Complete Dictionary Recovery Over the Sphere II: Recovery by Riemannian Trust-Region Method

Ju Sun, Student Member, IEEE, Qing Qu, Student Member, IEEE, and John Wright, Member, IEEE

Abstract—We consider the problem of recovering a complete (i.e., square and invertible) matrix \( A_0 \), from \( Y \in \mathbb{R}^{n \times p} \) with \( Y = A_0 X_0 \), provided \( X_0 \) is sufficiently sparse. This recovery problem is central to theoretical understanding of dictionary learning, which seeks a sparse representation for a collection of input signals and finds numerous applications in modern signal processing and machine learning. We give the first efficient algorithm that provably recovers \( A_0 \) when \( X_0 \) has \( O(n) \) nonzeros per column, under suitable probability model for \( X_0 \). Our algorithmic pipeline centers around solving a certain nonconvex optimization problem with a spherical constraint, and hence is naturally phrased in the language of manifold optimization. In a companion paper, we have showed that with high probability, our nonconvex formulation has no “spurious” local minimizers and around any saddle point, the objective function has a negative directional curvature. In this paper, we take advantage of the particular geometric structure and describe a Riemannian trust-region algorithm that provably converges to a local minimizer of the problem.

Index Terms—Dictionary learning, nonconvex optimization, spherical constraint, escaping saddle points, trust-region method, manifold optimization, function landscape, second-order geometry, inverse problems, structured signals, nonlinear approximation.

I. INTRODUCTION

RECENTLY, there is a surge of research studying nonconvex formulations and provable algorithms for a number of central problems in signal processing and machine learning, including, e.g., low-rank matrix completion/recovery [4]–[25], phase retrieval [26]–[36], tensor recovery [37]–[41], mixed regression [42], [43], structured element pursuit [41], [44], blind deconvolution [45]–[49], noisy phase synchronization and community detection [50]–[52], deep learning [53], [54], numerical linear algebra and optimization [55], [56]. The research efforts are fruitful in producing more practical and scalable algorithms and even significantly better performance guarantees than known convex methods.

 Manuscript received November 29, 2015; revised September 8, 2016; accepted October 23, 2016. Date of publication November 23, 2016. This work was supported in part by ONR under Grant N00014-13-1-0492, in part by NSF under Grant 1343282, Grant CCF 1527809, and Grant IIS 1546411, and in part by the Moore and Sloan Foundations.

J. Sun was with the Columbia University, New York, NY 10027 USA. He is now with the Mathematics Department, Stanford University, Stanford, CA 94305 USA (e-mail: sunju@stanford.edu).
Q. Qu and J. Wright are with the Electrical Engineering Department, Columbia University, New York, NY 10027 USA (e-mail: qq2105@columbia.edu; jw2966@columbia.edu). Communicated by C. Caramanis, Associate Editor for Machine Learning.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2016.2632149

In a companion paper [3], we set out to understand the surprising effectiveness of nonconvex heuristics on the dictionary learning (DL) problem. In particular, we have focused on the complete dictionary recovery (DR) setting: given \( Y = A_0 X_0 \), with \( A_0 \in \mathbb{R}^{n \times n} \) complete (i.e., square and invertible), and \( X_0 \in \mathbb{R}^{n \times p} \) obeying an i.i.d. Bernoulli-Gaussian (BG) model with rate \( \theta \) (i.e., \( |X_0|_{ij} = \Omega_{ij} z_{ij} \) with \( \Omega_{ij} \sim \mathrm{Ber}(\theta) \) and \( z_{ij} \sim \mathcal{N}(0, 1) \)), recover \( A_0 \) and \( X_0 \). In this setting, \( \text{row}(Y) = \text{row}(X_0) \), where \( \text{row}(\cdot) \) denotes the row space. To first recover rows of \( X_0 \), we have tried to find the sparsest vectors in \( \text{row}(Y) \), and proposed solving the nonconvex formulation

\[
\min_{q} f(q; Y) = \frac{1}{p} \sum_{k=1}^{p} h_{\mu}(q^\top \tilde{y}_k) \quad \text{s.t.} \quad q \in \mathbb{S}^{n-1},
\]

where \( \tilde{Y} \) is a proxy of \( Y \) (i.e., after appropriate processing), \( \tilde{y}_k \) is the \( k \)-th column of \( \tilde{Y} \), and \( h_{\mu}(z) = \mu \log \cosh(z/\mu) \) is a (convex) smooth approximation to the absolute-value function. The spherical constraint renders the problem nonconvex.

Despite the apparent nonconvexity, our prior analysis in [3] has showed that all local minimizers of (I.1) are qualitatively equally good, because each of them produces a close approximation of a certain row of \( X_0 \) (see Corollary II.4). So the central issue is how to escape from saddle points. Fortunately, our previous results ([3, Th. II.3]) imply that all saddle points under consideration are “ridable,” i.e., the associated Hessians have both strictly positive and strictly negative values (see the recapitulation in Section II-B). Particularly, eigenvectors of the negative eigenvalues are direction of negative curvature, which intuitively serve as directions of local descent.

Second-order methods can naturally exploit the curvature information to escape from ridable saddle points. To gain some intuition, consider an unconstrained optimization problem

\[
\min_{x \in \mathbb{R}^n} \phi(x).
\]

The second-order Taylor expansion of \( \phi \) at a saddle point \( x_0 \) is

\[
\tilde{\phi}(\delta; x_0) = \phi(x_0) + \Delta \delta^\top \nabla^2 \phi(x_0) \delta.
\]

When \( \delta \) is chosen to align with an eigenvector of a negative eigenvalue \( \lambda_{\text{neg}}[\nabla^2 \phi(x_0)] < 0 \), it holds that

\[
\tilde{\phi}(\delta; x_0) - \phi(x_0) \leq -\|\lambda_{\text{neg}}[\nabla^2 \phi(x)]\| \delta^2.
\]

Thus, minimizing \( \tilde{\phi}(\delta; x_0) \) returns a direction \( \delta_* \) that tends to decrease the objective \( \phi \), provided local approximation of \( \tilde{\phi} \) to \( \phi \) is reasonably accurate. Based on this intuition, we
derive a (second-order) Riemannian trust-region algorithm that exploits the second-order information to escape from saddle points and provably returns a local minimizer to (I.1), from arbitrary initializations. We provide rigorous guarantees for recovering a local minimizer in Section II.

Obtaining a local minimizer only helps approximate one row of $X_0$. To recover the row, we derive a simple linear programming rounding procedure that provably works. To recover all rows of $X_0$, one repeats the above process based on a carefully designed deflation process. The whole algorithmic pipeline and the related recovery guarantees are provided in Section III. Particularly, we show that when $p$ is reasonably large, with high probability (w.h.p.), our pipeline efficiently recovers $A_0$ and $X_0$, even when each column of $X_0$ contains $O(n)$ nonzeros.

A. Prior Arts and Connections

In Section II.E of the companion paper [3], we provide detailed comparisons of our results with prior theoretical results on DR; we conclude that this is the first algorithmic framework that guarantees efficient recovery of complete dictionaries when the coefficients have up to constant fraction of nonzeros. We also draw methodological connections to work on understanding nonconvex heuristics, and other nonconvex problems with similar geometric structures. Here we focus on drawing detailed connections to the optimization literature.

