2. 
\]
(47)

Therefore, we can conclude that
\[
\mathbb{E}[\bar{f}_z(W)] \leq 105 (1 + \eta^2)^4 = R_2. 
\]
(48)
Applying Lemma B.4 for Concentration. Now we apply Lemma B.4 with

1. \[ B = \ln p, \quad \mu(n, p) = 2np\theta \exp\left(-\frac{(\ln p)^2}{4}\right) + 2np(1 - \theta) \exp\left(-\frac{(\ln p)^2}{2}\right), \quad (49) \]

2. \[ L_f = 12n^2\theta(1 - \theta), \quad \tilde{L}_f = 4n^2(\ln p)^4, \quad (50) \]

3. \[ R_1 = n^2(\ln p)^4, \quad R_2 = 105(1 + \eta^2)^4n^2, \quad (51) \]

we have

\[ \mathbb{P}\left( \sup_{\mathbf{w} \in O(n; \mathbb{R})} \frac{1}{np} \left| \sum_{j=1}^{p} [f_{z_j}(\mathbf{W}) - \mathbb{E} f_{z_j}(\mathbf{W})] \right| \geq \delta \right) \]

\[ = \mathbb{P}\left( \sup_{\mathbf{w} \in O(n; \mathbb{R})} \frac{1}{np} \left| \|\mathbf{W}^*\mathbf{Y}_N\|_4^4 - \mathbb{E} \|\mathbf{W}^*\mathbf{Y}_N\|_4^4 \right| \geq \delta \right) \]

\[ \leq \exp \left[ \frac{3\delta^2}{2(32R_2 + 8R_1n\delta)3/3} + n^2 \ln \frac{12(\tilde{L}_f + \tilde{L}_f)}{n\delta} + \ln 2 \right] + \mu(n, p) \]

\[ \leq \exp \left[ \frac{-3\delta^2}{C(1 + \eta^2)^4 + 8n(\ln p)^2\delta} + n^2 \ln \frac{60(n\ln p)^4}{\delta} + \ln 2 \right] \]

\[ + 2np\theta \exp\left(-\frac{(\ln p)^2}{4}\right) + 2np(1 - \theta) \exp\left(-\frac{(\ln p)^2}{2}\right) < \frac{1}{p}, \quad (52) \]

for a constant \( C > 10^4 \), when \( p = \Omega ((1 + \eta^2)^4n^2 \ln n/\delta^2) \). ∎

A.3 PROOF OF PROPOSITION 3.3

Claim A.3 (Expectation of Objective with Outliers) \( \forall \theta \in (0, 1), \) let \( \mathbf{X}_o \in \mathbb{R}^{n \times p}, \) \( x_{i,j} \sim_{i\text{id}} \mathcal{BG}(\theta). \) Let \( \mathbf{D}_o \in O(n; \mathbb{R}) \) be an orthogonal matrix and assume \( \mathbf{Y} = \mathbf{D}_o \mathbf{X}_o. \) For any orthogonal matrix \( \mathbf{W} \in O(n; \mathbb{R}) \) and any random Gaussian matrix \( \mathbf{G}' \in \mathbb{R}^{n \times \tau p}, g_{i,j} \sim_{i\text{id}} \mathcal{N}(0, 1) \) independent of \( \mathbf{X}_o, \) let \( \mathbf{Y}_o = [\mathbf{Y}, \mathbf{G}'] \) denote the data with outliers \( \mathbf{G}' \). Then the expectation of \( \frac{1}{np} \|\mathbf{W}^*\mathbf{Y}_o\|_4^4 \) satisfies:

\[ \frac{1}{np} \mathbb{E}_{\mathbf{X}_o, \mathbf{G}'} \|\mathbf{W}^*\mathbf{Y}_o\|_4^4 = 3\theta(1 - \theta) \frac{\|\mathbf{W}^*\mathbf{D}_o\|_4^4}{n} + 3\theta^2 + 3\tau. \quad (53) \]

Proof Notice that

\[ \mathbb{E}_{\mathbf{X}_o, \mathbf{G}'} \|\mathbf{W}^*\mathbf{Y}_o\|_4^4 = \mathbb{E}_{\mathbf{X}_o} \|\mathbf{W}^*\mathbf{Y}\|_4^4 + \mathbb{E}_{\mathbf{G}'} \|\mathbf{W}^*\mathbf{G}'\|_4^4, \quad (54) \]

and

\[ \mathbb{E}_{\mathbf{X}_o} \|\mathbf{W}^*\mathbf{Y}\|_4^4 = \mathbb{E}_{\mathbf{X}_o} \|\mathbf{W}^*\mathbf{D}_o\mathbf{X}_o\|_4^4 = 3p\theta(1 - \theta) \|\mathbf{W}^*\mathbf{D}_o\|_4^4 + 3np\theta^2. \quad (55) \]

Moreover, the orthogonal rotation \( (\mathbf{W}^*\mathbf{G}') \) of a standard Gaussian matrix \( \mathbf{G}' \) is also a standard Gaussian matrix, therefore,

\[ \mathbb{E}_{\mathbf{G}'} \|\mathbf{W}^*\mathbf{G}'\|_4^4 = 3\tau np. \quad (56) \]

Hence,

\[ \frac{1}{np} \mathbb{E}_{\mathbf{X}_o, \mathbf{G}'} \|\mathbf{W}^*\mathbf{Y}_o\|_4^4 = 3\theta(1 - \theta) \frac{\|\mathbf{W}^*\mathbf{D}_o\|_4^4}{n} + 3\theta^2 + 3\tau, \quad (57) \]

which completes the proof. ∎
A.4  PROOF OF THEOREM 3.4

Claim A.4 (Concentration of Objective with Outliers) \( \forall \theta \in (0, 1), \) let \( X_\theta \in \mathbb{R}^{n \times p}, \ x_{i,j} \sim_{\text{iid}} \mathcal{N}(0, \theta^2). \) Let \( D_\theta \in \mathbb{O}(n; \mathbb{R}) \) be an orthogonal matrix and assume \( Y = D_\theta X_\theta. \) For any orthogonal matrix \( W \in \mathbb{O}(n; \mathbb{R}) \) and any random Gaussian matrix \( G' \in \mathbb{R}^{n \times np}, \ g'_{i,j} \sim_{\text{iid}} \mathcal{N}(0, 1) \) independent of \( X_\theta, \) let \( Y_O = \begin{bmatrix} Y \ G' \end{bmatrix} \) denote the input with outlier \( G', \) then:

\[
\mathbb{P} \left( \sup_{W \in \mathbb{O}(n; \mathbb{R})} \frac{1}{np} \left\| W^* Y_O \right\|_4^4 - \mathbb{E} \left\| W^* Y_O \right\|_4^4 \geq \delta \right) < \frac{1}{p}, \quad (58)
\]

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

**Proof** Notice that

\[
\begin{align*}
&\mathbb{P} \left( \sup_{W \in \mathbb{O}(n; \mathbb{R})} \frac{1}{np} \left\| W^* Y_O \right\|_4^4 - \mathbb{E} \left\| W^* Y_O \right\|_4^4 \geq \delta \right) \\
&\leq \mathbb{P} \left( \sup_{W \in \mathbb{O}(n; \mathbb{R})} \frac{1}{np} \left\| W^* Y \right\|_4^4 - \mathbb{E} \left\| W^* Y \right\|_4^4 \geq \delta \right) \\
&\quad + \mathbb{P} \left( \sup_{W \in \mathbb{O}(n; \mathbb{R})} \frac{1}{np} \left\| W^* G' \right\|_4^4 - \mathbb{E} \left\| W^* G' \right\|_4^4 \geq \delta \right).
\end{align*}
\]

Apply Lemma B.5 (substitute \( \delta \) with \( \delta/2 \)) to \( \Gamma_1, \) we know that

\[
\Gamma_1 = \mathbb{P} \left( \sup_{W \in \mathbb{O}(n; \mathbb{R})} \frac{1}{np} \left\| W^* Y \right\|_4^4 - \mathbb{E} \left\| W^* Y \right\|_4^4 \geq \delta \right) < \exp \left( - \frac{3p\delta^2}{c_1 \theta + 16n(\ln p)^4} + n^2 \ln \left( \frac{120np(\ln p)^4}{\delta} \right) \right) \quad (60)
\]

for two constants \( c_1 \geq 4 \times 10^4, c_2 > 13,440, \) and the last inequality holds when \( p = \Omega(\theta n^2 \ln n/\delta^2). \) Moreover, apply Lemma B.5 again (let \( \theta = 1, \ Y = G' \in \mathbb{R}^{n \times np} \) and substitute \( \delta \) with \( \delta/2 \)) we have

\[
\Gamma_2 = \mathbb{P} \left( \sup_{W \in \mathbb{O}(n; \mathbb{R})} \frac{1}{np} \left\| W^* G' \right\|_4^4 - \mathbb{E} \left\| W^* G' \right\|_4^4 \geq \delta \right) \quad (61)
\]

for two constants \( c_1 \geq 4 \times 10^4, c_2 > 13,440, \) and the last inequality holds when \( p = \Omega(\tau^2 n^2 \ln n/\delta^2). \) Combine \( \theta \leq 1 \) and \( \tau \) can be larger than 1, we conclude that

\[
\mathbb{P} \left( \sup_{W \in \mathbb{O}(n; \mathbb{R})} \frac{1}{np} \left\| W^* Y_O \right\|_4^4 - \mathbb{E} \left\| W^* Y_O \right\|_4^4 \geq \delta \right) < \Gamma_1 + \Gamma_2 < \frac{1}{p} \quad (62)
\]

when \( p = \Omega(\tau^2 n^2 \ln n/\delta^2). \)
A.5 Proof of Proposition 3.5

Claim A.5 (Expectation of Objective with Sparse Corruptions) \( \forall \theta \in (0, 1), \) let \( X_0 \in \mathbb{R}^{n \times p}, x_{i,j} \sim \text{id} \) \( \mathcal{B}(\theta) \). Let \( D_0 \in O(n; \mathbb{R}) \) be an orthogonal matrix and assume \( Y = D_0 X_0 \). For any orthogonal matrix \( W \in \mathbb{O}(n; \mathbb{R}) \) and any random Bernoulli matrix \( B \in \mathbb{R}^{n \times p}, b_{i,j} \sim \text{id} \) \( \mathcal{B}(\beta) \) independent of \( X_0 \), let \( Y_C = Y + \sigma B \circ S \) denote the data with sparse corruptions, and \( S \in \mathbb{R}^{n \times p} \) is defined in equation 15. Then the expectation of \( \frac{1}{np} \| W^* Y_C \|_4^4 \) satisfies:

\[
\frac{1}{np} \mathbb{E}_{X, B, S} \| W^* Y_C \|_4^4 = 3\theta (1 - \theta) \frac{\| W^* D_0 \|_4^4}{n} + \sigma^4 \beta (1 - 3\beta) \frac{\| W \|_4^4}{n} + 3\theta^2 + 6\sigma^2 \theta \beta + 3\sigma^4 \beta^2.
\]

(63)

Proof Let \( W^* D_0 = M \in \mathbb{O}(n; \mathbb{R}) \), notice that

\[
\| W^* Y_C \|_4^4 = \| M X_0 + \sigma W^* (B \circ S) \|_4^4,
\]

hence

\[
\mathbb{E}_{X, B, S} \| W^* Y_C \|_4^4 = \mathbb{E}_{X, B, S} \sum_{j=1}^{p} \sum_{i=1}^{n} \left( \sum_{k=1}^{n} m_{i,k} x_{k,j} + \sigma \sum_{k=1}^{n} w_{i,k} b_{k,j} s_{k,j} \right)^4
\]

\[
= \sum_{j=1}^{p} \sum_{i=1}^{n} \left( \sum_{k=1}^{n} m_{i,k} x_{k,j} \right)^4 + 6 \sum_{j=1}^{p} \sum_{i=1}^{n} \left( \sum_{k=1}^{n} m_{i,k} x_{k,j} \right)^2 \left( \sum_{k=1}^{n} w_{i,k} b_{k,j} s_{k,j} \right)^2
\]

\[
+ \left( \sum_{k=1}^{n} w_{i,k} b_{k,j} s_{k,j} \right)^4 + 4 \sum_{j=1}^{p} \sum_{i=1}^{n} \left( \sum_{k=1}^{n} m_{i,k} x_{k,j} \right)^3 \left( \sum_{k=1}^{n} w_{i,k} b_{k,j} s_{k,j} \right)^3
\]