Trust-region method (TRM) has a rich history dating back to 40’s; see the monograph [57] for accounts of the history and developments. The main motivation for early developments was to address limitations of the classic Newton’s method (see, e.g., [58, Sec. 3]). The limitations include the technical subtleties to establish local and global convergence results. Moreover, when the Hessian is singular or indefinite, the movement direction is either not well-defined, or does not improve the objective function. [59]–[64] initialized the line of work that addresses the limitations. Particularly, [58] and [65] proposed using local second-order Taylor approximation as model function in the trust-region framework for unconstrained optimization. They showed that under mild conditions, the trust-region iterate sequence has a limit point that is critical and has positive semidefinite Hessian; see also [57, Secs. 6.5 and 6.6]. Upon inspecting the relevant proofs, it seems not hard to strengthen the results to sequence convergence to local minimizers, under a ridable saddle condition. Nonetheless, for our particular function, noisy/stochastic gradient method guarantees tee convergence to a critical point. Nonetheless, for our particular problem, their result is considerably pessimistic, compared to our convergence result obtained from a specialized analysis.

Solving the trust-region subproblem exactly is expensive. Practically, often a reasonable approximate solution with controlled quality is adequate to guarantee convergence. In this regard, the truncated conjugate gradient (tCG) solver with a good initial search direction is commonly employed in practice (see, e.g., [57, Sec. 7.5]). To ensure ridable saddle points be properly escaped from, the eigenpoint idea (see, e.g., [57, Sec. 6.6]) is particularly relevant; see also [72, Algorithm 3] and [72, Lemma 10].

The benign function landscape we characterized in the first paper allows any reasonable iterative method that is capable of escaping from ridable saddles to find a local minimizer, with possibly different performance guarantees. The trust-region method we focus on here, and the curvilinear search method [61] are second-order methods that guarantee global optimization from arbitrary initializations. Typical first-order methods such as the vanilla gradient descent can only guarantee convergence to a critical point. Nonetheless, for our particular function, noisy/stochastic gradient method guarantees to find a local minimizer from an arbitrary initialization with high probability [74].

B. Notations, and Reproducible Research

We use bold capital and small letters such as $X$ and $x$ to denote matrices and vectors, respectively. Small letters are reserved for scalars. Several specific mathematical objects we will frequently work with: $O_k$ for the orthogonal group of order $k$, $S^{n-1}$ for the unit sphere in $\mathbb{R}^n$, $E^n$ for the unit ball in $\mathbb{R}^n$, and $[m] = \{1, \ldots, m\}$ for positive integers $m$. We use $(\cdot)^*$ for matrix transposition, causing no confusion as we will work entirely on the real field. We use superscript to index rows of a matrix, such as $x^i$ for the $i$-th row of the matrix $X$, and subscript to index columns, such as $x_j$. All vectors are defaulted to column vectors. So the $i$-th row of $X$ as a row vector will be written as $(x^i)^T$. For norms, $\|\cdot\|$ is the usual $\ell^2$ norm for a vector and the operator norm (i.e., $\ell^2 \to \ell^2$) for a matrix; all other norms will be indexed by subscript,
II. FINDING ONE LOCAL MINIMIZER VIA THE RIEmannIAN TRUST-REGION Method

We are interested to seek a local minimizer of (1.1). The presence of saddle points have motivated us to develop a second-order Riemannian trust-region algorithm over the sphere; the existence of descent directions at nonoptimal points drives the trust-region iteration sequence towards one of the minimizers asymptotically. We will prove that under our modeling assumptions, this algorithm with an arbitrary initialization efficiently produces an accurate approximation\(^1\) to one of the minimizers. Throughout the exposition, basic knowledge of Riemannian geometry is assumed. We will try to keep the technical requirement minimal possible; the reader can consult the excellent monograph [68] for relevant background and details.

A. Some Basic Facts About the Sphere and \(f\)

For any point \(q \in S^{n-1}\), the tangent space \(T_qS^{n-1}\) and the orthoprojector \(P_{T_qS^{n-1}}\) onto \(T_qS^{n-1}\) are given by

\[
T_qS^{n-1} = \{ \delta \in \mathbb{R}^n : q^* \delta = 0 \}, \\
P_{T_qS^{n-1}} = I - qq^* = UU^*,
\]

where \(U \in \mathbb{R}^{n \times (n-1)}\) is an arbitrary orthonormal basis for \(T_qS^{n-1}\) (note that the orthoprojector is independent of the basis \(U\) we choose). Consider any \(\delta \in T_qS^{n-1}\). The map

\[
\gamma(t) : t \mapsto q \cos(t \|\delta\|) + \frac{\delta}{\|\delta\|} \sin(t \|\delta\|)
\]

defines a smooth curve on the sphere that satisfies \(\gamma(0) = q\) and \(\gamma'(0) = \delta\). Geometrically, \(\gamma(t)\) is a segment of the great circle that passes \(q\) and has \(\delta\) as its tangent vector at \(q\). The exponential map for \(\delta\) is defined as

\[
\exp_q(\delta) = \gamma(1) = q \cos \|\delta\| + \frac{\delta}{\|\delta\|} \sin \|\delta\|. 
\]

\(^1\)By “accurate” we mean one can achieve an arbitrary numerical accuracy \(\varepsilon > 0\) with a reasonable amount of time. Here the running time of the algorithm is on the order of \(\log \log(1/\varepsilon)\) in the target accuracy \(\varepsilon\), and polynomial in other problem parameters.
Similarly, following the notation of [68], we denote the inverse tangent space like a full-rank matrix. If a matrix, whose restriction to \( T_q S^{n-1} \) is the inverse of the parallel translation operator (the detailed derivation can be found in [68, Ch. 8.1]):

\[
P^\tau_{\gamma}^{-0} : T_q S^{n-1} \rightarrow T_{\gamma}(T_q S^{n-1})
\]

denotes the parallel translation operator, which translates the tangent vector \( \delta \) at \( q = \gamma(0) \) to a tangent vector at \( \gamma(\tau) \), in a “parallel” manner. In the sequel, we identify \( P^\tau_{\gamma}^{-0} \) with the following \( n \times n \) matrix, whose restriction to \( T_q S^{n-1} \) is the parallel translation operator (the detailed derivation can be found in [68, Ch. 8.1]):

\[
P^\tau_{\gamma}^{-0} = \begin{pmatrix}
I - \frac{\delta \delta^*}{\| \delta \|^2} & -q \sin(\| \delta \|) \frac{\delta^*}{\| \delta \|} + \frac{\delta}{\| \delta \|} \cos(\| \delta \|) \frac{\delta^*}{\| \delta \|}
\end{pmatrix}
\]

\[
= I + (\cos(\| \delta \|) - 1) \frac{\delta \delta^*}{\| \delta \|^2} - \sin(\| \delta \|) \frac{q \delta^*}{\| \delta \|}.
\]

(II.8)

Similarly, following the notation of [68], we denote the inverse of this matrix by \( P^0_{\gamma - \tau} \), where its restriction to \( T_{\gamma}(T_q S^{n-1}) \) is the inverse of the parallel translation operator \( P^\tau_{\gamma}^{-0} \).

**B. The Geometric Results From [3]**

We reproduce the main geometric theorems from [3] here for the sake of completeness. To characterize the function landscape of \( f(q; X_0) \) over \( S^{n-1} \), we mostly work with the function

\[
g(w; X_0) = f(q; w; X_0) = \frac{1}{p} \sum_{k=1}^{p} h_{\mu}(q(w)^*X_0k),
\]

(II.9)

induced by the reparametrization

\[
q(w) = \left(w, \sqrt{1 - \|w\|^2}\right), \quad w \in S^{n-1}.
\]

(II.10)

Geometrically, this corresponds to projection of the function \( f \) above the equatorial section \( e_n^\perp \) onto \( e_n^\perp \) (see Fig. 2 (right) for illustration). In particular, we focus our attention to the smaller set of the ball:

\[
\Gamma = \left\{ w : \|w\| < \sqrt{\frac{4n-1}{4n}} \right\} \subseteq S^{n-1}
\]

(II.11)

because \( q(\Gamma) \) contains all points \( q \in S^{n-1} \) with \( n \in \arg \max_{e_i^\perp \in \{q^*e_i\}} \). We can similarly characterize other parts of \( f \) on \( S^{n-1} \) using projection onto other equatorial sections.

**Theorem 1 (High-Dimensional Landscape - Orthogonal Dictionary):** Suppose \( A_0 = I \) and hence \( Y = A_0X_0 = X_0 \). There exist positive constants \( c_a \) and \( C \), such that for any \( \theta \in (0, 1/2) \) and \( \mu < c_a \min\{\theta n^{-1}, n^{-5/4}\} \), whenever

\[
p \geq C \frac{n^3}{\mu^2} \log \frac{n}{\mu^2},
\]

(II.12)

the following hold simultaneously with probability at least \( 1 - c_b p^{-6} \):

\[
\nabla^2 g(w; X_0) \succeq \frac{c_a \theta}{4} I
\]

(II.13)

\[
\frac{w^* \nabla g(w; X_0)}{\|w\|} \succeq c_a \theta
\]

(II.14)

if \( \|w\| \leq \frac{\mu}{4\sqrt{2}} \) and \( \|w\| \leq \frac{1}{20\sqrt{5}} \),

---

2Note that the \( n \times n \) matrix \( \text{Hess} f(q; \tilde{Y}) \) has rank at most \( n - 1 \), as the nonzero \( q \) obviously is in its null space. When \( \text{Hess} f(q; \tilde{Y}) \) has rank \( n - 1 \), it has no null direction in the tangent space. Thus, in this case it acts on the tangent space like a full-rank matrix.
\[
\frac{w^* \nabla^2 g (w; X_0) w}{\|w\|^2} \leq -c_\theta 
\]
if \( \frac{1}{20\sqrt{5}} \leq \|w\| \leq \sqrt{\frac{4n-1}{4n}}, \tag{II.15} \]
and the function \( g(w; X_0) \) has exactly one local minimizer \( w_* \) over the open set \( \Gamma = \{ w : \|w\| < \sqrt{\frac{4n-1}{4n}} \} \), which satisfies
\[
\|w_* - 0\| \leq \min \left\{ \frac{c_\mu}{\theta} \sqrt{\frac{n \log p}{p}}, \frac{\mu}{16} \right\}. \tag{II.16} \]
Here \( c_\alpha \) through \( c_\epsilon \) are all positive constants.

Recall that the reason we just need to characterize the geometry for the case \( A_0 = I \) is that for other orthogonal \( A_0 \), the function landscape is simply a rotated version of that of \( A_0 = I \).

Theorem 2 (High-Dimensional Landscape - Complete Dictionary): Suppose \( A_0 \) is complete with its condition number \( \kappa(A_0) \). There exist positive constants \( c_\alpha \), (particularly, the same constant as in Theorem 1) and \( C \), such that for any \( \theta \in (0, 1/2) \) and \( \mu < c_\alpha \min \{\frac{1}{2}, \frac{5}{16} \} \), when
\[
p \geq \frac{C}{c_\alpha^2 \theta^2} \max \left\{ \frac{n^4}{\mu^4}, \frac{n^5}{\mu^5} \right\} \kappa^8(A_0) \log^4 \left( \frac{\kappa(A_0) n}{\mu} \right), \tag{II.17} \]
and \( Y = \sqrt{\mu} \cdot (Y^*)^{-1/2} Y, \ U \Sigma \Sigma^* = \Sigma^2 \cdot (A_0) \), the following hold simultaneously with probability at least \( 1 - c_b p^{-6} \):
\[
\nabla^2 g (w; UU^* Y) \geq \frac{c_\alpha \theta}{2 \mu} I \]
if \( \|w\| \leq \frac{\mu}{2 \sqrt{\theta}} \),
\[
\frac{w^* \nabla g (w; UU^* Y) w}{\|w\|^2} \geq \frac{1}{2} c_\theta \]
if \( \|w\| \leq \frac{\mu}{4 \sqrt{2}} \),
\[
\frac{w^* \nabla^2 g (w; UU^* Y) w}{\|w\|^2} \leq \frac{1}{2} c_\theta \]
if \( \|w\| \leq \frac{\mu}{20 \sqrt{5}} \),
and the function \( g(w; UU^* Y) \) has exactly one local minimizer \( w_* \) over the open set \( \Gamma = \{ w : \|w\| < \sqrt{\frac{4n-1}{4n}} \} \), which satisfies
\[
\|w_* - 0\| \leq \frac{\mu}{7}. \tag{II.21} \]
Here \( c_\alpha \) and \( c_b \) are both positive constants.

From the above theorems, it is clear that for any saddle point in the \( w \) space, the Hessian of \( g \) has at least one negative eigenvalue. Now the problem is whether all saddle points of \( f \) on \( \mathbb{S}^{n-1} \) are “ridable”, because as alluded to in previous discussion, we need to perform actual optimization in the \( q \) space. Instead of presenting a rigorous technical statement and detailed proof, we include here just an informal argument; our actual proof for algorithmic convergence runs back and forth in \( w \) and \( q \) space and such lack will not affect our arguments there.

It is very easy to verify the following fact (see proof of Lemma 11 on page 907):
\[
(\nabla f(q), q - e_n/q_n) = (w, \nabla g(w)).
\]
Thus, \( (\nabla f(q), q - e_n/q_n) \neq 0 \) if and only if \( (w, \nabla g(w)) \neq 0 \), implying that \( \nabla f(q) \) will never be zero in the spherical region corresponding to \( \{ w : \|w\| \leq \frac{\mu}{(4\sqrt{2})} \} \). Moreover, it is shown in Lemma 10 below that the Riemannian Hessian is positive definite for the spherical region corresponding to \( \{ w : \|w\| \leq \frac{\mu}{(4\sqrt{2})} \} \), so there is no saddle point in this region either. Over \( q(1) \cap \mathbb{S}^{n-1} \), potential saddle points lie only in the region corresponding to \( \{ w : \|w\| \leq \frac{\mu}{(4\sqrt{2})} \} \). Theorem 1 and Theorem 2 imply that around each point in this region, a cross section of the function \( g(w) \) is strictly concave locally. Intuitively, by the \( q(w) \) mapping the same happens in the \( q \) space, i.e., the Riemannian Hessian has a strictly negative eigenvalue.

C. The Riemannian Trust-Region Algorithm Over the Sphere

For a function \( f \) in the Euclidean space, the typical TRM starts from some initialization \( q^{(0)} \in \mathbb{R}^n \), and produces a sequence of iterates \( q^{(1)}, q^{(2)}, \ldots \), by repeatedly minimizing a quadratic approximation \( \hat{f} \) to the objective function \( f(q) \), over a ball centered around the current iterate.

For our \( f \) defined over \( \mathbb{S}^{n-1} \), given the previous iterate \( q^{(r-1)} \), the TRM produces the next movement by generating a solution \( \hat{\delta} \) to
\[
\min_{\delta \in T_{q^{(r-1)}} \mathbb{S}^{n-1}} \|\delta\| \leq \Delta \hat{f}(\delta; q^{(r-1)}), \tag{II.22} \]
where \( \hat{f}(\delta; q^{(r-1)}) \) is the local quadratic approximation defined in (II.1). The solution \( \hat{\delta} \) is then pulled back to \( \mathbb{S}^{n-1} \) from \( T_{q^{(r-1)}} \mathbb{S}^{n-1} \). If we choose the exponential map to pull back the movement \( \hat{\delta} \), the next iterate then reads
\[
q^{(r)} = q^{(r-1)} \cos \|\hat{\delta}\| + \frac{\hat{\delta}}{\|\hat{\delta}\|} \sin \|\hat{\delta}\|. \tag{II.23} \]
To solve the subproblem (II.22) numerically, we can take any matrix \( U \in \mathbb{R}^{n \times (n-1)} \), whose columns form an orthonormal basis for \( T_{q^{(r-1)}} \mathbb{S}^{n-1} \), and produce a solution \( \hat{\xi} \) to
\[
\min_{\|\xi\| \leq \Delta} \hat{f}(U\xi; q^{(r-1)}). \tag{II.24} \]
Solution to (II.22) can then be recovered as \( \hat{\delta} = U\hat{\xi} \).