Moreover,

- \( \Gamma_1 \):

\[
\mathbb{E}_{X, B, S} \Gamma_1 = \mathbb{E}_{X} \Gamma_1 = 3\theta \sum_{k=1}^{n} m_{i,k}^4 + 6\theta^2 \sum_{1 \leq k_1 < k_2 \leq n} m_{i,k_1}^2 m_{i,k_2}^2
\]

\[
= 3\theta (1 - \theta) \sum_{k=1}^{n} m_{i,k}^4 + 3\theta^2,
\]

(66)

- \( \Gamma_2 \):

\[
\mathbb{E}_{X, B, S} \Gamma_2 = \left[ \mathbb{E}_{X} \left( \sum_{k=1}^{n} m_{i,k} x_{k,j} \right)^2 \right] \left[ \mathbb{E}_{B, S} \left( \sum_{k=1}^{n} w_{i,k} b_{k,j} s_{k,j} \right)^2 \right]
\]

\[
= \theta \left( \sum_{k=1}^{n} m_{i,k}^2 k \right) \sigma^2 \beta \left( \sum_{k=1}^{n} w_{i,k}^2 k \right) = \sigma^2 \theta \beta,
\]

(67)

- \( \Gamma_3 \):

\[
\mathbb{E}_{X, B, S} \Gamma_3 = \mathbb{E}_{B, S} \Gamma_3 = \sigma^4 \beta \sum_{k=1}^{n} w_{i,k}^4 + 6\sigma^4 \beta^2 \sum_{1 \leq k_1 < k_2 \leq n} w_{i,k_1}^2 w_{i,k_2}^2
\]

\[
= \sigma^4 \beta (1 - 3\beta) \sum_{k=1}^{n} w_{i,k}^4 + 3\sigma^4 \beta^2,
\]

(68)
\[ \Gamma_4, \Gamma_5 : \quad E_{X_n, B, S} \Gamma_4 = 0, \quad E_{X_n, B, S} \Gamma_5 = 0. \quad (69) \]

Substitute \( \Xi \Gamma_1, \Xi \Gamma_2, \Xi \Gamma_3, \Xi \Gamma_4, \Xi \Gamma_5 \) back to equation (65), yields
\[ \frac{1}{np} E_{X_n, B, S} \| WY \|_4^4 = 3\theta (1-\theta) \frac{\| W^* D_o \|_4^4}{n} + \sigma^4 \beta (1 - 3\beta) \frac{\| W \|_4^4}{n} + 3\theta^2 + 6\sigma^2 \theta \beta + 3\sigma^4 \beta^2. \quad (70) \]

**Proof** of Theorem 3.6

**Claim A.6 (Concentration of Objective with Sparse Corruptions)** \( \forall \theta \in (0, 1), \text{let } X_o \in \mathbb{R}^{n \times p}, \) \( x_{i,j} \sim \text{iid } BG(\theta). \) Let \( D_o \in \mathbb{O}(n; \mathbb{R}) \) be an orthonormal matrix and assume \( Y = D_o X_o. \) For any orthonormal matrix \( W \in \mathbb{O}(n; \mathbb{R}) \) and any random Bernoulli matrix \( B \in \mathbb{R}^{n \times p}, b_{i,j} \sim \text{iid } Ber(\beta) \) independent of \( X_o, \) let \( Y_C = Y + \sigma B \circ S \) denote the input with sparse corruptions, and \( S \in \mathbb{R}^{n \times p} \) is defined in equation 15, then:
\[ P \left( \sup_{W \in \mathbb{O}(n; \mathbb{R})} \left| \frac{1}{np} \| W^* Y_C \|_4^4 - E \| W^* Y_C \|_4^4 \right| \geq \delta \right) < \frac{1}{p}. \quad (71) \]
when \( p = \Omega \left( \sigma^8 \beta n^2 \ln n / \delta^2 \right). \)

**Proof** According to Proposition 3.5, we know that
\[ \frac{1}{np} E_{X_n, B, S} \| W^* Y_C \|_4^4 = 3\theta (1-\theta) \frac{\| W^* D_o \|_4^4}{n} + \sigma^4 \beta (1 - 3\beta) \frac{\| W \|_4^4}{n} + 3\theta^2 + 6\sigma^2 \theta \beta + 3\sigma^4 \beta^2. \quad (72) \]
Moreover, since \( Y_C = [y_1 + \sigma b_1 \circ s_1, y_2 + \sigma b_2 \circ s_2, \ldots, y_p + \sigma b_p \circ s_p], \) whose columns are independent. Let
\[ Z = [z_1, z_2, \ldots, z_p] = D^* o Y_C = X_o + \sigma C, \quad (73) \]
where \( C = D^* (B \circ S). \) Note that the columns of \( Z \) are independent, we can define function \( f_z(W) : \mathbb{O}(n; \mathbb{R}) \mapsto \mathbb{R} \) as
\[ f_z(W) = \| W^* z \|_4^4. \quad (74) \]
Next, we will check assumption 1, 2, and 3 then apply Lemma B.4.