The problem (II.24) is an instance of the classic trust region subproblem, i.e., minimizing a quadratic function subject to a single quadratic constraint. Albeit potentially nonconvex, this notable subproblem can be solved in polynomial time.

\[\text{The exponential map is only one of the many possibilities; also for general manifolds other retraction schemes may be more practical. See exposition on retraction in [68, Ch. 4].}\]
by several numerical methods [57], [65], [75]–[78]. Approximate solution of the subproblem suffices to guarantee convergence in theory, and lessens the storage and computational burden in practice. We will deploy the approximate version in simulations. For simplicity, however, our subsequent analysis assumes the subproblem is solved exactly. We next briefly describe how one can deploy the semidefinite programming (SDP) approach [75]–[78] to solve the subproblem exactly. This choice is due to the well-known effectiveness and robustness of the SDP approach on this problem. We introduce

$$
\hat{\xi} = [\xi^*, 1]^\top, \quad \Theta = \hat{\xi} \hat{\xi}^\top, \quad M = \begin{bmatrix} A & b \end{bmatrix} b^* 0
$$

(II.25)

where $A = \nabla f(q^{(r-1)}; \hat{Y}) U$ and $b = U^* \nabla f(q^{(r-1)}; \hat{Y})$. The resulting SDP to solve is

$$
\min_{\Theta, \lambda} \frac{1}{2} \Tr(\Theta) \leq \Delta^2 + 1, \quad \langle E_{n+1}, \Theta \rangle = 1, \quad \Theta \succeq 0.
$$

(II.26)

where $E_{n+1} = e_{n+1} e_{n+1}^\top$. Once the problem (II.26) is solved to its optimum $\Theta_*$, one can provably recover the minimizer $\hat{\xi}_*$ of (II.24) by computing the SVD of $\Theta_* = U \Sigma \hat{V}^\top$, and extracting as a subvector the first $n - 1$ coordinates of the principal eigenvector $\hat{u}_1$ (see [79, Appendix B]).

### D. Main Convergence Results

Using general convergence results on Riemannian TRM (see, e.g., [68, Ch. 7]), it is not difficult to prove that the gradient sequence $\nabla f(q^{(r)}; \hat{Y})$ produced by TRM converges to zero (i.e., global convergence), or the sequence converges (at quadratic rate) to a local minimizer if the initialization is already close a local minimizer (i.e., local convergence). In this section, we show that under our probabilistic assumptions, these results can be substantially strengthened. In particular, the algorithm is guaranteed to produce an accurate approximation to a local minimizer of the objective function, in a number of iterations that is polynomial in the problem size, from arbitrary initializations. The arguments in the companion paper [3] showed that w.h.p. every local minimizer of $f$ produces a close approximation to a row of $X_0$. Taken together, this implies that the algorithm efficiently produces a close approximation to one row of $X_0$.

Thorough the analysis, we assume the trust-region subproblem is exactly solved and the step size parameter $\Delta$ is fixed. Our next two theorems summarize the convergence results for orthogonal and complete dictionaries, respectively.

#### Theorem 3 (TRM Convergence - Orthogonal Dictionary): Suppose the dictionary $A_0$ is orthogonal. There exists a positive constant $C$, such that for all $\theta \in (0, 1/2)$ and $\mu < c_\theta \min \{\theta n^{-1}, n^{-5/4}\}$, whenever

$$
p \geq \frac{C}{\mu \theta^2} n^3 \log n \mu \theta,
$$

with probability at least $1 - c_\theta p^{-6}$, the Riemannian trust-region algorithm with input data matrix $Y = Y_0$, any initialization $q^{(0)}$ on the sphere, and a step size satisfying

$$
\Delta \leq \frac{c_\theta c_\Delta^2 \mu^2}{n^{7/2} \log^{7/2}(np)}
$$

(II.27)

returns a solution $\hat{q} \in S^{n-1}$ which is $e$ near to one of the local minimizers $q_*$, i.e., $\|\hat{q} - q_*\| \leq e$ in at most

$$
\max \left\{ \frac{c_\theta n^6 \log^3(n p)}{c_\theta^2 \theta^3 \mu \Delta^4}, \frac{c_\theta n}{c_\theta^2 \theta^2 \Delta^2} \right\} f(q^{(0)}) + \log \log \frac{c_\theta c_\Delta^2 \mu}{n^{7/2} \log^{7/2}(np)}
$$

(II.28)

iterations. Here $c_\theta$ is as defined in Theorem 1, and $c_\Delta$ through $c_f$ are all positive constants.

#### Theorem 4 (TRM Convergence - Complete Dictionary): Suppose the dictionary $A_0$ is complete with condition number $\kappa(A_0)$. There exists a positive constant $C$, such that for all $\theta \in (0, 1/2)$, and $\mu < c_\theta \min \{\theta n^{-1}, n^{-5/4}\}$, whenever

$$
p \geq \frac{C}{\mu \theta^2} \max \left\{ n^3 \mu^5 \kappa^8(A_0) \log \left( \frac{\kappa(A_0) n}{\mu \theta} \right) \right\}
$$

with probability at least $1 - c_\theta p^{-6}$, the Riemannian trust-region algorithm with input data matrix $Y = Y_0$, any initialization $q^{(0)}$ on the sphere and a step size satisfying

$$
\Delta \leq \frac{c_\theta c_\Delta^2 \mu^2}{n^{7/2} \log^{7/2}(np)}
$$

(II.29)

returns a solution $\hat{q} \in S^{n-1}$ which is $e$ near to one of the local minimizers $q_*$, i.e., $\|\hat{q} - q_*\| \leq e$ in at most

$$
\max \left\{ \frac{c_\theta n^6 \log^3(n p)}{c_\theta^2 \theta^3 \mu \Delta^4}, \frac{c_\theta n}{c_\theta^2 \theta^2 \Delta^2} \right\} f(q^{(0)}) + \log \log \frac{c_\theta c_\Delta^2 \mu}{n^{7/2} \log^{7/2}(np)}
$$

(II.30)

iterations. Here $c_\theta$ is as in Theorem 1, and $c_\Delta$ through $c_f$ are all positive constants.

Our convergence result shows that for any target accuracy $e > 0$ the algorithm terminates within polynomially many steps. Specifically, the first summand in (II.28) or (II.30) is the number of steps the sequence takes to enter the strongly convex region and be “reasonably” close to a local minimizer. All subsequent trust-region subproblems are then unconstrained (proved below) – the constraint is inactive at optimal point, and hence the steps behave like Newton steps. The second summand reflects the typical quadratic local convergence of the Newton steps.

Our estimate of the number of steps is pessimistic: the running time is a relatively high-degree polynomial in $p$ and $n$. We will discuss practical implementation details that help speed up in Section IV. Our goal in stating the above results is not to provide a tight analysis, but to prove that the Riemannian TRM algorithm finds a local minimizer in polynomial time. For nonconvex problems, this is not entirely trivial – results of [80] show that in general it is NP-hard to find a local minimizer of a nonconvex function.