**Assumption 1:** \( \mu(n, p). \) Since \( Z = X_o + \sigma C, \) we have
\[ \{ \max_{i,j} |z_{i,j}| > B \} = \{ \max_{i,j} |x_{i,j} + \sigma c_{i,j}| > B \} \subseteq \{ \max_{i,j} |x_{i,j}| > B - \sigma \max_{i,j} |c_{i,j}| \}. \quad (75) \]
Moreover, since
\[ \max_{i,j} |c_{i,j}| \leq \max_{i} \| d_i \|_1 \leq \sqrt{n}, \quad (76) \]
we have
\[ \{ \max_{i,j} |z_{i,j}| > B \} \subseteq \{ \max_{i,j} |x_{i,j}| > B - \sigma \sqrt{n} \}. \quad (77) \]
Thus we know that
\[ P \left( \max_{i,j} |z_{i,j}| > B \right) \leq P \left( \max_{i,j} |x_{i,j}| > B - \sigma \sqrt{n} \right) \leq 2np\theta \exp \left( -\frac{(B - \sigma \sqrt{n})^2}{2} \right). \quad (78) \]
Specifically, we set \( B = \frac{p^2 \dagger}{2}, \) which yields
\[ \mu(n, p) = 2np\theta \exp \left( -\frac{\left( \frac{p^2 \dagger}{2} - \sigma \sqrt{n} \right)^2}{2} \right). \quad (79) \]
Assumption 2: Lipschitz Constant $L_f$. By Proposition 3.5, $\forall W \in O(n; \mathbb{R})$, we know that
\[
E f_z(W) = 3\theta(1 - \theta) \|W^*D_o\|_4^2 + \sigma^4 \beta(1 - 3\beta) \|W\|_4^2 + 3n\theta^2 + 6n\sigma^2 \theta \beta + 3n\sigma^4 \beta^2. \tag{80}
\]
Hence $\forall W_1, W_2 \in O(n; \mathbb{R})$, we have
\[
|E \|W_1^*z\|_4^4 - E \|W_2^*z\|_4^4| \leq 3\theta(1 - \theta) \|W_1^*D_o\|_4^2 - \|W_1^*D_o\|_4^2 + \sigma^4 \beta(1 - 3\beta) \|W_1\|_4^4 - \|W_2\|_4^4, \tag{81}
\]
from the derivation of equation 34, we have
\[
\|W_1^*D_o\|_4^2 - \|W_2^*D_o\|_4^2 \leq 4n^2 \|W_1 - W_2\|_2, \tag{82}
\]
and therefore, we have
\[
|E \|W_1^*z\|_4^4 - E \|W_2^*z\|_4^4| \leq [12n^2\theta(1 - \theta) + 4n^2\sigma^4 \beta(1 - 3\beta)] \|W_1 - W_2\|_2, \tag{83}
\]
which yields
\[
L_f = 12n^2\theta(1 - \theta) + 4n^2\sigma^4 \beta(1 - 3\beta). \tag{84}
\]
Assumption 2: Lipschitz Constant $\bar{L}_f$. Since $f_z(W) = \|W^*\bar{z}\|_4^4$, apply the same derivation as equation 37, we have
\[
\|W_1^*\bar{z}\|_4^4 - \|W_2^*\bar{z}\|_4^4 \leq 4n^2B^4 \|W_1 - W_2\|_2, \tag{85}
\]
which yields
\[
\bar{L}_f = 4n^2B^4. \tag{86}
\]
Assumption 3: Upper Bound $R_1$. Notice that $\forall W \in O(n; \mathbb{R})$,
\[
f_z(W) = \|W^*\bar{z}\|_4^4 = \|W^*\bar{z}\|_4^4 = \|W^*\bar{z}\|_2^4 \|W^*\bar{z}\|_2^4 \leq \|W^*\bar{z}\|_2^4 \leq n^2B^4 = R_1. \tag{87}
\]
Assumption 3: Upper Bound $R_2$. Assume the support for $x$ is $S$, so we know that $\forall i \in [n]$,
\[
z_i = \begin{cases} v_i + c_i, & \text{if } i \in S, \\ c_i, & \text{otherwise,} \end{cases} \tag{88}
\]
where $v \sim \mathcal{N}(0, I)$ is a Gaussian vector and $c = D_o^*(b \circ s)$ is the sparse corruption vector after orthogonal rotation $D_o^*$. Let $P_S : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a projection onto the support $S$, that is, $\forall q \in \mathbb{R}^n$:
\[
(P_S q)_i = \begin{cases} q_i, & \text{if } i \in S, \\ 0, & \text{otherwise.} \end{cases} \tag{89}
\]
Let $M = D_oW$, and $h = b \circ s$, we have
\[
E f_z^2(W) = E \left( \|W^*z\|_4^8 \right) = \mathbb{E} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ \left( \langle w_i, x \rangle + \sigma \langle m_i, h \rangle \right)^4 \left( \langle w_j, x \rangle + \sigma \langle m_j, h \rangle \right)^4 \right] \leq \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ \mathbb{E} \left( \langle w_i, x \rangle + \sigma \langle m_i, h \rangle \right)^8 \mathbb{E} \left( \langle w_j, x \rangle + \sigma \langle m_j, h \rangle \right)^8 \right]^{\frac{1}{2}} \leq \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ \mathbb{E} \left( \langle P_S w_i, v \rangle + \sigma \langle m_i, h \rangle \right)^8 \mathbb{E} \left( \langle P_S w_j, v \rangle + \sigma \langle m_j, h \rangle \right)^8 \right]^{\frac{1}{2}}. \tag{90}
\]
Moreover, we have:

\[
\mathbb{E} \left( \left( \mathbf{P}_S \mathbf{w}_i, \mathbf{v} \right)^8 \right) + 28 \sigma^2 \mathbb{E} \left( \left( \mathbf{P}_S \mathbf{w}_i, \mathbf{v} \right)^6 \left( \mathbf{m}_i, \mathbf{h} \right)^2 \right) + 70 \sigma^4 \mathbb{E} \left( \left( \mathbf{P}_S \mathbf{w}_i, \mathbf{v} \right)^4 \left( \mathbf{m}_i, \mathbf{h} \right)^4 \right) + 28 \sigma^6 \mathbb{E} \left( \left( \mathbf{P}_S \mathbf{w}_i, \mathbf{v} \right)^6 \left( \mathbf{m}_i, \mathbf{h} \right)^2 \right) + \sigma^8 \mathbb{E} \left( \left( \mathbf{m}_i, \mathbf{h} \right)^8 \right).
\]

We now provide upper bound for \( \mathbb{E} \left( \left( \mathbf{P}_S \mathbf{w}_i, \mathbf{v} \right)^r \right) \) and \( \mathbb{E} \left( \left( \mathbf{m}_i, \mathbf{h} \right)^r \right) \) for \( r = 2, 4, 6, 8 \) respectively. Since \( \mathbb{E} \mathbf{h} = \mathbf{0} \) and \( \mathbf{h} = \mathbf{b} \circ \mathbf{s} \). Let \( \mathcal{S}' \) denote the support of \( \mathbf{h} \), we have

\[
\mathbb{E} \left( \left( \mathbf{m}_i, \mathbf{h} \right)^8 \right) = \mathbb{E} \prod_{i=1}^{n} \left( \sum_{k_i=1}^{m} \mathbf{m}_{k_i,i} \mathbf{h}_i \right)
= \sum_{k_1,k_2,k_3,k_4} \mathbb{E}_{\mathcal{S}'} \left( m_{k_1,i}^2 I_{k_1 \in \mathcal{S}'} m_{k_2,i}^2 I_{k_2 \in \mathcal{S}'} m_{k_3,i}^2 I_{k_3 \in \mathcal{S}'} m_{k_4,i}^2 I_{k_4 \in \mathcal{S}'} \right).
\]