#### E. Sketch of Proof for Orthogonal Dictionaries

The reason that our algorithm is successful derives from the geometry formalized in Theorem 1. Basically, the sphere $S^{n-1}$ can be divided into three regions. Near each local minimizer,
the function is strongly convex, and the algorithm behaves like a standard (Euclidean) TRM algorithm applied to a strongly convex function — in particular, it exhibits a quadratic asymptotic rate of convergence. Away from local minimizers, the function always exhibits either a strong gradient, or a direction of negative curvature (i.e., the Hessian has a strictly negative eigenvalue). The Riemannian TRM algorithm is capable of exploiting these quantities to reduce the objective value by at least a fixed amount in each iteration. The total number of iterations spent away from the vicinity of the local minimizers can be bounded by comparing this amount to the initial objective value. Our proofs follow exactly this line and make the various quantities precise.

Note that for any orthogonal $A_0$, $f(q; A_0x_0) = f(A_0^Tq; x_0)$. In words, this is the established fact that the function landscape of $f(q; A_0x_0)$ is a rotated version of that of $f(q; x_0)$. Thus, any local minimizer $q_*$ of $f(q; x_0)$ is rotated to $A_0q_*$, a local minimizer of $f(q; A_0x_0)$. Also if our algorithm generates iteration sequence $q_0, q_1, q_2, \ldots$ for $f(q; x_0)$ upon initialization $q_0$, it will generate the iteration sequence $A_0q_0, A_0q_1, A_0q_2, \ldots$ for $f(q; A_0x_0)$.

We partition the sphere into three regions, for which we label as $R_1, R_2, R_3$, corresponding to the strongly convex, nonzero gradient, and negative curvature regions, respectively (see Theorem 1). That is, $R_1$ consists of a union of $2n$ spherical caps of radius $\mu/(4\sqrt{2})$, each centered around a signed standard basis vector $\pm e_i$. $R_2$ consists of the set difference of a union of $2n$ spherical caps of radius $1/(20\sqrt{5})$, centered around the standard basis vectors $\pm e_i$, and $R_3$. Finally, $R_3$ covers the rest of the sphere. We say a trust-region step takes an $R_1$ step if the current iterate is in $R_2$; similarly for $R_2$ and $R_3$ steps. Since we use the geometric structures derived in [3, Th. 1] and [3, Corollary II.2], the conditions

$$\theta \in (0, 1/2), \quad \mu < c \min \{\theta^{-1}, n^{-5/4}\},$$

$$p \geq \frac{c L}{\mu^2 n^3} \log \frac{n}{\mu \theta}$$

are always in force.

At step $r$ of the algorithm, suppose $\delta^{(r)}$ is the minimizer of the trust-region subproblem (II.22). We call the step “constrained” if $\|\delta^{(r)}\| = \Delta$ (the minimizer lies on the boundary and hence the constraint is active), and call it “unconstrained” if $\|\delta^{(r)}\| < \Delta$ (the minimizer lies in the relative interior and hence the constraint is not in force). Thus, in the unconstrained case the optimality condition is (II.7).

The next lemma provides some estimates about $\nabla f$ and $\nabla^2 f$ that are useful in various contexts.

**Lemma 5:** We have the following estimates about $\nabla f$ and $\nabla^2 f$:

$$\sup_{q \in S^{n-1}} \|\nabla f(q)\| \preceq M \sqrt{n} \|x_0\|, \quad \sup_{q \in S^{n-1}} \|\nabla^2 f(q)\| \preceq M \sqrt{n} \|x_0\|^2.$$
Lemma 7: A trust-region step decreases the objective value by at least a given iterate $r$, where $c_a$ to $c_e$ are positive constants, and $c_\star$ as defined in Theorem 1.

Proof: We only consider the symmetric section in the vicinity of $e_n$ and the claims carry on to others by symmetry. If the current iterate $q^{(r)}$ is in the region $R_{11}$, by Theorem 1, w.h.p., we have $w^\top g(w)/\|w\|^2 \geq c_e \theta$ for the constant $c_e$. By Proposition 6 and Lemma 31, w.h.p., $w^\top g(w)/\|w\|^2$ is $C_{n^2} \log (np)/\mu$-Lipschitz. Therefore, By Lemma 6 and Lemma 7, a trust-region step decreases the objective value by at least

$$d_{11} = \frac{1}{2} \min \left( \frac{c_e^2 c_a \theta^2 \mu^2}{2 (3 \Delta^2 \log (np)/\mu^2)}, \frac{3 \Delta^2 c_a \theta}{8 \pi^2} \right),$$

$$d_{111} = \frac{1}{2} \min \left( \frac{c_e^2 c_a \theta^2 \mu^2}{2 (3 \Delta^2 \log (np)/\mu^2)}, \frac{3 \Delta^2 c_a \theta}{8 \pi^2} \right) \tag{I.32}$$

respectively, provided that

$$\Delta < \frac{c_e c_a \theta^2 \mu^2}{n^{5/2} \log (np)/\mu^2},$$

(II.33)

where $c_a$ to $c_e$ are positive constants, and $c_\star$ as defined in Theorem 1.

Proposition 1: Assume (II.31). In regions $R_{11}$ and $R_{111}$, each trust-region step decreases the objective value by at least

$$f(\exp_{q^{(r)}}(\delta)) \leq f(q^{(r)}) - \frac{m_H \Delta^2}{M_H} + \frac{1}{6} \eta f \Delta^3,$$

where $\eta f$ is defined the same as Lemma 6.

Proof: See Page 903 under Section VI.

The next lemma provides an estimate of $m_H$. Again we will only state the result for the “canonical” section with the “canonical” $q(w)$ mapping.

Lemma 10: There exists a positive constant $C$, such that for all $\theta \in (1/2, 1/12)$ and $\mu < \theta /10$, whenever $p \geq C n^3 \log \mu^2 / (\mu^2)$, it holds with probability at least $1 - cp^{-\eta}$ that for all $q$ with $\|w(q)\| \leq \mu / (4 \sqrt{2})$.

$$\text{Hess } f(\delta) \geq (\theta \mu / \mu^2) \cdot \mathcal{P}_{q(w)},$$

where $c_\star$ is as in Theorem 1 and Theorem 2, and $c > 0$ is another constant.

Proof: See Page 904 under Section VI.

We know that $\|X_0\| \leq 4 \log (1/2) (np)$ w.h.p., and hence by the definition of Riemannian Hessian and Lemma 5,

$$M_H \geq \|\text{Hess } f(q)\| \leq \|w^2 f(q)\| + \|\nabla f(q)\| \leq M_{q^2} + M_{q^2} \leq \frac{2n}{\mu} \|X_0\|^2 \leq \frac{16n}{\mu} \log (np).$$

Combining this estimate and Lemma 10 and Lemma 6, we obtain a concrete lower bound for the reduction of objective value for each constrained $R_1$ step.

Proposition 2: Assume (II.31). Each constrained $R_1$ trust-region step (i.e., $\|\delta\| = \Delta$) reduces the objective value by at least

$$d_\Delta = \frac{c e^2 \theta^2 \mu^2}{3 \mu^2} (\mu / \mu)^2,$$

provided

$$\Delta \leq \frac{c_e^2 \theta^2 \mu^2}{n^{5/2} \log (np)/\mu^2},$$

(II.35)

where $c_\star$ and $c_e$ are positive constants.