Now we discuss these cases separately:

- With probability \( c_1 \beta^4 (c_1 \leq 1) \), all \( k_1, k_2, k_3, k_4 \) are in \( \mathcal{S} \), in this case, we have

\[
\mathbb{E} \left( \left( \mathbf{m}_i, \mathbf{h} \right)^8 \right) = \sum_{k_1,k_2,k_3,k_4} \mathbb{E}_{\mathcal{S}'} \left( m_{k_1,i}^2 m_{k_2,i}^2 m_{k_3,i}^2 m_{k_4,i}^2 I_{k_1 \in \mathcal{S}} I_{k_2 \in \mathcal{S}} I_{k_3 \in \mathcal{S}} I_{k_4 \in \mathcal{S}} \right) = 1.
\]

- With probability \( c_2 \beta^3 (c_2 \leq 1) \), only three among \( k_1, k_2, k_3, k_4 \) are in \( \mathcal{S}' \), in this case, we have

\[
\mathbb{E} \left( \left( \mathbf{m}_i, \mathbf{h} \right)^8 \right) = \sum_{k_1,k_2,k_3,k_4} \mathbb{E}_{\mathcal{S}'} \left( m_{k_1,i}^2 m_{k_2,i}^2 m_{k_3,i}^4 I_{k_1 \in \mathcal{S}} I_{k_2 \in \mathcal{S}} I_{k_3 \in \mathcal{S}} I_{k_4 \in \mathcal{S}} \right) = \sum_{k_1,k_2,k_3} \left( m_{k_1,i}^2 m_{k_2,i}^2 m_{k_3,i}^4 \right) = \| \mathbf{m}_i \|_4^4 \leq 1.
\]

- With probability \( c_3 \beta^2 (c_3 \leq 1) \), only two among \( k_1, k_2, k_3, k_4 \) are in \( \mathcal{S}' \), in this case, we have

\[
\mathbb{E} \left( \left( \mathbf{m}_i, \mathbf{h} \right)^8 \right) = \sum_{k_1,k_2,k_3,k_4} \mathbb{E}_{\mathcal{S}'} \left( m_{k_1,i}^4 m_{k_2,i}^4 I_{k_1 \in \mathcal{S}} I_{k_2 \in \mathcal{S}} I_{k_3 \in \mathcal{S}} I_{k_4 \in \mathcal{S}} \right) = \mathbb{E} \left( \left( \mathbf{m}_i \right)^4 \right) \leq \| \mathbf{m}_i \|_4^4 + \| \mathbf{m}_i \|_6^6 \leq 2.
\]

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

\[
\mathbb{E} \left( \left( \mathbf{m}_i, \mathbf{h} \right)^8 \right) = \sum_{k_1,k_2,k_3,k_4} \mathbb{E}_{\mathcal{S}'} \left( m_{k_1,i}^8 I_{k_1 \in \mathcal{S}} I_{k_2 \in \mathcal{S}} I_{k_3 \in \mathcal{S}} I_{k_4 \in \mathcal{S}} \right) = \sum_{k_1} m_{k_1,i}^8 = \| \mathbf{m}_i \|_8^8 \leq 1.
\]

Hence, combine equation \( 93 \), equation \( 94 \), equation \( 95 \), and equation \( 96 \), we know that

\[
\mathbb{E} \left( \left( \mathbf{m}_i, \mathbf{h} \right)^8 \right) \leq c \beta,
\]

for a constant \( c > 1 \). Similarly, one can conclude that

\[
\mathbb{E} \left( \left( \mathbf{m}_i, \mathbf{h} \right)^r \right) \leq c \beta, \quad \forall r = 2, 4, 6.
\]

Moreover, since \( \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \), we know that

\[
\mathbb{E} \left( \left( \mathbf{P}_S \mathbf{w}_i, \mathbf{v} \right)^r \right) = \left\| \mathbf{P}_S \mathbf{w}_i \right\|_2^r, \quad \forall r = 2, 4, 6, 8.
\]
Hence, with the similar technique applied above, we have
\[
\mathbb{E} \left( \langle P_S w_i, v \rangle^8 \right) = 105 \mathbb{E} \| P_S w_i \|_2^8 \\
= 105 \sum_{k_1, k_2, k_3, k_4} \mathbb{E} \left[ \left( w_{k_1}^2 \mathbb{1}_{k_1 \in S} + w_{k_2}^2 \mathbb{1}_{k_2 \in S} + w_{k_3}^2 \mathbb{1}_{k_3 \in S} + w_{k_4}^2 \mathbb{1}_{k_4 \in S'} \right)^8 \right] \leq 105 c \theta,
\]
for a constant \( c > 1 \). Similarly, we have
\[
\mathbb{E} \left( \langle P_S w_i, v \rangle^r \right) = \| P_S w_i \|_2^r \leq c (r - 1) \theta^r, \quad \forall r = 2, 4, 6,
\]
for a constant \( c > 1 \). Therefore, combine equation 91, we have
\[
\mathbb{E} \left( \langle P_S w_i, v \rangle^6 \right) + 28 \sigma^2 \mathbb{E} \left( \langle P_S w_i, v \rangle^6 \langle m_i, h \rangle^2 \right) + 70 \sigma^4 \mathbb{E} \left( \langle P_S w_i, v \rangle^4 (m_i, h)^4 \right) \\
+ 28 \sigma^6 \mathbb{E} \left( \langle P_S w_i, v \rangle^6 (m_i, h)^2 \right) + \sigma^8 \mathbb{E} \left( (m_i, h)^8 \right) \\
\leq c \theta + 28 \sigma^2 \theta \beta + 70 \sigma^4 \theta \beta + 28 \sigma^6 \theta \beta + \sigma^8 \beta \leq \sigma^8 \beta,
\]
for a constant \( c > 1 \). Hence, combine equation 90, we have
\[
\mathbb{E} f_2^2(W) = \mathbb{E} \left( \| W^* z \|_4^8 \right) \\
= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E} \left( \langle P_S w_i, v \rangle + \sigma \langle m_i, h \rangle \right)^8 \mathbb{E} \left( \langle P_S w_j, v \rangle + \sigma \langle m_j, h \rangle \right)^4 \leq \sigma^8 \beta n^2,
\]
for a constant \( c > 1 \). Therefore, we can conclude that
\[
R_2 = c \sigma^8 \beta n^2,
\]
for a constant \( c > 1 \).

**Applying Lemma B.4 for Concentration.** Now we apply Lemma B.4 with

1. \( B = p^4 \), \( \mu(n, p) = 2np \theta \exp \left( -\frac{(p^4 - \sigma \sqrt{n})^2}{2} \right) \), \( (105) \)

2. \( L_f = 12 n^2 \theta (1 - \theta) + 4n^2 \sigma^2 \beta (1 - 3 \beta) \), \( \tilde{L}_f = 4n^2 p \), \( (106) \)

3. \( R_1 = n^2 p \), \( R_2 = c \sigma^8 \beta n^2 \), \( (107) \)

for a constant \( c > 1 \), we have
\[
\begin{aligned}
\mathbb{P} \left( \sup_{W \in O(n; \mathbb{R})} \frac{1}{np} \sum_{j=1}^p \left[ f_{z_j}(W) - \mathbb{E} f_{z_j}(W) \right] \geq \delta \right) = & \mathbb{P} \left( \sup_{W \in O(n; \mathbb{R})} \frac{1}{np} \left[ \| W^* Y_C \|_4^4 - \mathbb{E} \| W^* Y_C \|_4^4 \right] \geq \delta \right) \\
< & \exp \left[ -\frac{p \mu(n, p)^2 \delta^2}{32 R_2 + 8 R_1 n \delta / 3} + n^2 \ln \left( \frac{12 (L_f + \tilde{L}_f)}{n \delta} \right) + \ln 2 \right] + \mu(n, p)
\end{aligned}
\]
\[
< \exp \left[ -\frac{3 \mu(n, p)^2 \delta^2}{C \sigma^8 \beta + 8 np \delta} + n^2 \ln \left( \frac{60 np}{\delta} \right) + \ln 2 \right] + 2np \theta \exp \left( -\frac{(p^4 - \sigma \sqrt{n})^2}{2} \right) < \frac{1}{p},
\]
for a constant \( C > 96 \), when \( p = \Omega \left( \sigma \beta n^2 / n \delta^2 \right) \). 

\[ \square \]
\section{Related Lemmas and Inequalities}

\begin{lemma}[Two-sided Bernstein’s Inequality]
Given \( p \) random variables \( x_1, x_2, \ldots, x_p \), if \( \forall i \in [p], |x_i| \leq b \) almost surely, then
\[
    P \left( \frac{1}{p} \sum_{i=1}^{p} |x_i - \mathbb{E}[x_i]| \geq t \right) \leq 2 \exp \left( -\frac{2pt^2}{p\sum_{i=1}^{p} \mathbb{E}[x_i^2] + 2bt/3} \right). \tag{109}
\]
\end{lemma}

\textbf{Proof} See Proposition 2.14 in Wainwright (2019), one can easily generalize it to two-sided case. ■

\begin{lemma}[Entry-wise Truncation of a Bernoulli Gaussian Matrix]
Let \( X \in \mathbb{R}^{n \times p} \), where \( x_{i,j} \sim_{iid} \text{BG}(\theta) \) and let \( \| \cdot \|_{\infty} \) denote the maximum element (in absolute value) of a matrix, then
\[
    P \left( \max_{i,j} |x_{i,j}| \geq t \right) \leq 2np\theta \exp \left( -\frac{t^2}{2} \right). \tag{110}
\]
\end{lemma}

\textbf{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
\[
    P \left( |x_{i,j}| \geq t \right) = \theta \cdot P \left( |g_{i,j}| \geq t \right) \leq 2\theta \exp \left( -\frac{t^2}{2} \right). \tag{111}
\]

By union bound, we have:
\[
    P \left( \max_{i,j} |x_{i,j}| \geq t \right) \leq \sum_{i=1}^{n} \sum_{j=1}^{p} P \left( |x_{i,j}| \geq t \right) \leq 2np\theta \exp \left( -\frac{t^2}{2} \right). \tag{112}
\]

\begin{lemma}[\( \epsilon \)-Net Covering of Stiefel Manifolds]
There is a covering \( \epsilon \)-net \( S_{\epsilon} \) for Stiefel manifold \( M = \{ W \in \mathbb{R}^{n \times p} | W^T W = I \}, (n \geq r) \) in operator norm
\[
    \forall W \in M, \ \exists W' \in S_{\epsilon} \text{ subject to } \| W - W' \|_2 \leq \epsilon, \tag{113}
\]
of size \( |S_{\epsilon}| \leq \left( \frac{5}{\epsilon} \right)^{nr} \).
\end{lemma}

\textbf{Proof} See Lemma D.4 in Zhai et al. (2019b). ■

\begin{lemma}[Uniform Concentration bound over \( \mathcal{O}(n; \mathbb{R}) \)]
Let \( Z \in \mathbb{R}^{n \times p} \) be a random matrix whose columns \( z_1, z_2, \ldots, z_p \) are i.i.d. drawn from a distribution \( \mathcal{P} \). \( \forall z \sim \mathcal{P} \), let \( f_z(\cdot) : \mathcal{O}(n; \mathbb{R}) \rightarrow \mathbb{R} \) denote a function that maps \( \mathcal{O}(n; \mathbb{R}) \) to \( \mathbb{R} \). \( \forall B > 0 \), let \( Z \) denote the truncation of \( Z \):
\[
    \tilde{z}_{i,j} = \begin{cases} 
        z_{i,j} & \text{if } |z_{i,j}| \leq B, \\
        0 & \text{otherwise.} 
    \end{cases} \tag{144}
\]
Assume that:
\begin{enumerate}
    \item \( P(\max_{i,j} |z_{i,j}| > B) < \mu(n, p) \), where \( \mu(n, p) \rightarrow 0 \) as \( p \) increase, \( B \) depends on \( p \).
    \item \( Ef_z(\cdot) \) is \( L_f \)-Lipschitz and \( f_z(\cdot) \) is \( \tilde{L}_f \)-Lipschitz, that is, \( \forall W_1, W_2 \in \mathcal{O}(n; \mathbb{R}) \):
\[
    |Ef_z(W_1) - Ef_z(W_2)| \leq L_f \| W_1 - W_2 \|_2, \\
    |f_z(W_1) - f_z(W_2)| \leq \tilde{L}_f \| W_1 - W_2 \|_2. \tag{115}
\]
    \item \( \forall W \in \mathcal{O}(n; \mathbb{R}), f_z(W) \leq R_1, Ef_z^2(W) \leq R_2. \)
\end{enumerate}
\end{lemma}