Proof: We only consider the symmetric section in the vicinity of $e_n$ and the claims carry on to others by symmetry. We have that w.h.p.

$$\|\text{Hess } f(q)\| \leq \frac{16n}{\mu} \log (np),$$

$$\|\text{Hess } f(q)\| \geq (\theta \mu / \mu^2) \cdot \mathcal{P}_{q(w)}.$$
Finally, by the condition on $\Delta$ in (II.35) and the assumed conditions (II.31), we obtain
\[
d_t \geq \frac{c_2^2 \theta^2}{2 \mu \log(np)} \Delta^2 - \frac{c_0 n^{3/2} \log^{3/2}(np)}{\mu^2} \Delta^3
\geq \frac{c_2^2 \theta^2}{4 \mu n \log(np)} \Delta^2,\
\]
as desired.

By the proof strategy for $R_1$ we sketched before Lemma 9, we expect the iteration sequence ultimately always takes unconstrained steps when it moves very close to a local minimizer. We will show that the following is true: when $\Delta$ is small enough, once the iteration sequence starts to take unconstrained $R_1$ step, it will take consecutive unconstrained $R_1$ steps afterwards. It takes two steps to show this: (1) upon an unconstrained $R_1$ step, the next iterate will stay in $R_\Delta$; (2) when $\Delta$ is small enough, the step is in fact unconstrained. Again we will only state the result for the “canonical” section with the “canonical” $q(w)$ mapping. The next lemma exhibits an absolute lower bound for magnitudes of gradients in $R_\Delta$.

**Lemma 11:** For all $q$ satisfying $\mu/(4\sqrt{2}) \leq \|w(q)\| \leq 1/(20\sqrt{3})$, it holds that
\[
\|\nabla f(q)\| \geq \frac{9}{10} \frac{w^* \nabla g(w)}{\|w\|}.
\]

**Proof:** See Page 907 under Section VI.

Assuming (II.31), Theorem 1 gives that w.h.p. $w^* \nabla g(w)/\|w\| \geq c_\theta$. Thus, w.h.p., $\|\nabla f(q)\| \geq 9c_\theta/10$ for all $q \in R_\Delta$. The next lemma compares the magnitudes of gradients before and after taking one unconstrained $R_1$ step. This is crucial to providing upper bound for magnitude of gradient for the next iterate, and also to establishing the ultimate (quadratic) sequence convergence.

**Lemma 12:** Suppose the trust-region size $\Delta \leq 1$, and at a given iterate $r$, $\text{Hess} f(q^{(r)}) \succeq m_H P_{q^{(r)}} \delta^{g-1}$, and that the unique minimizer $\delta_* \in T_{q^{(r)}} \delta^{g-1}$ to the trust region subproblem (II.22) satisfies $\|\delta_*\| < \Delta$ (i.e., the constraint is inactive). Then, for $q^{(r+1)} = \exp_{q^{(r)}}(\delta_*)$, we have
\[
\|\nabla f(q^{(r+1)})\| \leq \frac{L_H}{2m_H^2} \|\nabla f(q^{(r)})\|^2,
\]
where $L_H = 5n^{3/2}(2\mu^2) \|X_0\|^2_\infty + 9n/\mu \|X_0\|^2_\infty + 9\sqrt{n} \|X_0\|_\infty$.

**Proof:** See Page 907 under Section VI.

We can now bound the Riemannian gradient of the next iterate as
\[
\|\nabla f(q^{(r+1)})\| \leq \frac{L_H}{2m_H^2} \|\nabla f(q^{(r)})\|^2
\leq \frac{L_H}{2m_H^2} \|U^* \text{Hess} f(q^{(r)}) U\|^2
\leq \frac{L_H}{2m_H^2} \|\text{Hess} f(q^{(r)})\|^2 \Delta^2 = \frac{L_H M^2_H}{2m_H^2} \Delta^2.
\]

Obviously, one can make the upper bound small by tuning down $\Delta$. Combining the above lower bound for $\|\nabla f(q)\|$ for $q \in R_\Delta$, one can conclude that when $\Delta$ is small, the next iterate $q^{(r+1)}$ stays in $R_\Delta$. Another application of the optimality condition (II.7) gives conditions on $\Delta$ that guarantees the next trust-region step is also unconstrained. Detailed argument can be found in proof of the following proposition.

**Proposition 3:** Assume (II.31). W.h.p, once the trust-region algorithm takes an unconstrained $R_1$ step (i.e., $\|\delta\| < \Delta$), it always takes unconstrained $R_1$ steps, provided that
\[
\Delta \leq \frac{cc_2^2 \theta^2 \mu}{n^{7/2} \log^{7/2}(np)}.
\]

Here $c_\star$ is as in Theorem 1 and Theorem 2, and $c > 0$ is another constant.

**Proof:** We only consider the symmetric section in the vicinity of $e_n$ and the claims carry on to others by symmetry. Suppose that step $k$ is an unconstrained $R_1$ step. Then
\[
\|w(q^{(r+1)}) - w(q^{(r)})\| \leq \|q^{(r+1)} - q^{(r)}\| = \|\exp_{q^{(r)}}(\delta) - q^{(r)}\|
= \sqrt{2 - 2 \cos \|\delta\| / 2} \leq \|\delta\| < \Delta.
\]
Thus, if $\Delta \leq \frac{1}{20\sqrt{3}} - \frac{\mu}{4\sqrt{2}}$, $q^{(r+1)}$ will be in $R_\Delta \cup R_\Delta$. Next, we show that if $\Delta$ is sufficiently small, $q^{(r+1)}$ will be indeed in $R_\Delta$. By Lemma 12,
\[
\|\nabla f(q^{(r+1)})\| \leq \frac{L_H}{2m_H^2} \|\nabla f(q^{(r)})\|^2
\leq \frac{L_H M^2_H}{2m_H^2} \|\text{Hess} f(q^{(r)})\|^2 \Delta^2,
\]
where we have used the fact that
\[
\|\delta^{(r)}\| = \|U^* \text{Hess} f(q^{(r)}) U\|^{-1} U^* \nabla f(q^{(r)}) < \Delta,
\]
and as the step is unconstrained. Therefore, on the other hand, by Theorem 1 and Lemma 11, w.h.p.
\[
\|\nabla f(q)\| \geq \beta_{\text{grad}} = \frac{9}{10} c_\theta, \quad \forall q \in R_\Delta.
\]
Hence, provided
\[
\Delta < \frac{m_H}{M_H} \sqrt{\frac{2\beta_{\text{grad}}}{L_H}},
\]  
we have \(q^{(r+1)} \in R_\zeta\).

We next show that when \(\Delta\) is small enough, the next step is also unconstrained. Straightforward calculations give
\[
\|U^* \text{Hess} f(q^{(r+1)}) U\|^{-1} \|U^* \text{grad} f(q^{(r+1)})\| \leq \frac{L_H M_H^2}{2m_H^2} \Delta^2.
\]
Hence, provided that
\[
\Delta < \frac{2m_H^3}{L_H M_H^2},
\]  
we will have
\[
\left\| U^* \text{Hess} f(q^{(r+1)}) U \right\|^{-1} \|U^* \text{grad} f(q^{(r+1)})\| < \Delta;
\]
in words, the minimizer to the trust-region subproblem for the next step lies in the relative interior of the trust region - the constraint is inactive. By Lemma 12 and Lemma 31, we have
\[
L_H = C_H r^{3/2} \log^{3/2}(np) / \mu^2,
\]  
with high probability. Combining this with our previous estimates of \(m_H, M_H\), we conclude whenever
\[
\Delta \leq \min \left\{ \frac{1}{20\sqrt{\pi}}, \frac{\mu}{4\sqrt{2}}, \frac{c_1 c_2 r^{3/2}}{n^{7/4} \log^{1/4}(np)}, \frac{c_2 \mu c_3 r^3}{n^{7/2} \log^{1/2}(np)} \right\},
\]
w.h.p., our next trust-region step is also an unconstrained \(R_\zeta\) step. Simplifying the above bound completes the proof.