\textsuperscript{11}A similar result can be found in Lemma 4.5 of Recht et al. (2010).
Then:

\[
P\left( \sup_{W \in O(n; \mathbb{R})} \frac{1}{np} \sum_{j=1}^{p} \left| f_{z_j}(W) - \mathbb{E}f_{z_j}(W) \right| \geq \delta \right) < \exp \left[ -\frac{pn^2\delta^2}{32R_2 + 8R_1 n\delta/3} + n^2 \ln \left( \frac{12(L_f + \bar{L}_f)}{n\delta} \right) + \ln 2 \right] + \mu(n, p),
\]

(116)

when \( p > \rho \), where \( \rho \) depends on \( n \).

**Proof** By assumption 1, we have

\[
P\left( \sup_{W \in O(n; \mathbb{R})} \frac{1}{np} \sum_{j=1}^{p} \left| f_{z_j}(W) - \mathbb{E}f_{z_j}(W) \right| \geq \delta \right) \leq \mathbb{P}(Z \neq \bar{Z}) + \mathbb{P}\left( \sup_{W \in O(n; \mathbb{R})} \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}_j}(W) - \mathbb{E}f_{\bar{z}}(W) \geq n\delta \right)
\]

(117)

\[
\leq \mu(n, p) + \mathbb{P}\left( \sup_{W \in O(n; \mathbb{R})} \left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}_j}(W) - \mathbb{E}f_{\bar{z}}(W) \right| \geq n\delta \right).
\]

**Uniform bound over** \( O(n; \mathbb{R}) \). For any \( \epsilon > 0 \), lemma B.3 shows there exists an \( \epsilon \)-nets \( S_\epsilon \) that covers \( O(n; \mathbb{R}) \):

\[
S_\epsilon = \{ W_1, W_2, \ldots, W_{|S_\epsilon|} \}, \quad O(n; \mathbb{R}) \subset \bigcup_{l=1}^{|S_\epsilon|} \mathbb{B}(W_l, \epsilon),
\]

(118)

and \( |S_\epsilon| \) satisfies \( |S_\epsilon| \leq (6/\epsilon)^n \). Together with Lipschitz assumption 2, we know that

\[
\sup_{W \in \mathbb{B}(W_i, \epsilon)} \left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}_j}(W) - \mathbb{E}f_{\bar{z}}(W) \right|
\]

\[
\leq \sup_{W \in \mathbb{B}(W_i, \epsilon)} \left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}_j}(W_i) - \mathbb{E}f_{\bar{z}}(W_i) \right| + \sup_{W \in \mathbb{B}(W_i, \epsilon)} |\mathbb{E}f_{\bar{z}}(W_i) - \mathbb{E}f_{\bar{z}}(W)|
\]

\[
+ \sup_{W \in \mathbb{B}(W, \epsilon)} \left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}_j}(W_i) - \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}_j}(W) \right|
\]

\[
\leq \sup_{W \in \mathbb{B}(W_i, \epsilon)} \left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}_j}(W_i) - \mathbb{E}f_{\bar{z}}(W_i) \right| + (L_f + \bar{L}_f)\epsilon.
\]

(119)

Hence, let

\[
\epsilon = \frac{n\delta}{2(L_f + \bar{L}_f)}.
\]

(120)
we have
\[
\mathbb{P}\left( \sup_{W \in \mathcal{O}(n; \mathbb{R})} \left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}}(W) - \mathbb{E} f_{\bar{z}}(W) \right| \geq n\delta \right) \\
\leq \sum_{l=1}^{|S_x|} \mathbb{P}\left( \sup_{W \in \mathcal{B}(W_i; \epsilon)} \left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}}(W_l) - \mathbb{E} f_{\bar{z}}(W_l) \right| \geq n\delta \right) \\
< \left( \frac{6}{e} \right)^{n^2} \mathbb{P}\left( \sup_{W \in \mathcal{B}(W_i; \epsilon)} \left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}}(W_i) - \mathbb{E} f_{\bar{z}}(W_i) \right| \geq n\delta - (L_f + \bar{L}_f)\epsilon \right) \\
= \exp\left[ n^2 \ln \left( \frac{12(L_f + \bar{L}_f)}{n\delta} \right) \right] \mathbb{P}\left( \sup_{W \in \mathcal{B}(W_i; \epsilon)} \left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}}(W_i) - \mathbb{E} f_{\bar{z}}(W_i) \right| \geq \frac{n\delta}{2} \right). (121)
\]

**Tail bound within each** $\mathcal{B}(W_i, \epsilon)$. Then, \( \forall l \in [S_x] \), we apply point-wise control to a given point $W_l \in S$. Later we will provide a uniform concentration bound over $\mathcal{O}(n; \mathbb{R})$. By triangle inequality, we have
\[
\left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}}(W_l) - \mathbb{E} f_{\bar{z}}(W_l) \right| \leq \left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}}(W_l) - \mathbb{E} f_{\bar{z}}(W_l) \right| + \left| \mathbb{E} f_{\bar{z}}(W_l) - \mathbb{E} f_{\bar{z}}(W_l) \right|. (122)
\]

\( \forall W \in \mathcal{B}(W_i, \epsilon) \), by assumption 1 and 3, we have
\[
\left| \mathbb{E} f_{\bar{z}}(W) - \mathbb{E} f_{\bar{z}}(W) \right| = \left| \mathbb{E} \left[ f_{\bar{z}}(W) \cdot 1_{Z \neq \bar{z}} \right] \right| \leq \sqrt{\mathbb{E}^2 f_{\bar{z}}(W)} \left\| \mathbb{E} 1_{Z \neq \bar{z}} \right\| = \sqrt{\mathbb{E}^2 f_{\bar{z}}(W)} P(\max_{i,j} |z_{i,j}| > B) \leq \sqrt{R_2 \mu(n, p)}. \tag{123}
\]

Let $\rho$ be the lower bound of $p$, such that $\forall p > \rho$, we have $\sqrt{R_2 \mu(n, p)} < \frac{n\delta}{4}$. Hence, when $p > \rho$, we have
\[
\left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}}(W_l) - \mathbb{E} f_{\bar{z}}(W_l) \right| \leq \left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}}(W_l) - \mathbb{E} f_{\bar{z}}(W_l) \right| + \frac{n\delta}{4},. \tag{124}
\]

which implies
\[
\mathbb{P}\left( \left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}}(W) - \mathbb{E} f_{\bar{z}}(W) \right| \geq \frac{n\delta}{2} \right) \leq \mathbb{P}\left( \left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}}(W) - \mathbb{E} f_{\bar{z}}(W) \right| \geq \frac{n\delta}{4} \right) \leq 2 \exp\left( -\frac{pm^2\delta^2}{32R_2 + 8R_1 n\delta/3} \right). \tag{125}
\]

where the last inequality is achieved by Bernstein’s inequality (Lemma B.1), along with assumption 3 and $\mathbb{E} f_{\bar{z}}^2(W) \leq R_2$. Combine equation 125 and equation 121, we have
\[
\mathbb{P}\left( \sup_{W \in \mathcal{O}(n; \mathbb{R})} \left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}}(W) - \mathbb{E} f_{\bar{z}}(W) \right| \geq n\delta \right) \\
< \exp\left[ n^2 \ln \left( \frac{12(L_f + \bar{L}_f)}{n\delta} \right) \right] \mathbb{P}\left( \sup_{W \in \mathcal{B}(W_i; \epsilon)} \left| \frac{1}{p} \sum_{j=1}^{p} f_{\bar{z}}(W_i) - \mathbb{E} f_{\bar{z}}(W_i) \right| \geq \frac{n\delta}{2} \right) \leq 2 \exp\left( -\frac{pm^2\delta^2}{32R_2 + 8R_1 n\delta/3} \right) + n^2 \ln \left( \frac{12(L_f + \bar{L}_f)}{n\delta} \right) + \ln 2. \tag{126}
\]

26
Summary. Therefore, we conclude that

\[
P \left( \sup_{W \in O(n; \mathbb{R})} \frac{1}{np} \left| \sum_{j=1}^{p} [f_{z_j}(W) - \mathbb{E}f_{z_j}(W)] \right| \geq \delta \right) < \exp \left( -\frac{3p\delta^2}{c_1\theta + 8n(\ln p)^4} + n^2 \ln \left( \frac{60np(\ln p)^4}{\delta} \right) \right)
\]

\[
+ \exp \left( -\frac{p\delta^2}{c_2\theta} + n^2 \ln \left( \frac{60np(\ln p)^4}{\delta} \right) \right) + 2np\theta \exp \left( -\frac{(\ln p)^2}{2} \right) \leq \frac{1}{p},
\]

when \( p > \rho \).

\[ (127) \]

**Lemma B.5** (Concentration Bound of the Clean Objective over \( O(n; \mathbb{R}) \) \( \forall \theta \in (0, 1] \), if \( X \in \mathbb{R}^{n \times p}, x_{i,j} \sim_{iid} \text{BG}(\theta) \), for any \( \delta > 0 \), the following inequality holds

\[
P \left( \sup_{W \in O(n; \mathbb{R})} \frac{1}{np} \left\|WX\right\|_4^4 - \mathbb{E}\left\|WX\right\|_4^4 \left| \geq \delta \right. \right) < \exp \left( -\frac{3p\delta^2}{c_0\theta + 8n(\ln p)^4} + n^2 \ln \left( \frac{60np(\ln p)^4}{\delta} \right) \right)
\]

\[
+ \exp \left( -\frac{p\delta^2}{c_2\theta} + n^2 \ln \left( \frac{60np(\ln p)^4}{\delta} \right) \right) + 2np\theta \exp \left( -\frac{(\ln p)^2}{2} \right),
\]

for some constants \( c_1 > 10^4, c_2 > 3360 \). Moreover

\[
\exp \left( -\frac{3p\delta^2}{c_1\theta + 8n(\ln p)^4} + n^2 \ln \left( \frac{60np(\ln p)^4}{\delta} \right) \right)
\]

\[
+ \exp \left( -\frac{p\delta^2}{c_2\theta} + n^2 \ln \left( \frac{60np(\ln p)^4}{\delta} \right) \right) + 2np\theta \exp \left( -\frac{(\ln p)^2}{2} \right) \leq \frac{1}{p},
\]

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

**Proof** See Lemma 2.2 in Zhai et al. (2019b), note that the sparsity condition \( \theta \in (0, 1) \) of the original Lemma in Zhai et al. (2019b) can be easily generalized to \( \theta = 1 \).

\[ (129) \]

C ADDITIONAL EXPERIMENTAL RESULTS

![Figure 8: Representations of three 16 x 16 patches in both the clean and noisy images. Each selected patch is visualized, both with and without noise, and the 6 corresponding bases with largest absolute coefficients are shown.](image)
Figure 9: Representations of three $8 \times 8 \times 3$ colored patches in both the clean and noisy images. Each selected patch is visualized, both with and without noise, and the 6 corresponding bases with largest absolute coefficients are shown.

Figure 10: All $8 \times 8 \times 3 = 192$ bases learned from 100,000 random $8 \times 8$ colored patches sampled from the CIFAR-10 data-set. (a) Learned Bases from clean CIFAR-10; (b) Learned Bases from CIFAR-10 with Gaussian noise, SNR $= 6.23$; (c) Learned Bases from CIFAR-10 with 20% of Gaussian outliers; (d) Learned Bases from CIFAR-10 with 50% of sparse corruptions. For all learned bases, the resulting atoms are sorted according to the $\ell^1$-norm of their coefficients in the sparse code.