Finally, we want to show that ultimate unconstrained \(R_\zeta\) iterates actually converge to one nearby local minimizer rapidly. Lemma 12 has established the gradient is diminishing. The next lemma shows the magnitude of gradient serves as a good proxy for distance to the local minimizer.

**Lemma 13:** Let \(q_\ast \in \mathbb{R}^{n-1}\) such that \(\text{grad} f(q_\ast) = 0\), and \(\delta \in S_q \mathbb{R}^{n-1}\). Consider a geodesic \(\gamma(t) = \text{exp}_{q_\ast}(t\delta)\), and suppose that on \([0, \varepsilon]\), \(\text{Hess} f(\gamma(t)) \geq m_H \mathcal{P}_{T_{\gamma(t)}}\mathbb{R}^{n-1}\). Then
\[
\|\text{grad} f(\gamma(\varepsilon))\| \geq m_H \varepsilon \|\delta\|.
\]  
**Proof:** See Page 907 under Section VI.

To see this relates the magnitude of gradient to the distance away from the nearby local minimizer, w.l.o.g., one can assume \(\varepsilon = 1\) and consider the point \(q = \exp_{q_\ast}(\delta)\). Then
\[
\|q_\ast - q\| = \|\exp_{q_\ast}(\delta) - q\| = \sqrt{2 - 2\cos \|\delta\|} = 2 \sin(\|\delta\| / 2) \leq \|\delta\| \leq \|\text{grad} f(q)\| / m_H,
\]
where at the last inequality above we have used Lemma 13. Hence, combining this observation with Lemma 12, we can derive the asymptotic sequence convergence rate as follows.

**Proposition 4:** Assume (II.31) and the conditions in Lemma 3. Let \(q^{(r_0)} \in R_\zeta\) and the \(r_0\)-th step the first unconstrained \(R_\zeta\) step and \(q_\ast\), be the unique local minimizer of \(f\) over one connected component of \(R_\zeta\) that contains \(q^{(r_0)}\). Then w.h.p., for any positive integer \(r' \geq 1\),
\[
\|q^{(r_0+r')} - q_\ast\| \leq \frac{c c_2 \mu}{n^{3/2} \log^{3/2}(np)} 2^{-2r'},
\]  
provided that
\[
\Delta \leq \frac{c c_2 \mu}{n^{3/2} \log^{3/2}(np)}.
\]  
Here \(c_\ast\) is as in Theorem 1 and Theorem 2, and \(c, c'\) are both positive constants.

**Proof:** By the geometric characterization in [3, Th. 1] and [3, Corollary II.2], \(f\) has \(2n\) separated local minimizers, each located in \(R_\zeta\) and within distance \(\sqrt{\mu}/16\) of one of the \(2n\) signed basis vectors \([\pm e_i]_{i \in [n]}\). Moreover, it is obvious when \(n \leq 1\), \(R_\zeta\) consists of \(2n\) disjoint connected components.

We only consider the symmetric component in the vicinity of \(e_i\) and the claims carry on to others by symmetry.

Suppose that \(r_0\) is the index of the first unconstrained iterate in region \(R_\zeta\), i.e., \(q^{(r_0)} \in R_\zeta\). By Lemma 12, for any integer \(r' \geq 1\), we have
\[
\|\text{grad} f(q^{(r_0+r')})\| \leq \frac{2m_H^2}{L_H} \left(\frac{L_H}{2m_H^2} \|\text{grad} f(q^{(r_0)})\|\right)^{2r'}.
\]  
where \(L_H\) is as defined in Lemma 12, \(m_H\) as the strong convexity parameter for \(R_\zeta\) defined above.

Now suppose \(q_\ast\) is the unique local minimizer of \(f\), lies in the same \(R_\zeta\) component that \(q^{(r_0)}\) is located. Let \(\gamma_r(t) = \text{exp}_{q_\ast}(t\delta)\) to be the unique geodesic that connects \(q_\ast\) and \(q^{(r_0+r')}\) with \(\gamma_r(0) = q_\ast\) and \(\gamma_r(1) = q^{(r_0+r')}\). We have
\[
\|q^{(r_0+r')} - q_\ast\| \leq \|\text{exp}_{q_\ast}(\delta) - q_\ast\| = \sqrt{2 - 2\cos \|\delta\|} = 2 \sin(\|\delta\| / 2)
\]
\[
\|\delta\| \leq \frac{1}{m_H} \|\text{grad} f(q^{(r_0+r')})\| \leq \frac{2m_H}{L_H} \left(\frac{L_H}{2m_H} \|\text{grad} f(q^{(r_0)})\|\right)^{2r'},
\]
where at the second line we have repeatedly applied Lemma 13.

By the optimality condition (II.7) and the fact that \(\|\delta^{(r_0)}\| < \Delta\), we have
\[
\frac{L_H}{2m_H} \|\text{grad} f(q^{(r_0)})\| \leq \frac{L_H}{2m_H} \|\text{grad} f(q^{(r_0)})\| \leq \frac{2m_H}{L_H} \Delta.
\]  
Thus, provided
\[
\Delta < \frac{m_H^2}{L_H M_H},
\]  
we can combine the above results and obtain
\[
\|q^{(r_0+r')} - q_\ast\| \leq \frac{2m_H}{L_H} 2^{-2r'}.
\]
Based on the previous estimates for $m_H$, $M_H$ and $L_H$, we obtain that w.h.p.,
\[
\|q^{(n+q')} - q_*\| \leq \frac{c_1 c_\theta \mu}{n^{3/2} \log^{3/2}(np)} 2^{-2'}.
\]
Moreover, by (II.46), w.h.p., it is sufficient to have the trust region size
\[
\Delta \leq \frac{c_2 c^2 \theta^2 \mu}{n^{5/2} \log^{7/2}(np)}.
\]
Thus, we complete the proof.

Now we are ready to piece together the above technical proposition to prove Theorem 3.

Proof: (of Theorem 3) Assuming (II.31) and in addition that
\[
\Delta < \frac{c_1 c^3 \theta^3 \mu^2}{n^{7/2} \log^{11/2}(np)},
\]
it can be verified that the conditions of all the above propositions are satisfied.

By the preceding four propositions, a step will either be $R_{III}$, $R_{II}$, or constrained $R_I$ step that decreases the objective value by at least a certain fixed amount (we call this Type A), or be an unconstrained $R_I$ step (Type B), such that all future steps are unconstrained $R_I$ and the sequence converges to a local minimizer quadratically. Hence, regardless of the initialization, the whole iteration sequence consists of consecutive Type A steps, followed by consecutive Type B steps. Depending on the initialization, either the Type A phase or the Type B phase can be absent. In any case, from $q^{(0)}$ it takes at most (note $f(q) \geq 0$ always holds)
\[
f(q^{(0)}) \leq \frac{c_1 c \theta \mu}{\epsilon n^{3/2} \log^{3/2}(np)}
\]
steps for the iterate sequence to start take consecutive unconstrained $R_I$ steps, or to already terminate. In case the iterate sequence continues to take consecutive unconstrained $R_I$ steps, Proposition 4 implies that it takes at most
\[
\log \left( \frac{c_2 c \theta \mu}{\epsilon n^{3/2} \log^{3/2}(np)} \right)
\]
steps to obtain an $\epsilon$-near solution to the $q_*$ that is contained in the connected subset of $R_I$ that the sequence entered.

Thus, the number of iterations to obtain an $\epsilon$-near solution to $q_*$ can be grossly bounded by
\[
\#\text{Iter} \leq \frac{f(q^{(0)})}{\min\{d_1, d_{III}, d_{III}\}} + \log \log \left( \frac{c_2 c \theta \mu}{\epsilon n^{3/2} \log^{3/2}(np)} \right)
\]
\[
\leq \left[ \min \left\{ \frac{c_3 c^3 \theta^4 \mu^4}{n^6 \log^{10}(np)} , \frac{c_4 c^2 \theta^2}{n^2} \right] - \Delta^2 \right]^{-1} f(q^{(0)})
\]
\[
+ \log \log \left( \frac{c_2 c \theta \mu}{\epsilon n^{3/2} \log^{3/2}(np)} \right).
\]
Finally, the claimed failure probability comes from a simple union bound with careful bookkeeping.

F. Extending to Convergence for Complete Dictionaries

Recall that in this case we consider the preconditioned input
\[
\bar{Y} = \sqrt{\rho} (Y Y^*)^{-1/2} Y.
\]
Note that for any complete $A_0$ with condition number $\kappa(A_0)$, from Lemma 32 we know when $p$ is large enough, w.h.p. one can write the preconditioned $\bar{Y}$ as
\[
\bar{Y} = U V^* X_0 + \Xi X_0
\]
for a certain $\Xi$ with small magnitude, and $U \Sigma V^* = \text{SVD}(A_0)$. Particularly, when $p$ is chosen by Theorem 2, the perturbation is bounded as
\[
\|\Xi\| \leq c c_\theta \theta \left( \max \left\{ \frac{n^{3/2}}{\mu^2}, \frac{n^2}{\mu} \right\} \log^{3/2}(np) \right)^{-1}
\]
for a certain constant $c$ which can be made arbitrarily small by making the constant $C$ in $p$ large. Since $UV^*$ is orthogonal,
\[
f(q; U V^* X_0 + \Xi X_0) = f(V U^* q; X_0 + V U^* \Xi X_0).
\]
In words, the function landscape of $f(q; U V^* X_0 + \Xi X_0)$ is a rotated version of that of $f(q; X_0 + V U^* \Xi X_0)$. Thus, any local minimizer $q_*$ of $f(q; X_0 + V U^* \Xi X_0)$ is rotated to $UV^* q_*$, one minimizer of $f(q; UV^* X_0 + \Xi X_0)$. Also if our algorithm generates iteration sequence $q_0, q_1, q_2, \ldots$ for $f(q; X_0 + V U^* \Xi X_0)$ upon initialization $q_0$, it will generate the iteration sequence $UV^* q_0, UV^* q_1, UV^* q_2, \ldots$ for $f(q; U V^* X_0 + \Xi X_0)$. So w.l.o.g. it is adequate that we prove the convergence results for the case $f(q; X_0 + V U^* \Xi X_0)$, corresponding to $A_0 = I$ with perturbation $\Xi = V U^* \Xi$. So in this section (Section II-F), we write $f(q; X_0)$ to mean $f(q; X_0 + \Xi X_0)$.

Theorem 2 has shown that when
\[
\theta \in \left( 0, \frac{1}{2} \right), \mu \leq c \min \left\{ \frac{\theta}{n}, \frac{1}{n^{5/4}} \right\},
\]
\[
p \geq \frac{C}{c^2 \theta^2} \max \left\{ \frac{n^4}{4 \mu^2} , \frac{n^5}{\mu^2} \right\} \kappa^2 (A_0) \log^{4} \left( \frac{\kappa (A_0) n}{\mu \theta} \right),
\]
the geometric structure of the landscape is qualitatively unchanged from the orthogonal case, and the parameter $c_\theta$ constant can be replaced with $c_{\theta}/2$. Particularly, for this choice of $p$, Lemma 32 implies
\[
\|\Xi\| = \|V U^* \Xi\| \leq c c_\theta \theta \left( \max \left\{ \frac{n^{3/2}}{\mu^2}, \frac{n^2}{\mu} \right\} \log^{3/2}(np) \right)^{-1}
\]
for a constant $c$ that can be made arbitrarily small by setting the constant $C$ in $p$ sufficiently large. The whole proof is quite similar to that of orthogonal case in the last section. We will only sketch the major changes below. To distinguish with the corresponding quantities in the last section, we use $\tilde{q}$ to denote the corresponding perturbed quantities here.

- **Lemma 5**: Note that
\[
\|X_0 + \Xi X_0\|_{\infty} \leq \|X_0\|_{\infty} + \|\Xi X_0\|_{\infty}
\]
\[
\leq \|X_0\|_{\infty} + \sqrt{n} \|\Xi\| \|X_0\|_{\infty}
\]
\[
\leq 3 \|X_0\|_{\infty}/2,
\]
where by (II.52) we have used \( \| \widetilde{\mathbf{E}} \| \leq 1/(2\sqrt{n}) \) to simplify the above result. So we obtain

\[
\widetilde{M}_V \leq \frac{3}{2} M_V, \quad \widetilde{M}_{V^2} \leq \frac{9}{4} M_{V^2}, \\
\widetilde{L}_V \leq \frac{9}{4} L_V, \quad \widetilde{L}_{V^2} \leq \frac{27}{8} L_{V^2}.
\]

- Lemma 6: Now we have

\( \bar{\eta}_f = \widetilde{M}_V + 2\widetilde{M}_{V^2} + \widetilde{L}_V + \widetilde{L}_{V^2} \leq 4\eta_f. \)

- Lemma 7 and Lemma 8 are generic and nothing changes.

- Proposition 1: We have now \( w^* g(w; \widetilde{X}_0)/\|w\| \geq c_4 \theta/2 \) by Theorem 2, w.h.p. \( w^* \nabla g(w; \widetilde{X}_0)/\|w\| \) is \( C_1 n^2 \log(np)/\mu \)-Lipschitz by Proposition 6, and \( \| X_0 + \widetilde{\mathbf{E}} X_0 \| \leq \frac{1}{3} \| X_0 \| / \sqrt{2} \) as shown above. Similarly, \( w^* g(w; \widetilde{X}_0)/\|w\| \leq -c_4 \theta/2 \) by Theorem 2, and \( w^* \nabla^2 g(w; \widetilde{X}_0)w/w^2 \) is \( C_2 n^3 \log^{3/2}(np)/\mu^2 \)-Lipschitz. Moreover, \( \bar{\eta}_f \leq 4\eta_f \) as shown above. Since there are only multiplicative constant changes to the various quantities, we conclude

\[
\bar{d}_{\Pi} = c_1 d_{\Pi}, \quad \bar{d}_{\Pi \Pi} = c_1 d_{\Pi \Pi} \quad (II.53)
\]

provided

\[
\Delta < \frac{c_2 c_4 \theta \mu^2}{n^{3/2} \log^{3/2}(np)}. \quad (II.54)
\]

- Lemma 9: \( \eta_f \) is changed to \( \bar{\eta}_f \) with \( \bar{\eta}_f \leq 4\eta_f \) as shown above.

- Lemma 10: By (II.3), we have

\[
\| \nabla^2 f(q; X_0) - \nabla^2 f(q; \widetilde{X}_0) \|
\leq \frac{1}{p} \sum_{k=1}^{p} \left[ L_{h^*_k} \| \widetilde{\mathbf{E}} \| \| (x_0)_k \| \right]
\]

\[
+ \frac{1}{\mu} \left( \| (x_0)_k \| - \| \widetilde{\mathbf{E}} (x_0)_k \| \right)
\]

\[
\leq \| \widetilde{\mathbf{E}} \| (L_{h^*_k} + 2/\mu + \| \widetilde{\mathbf{E}} \| / \mu) \sum_{k=1}^{p} \| (x_0)_k \| ^2,
\]

where \( L_{h^*_k} \) is the Lipschitz constant for the function \( \tilde{h}_k \) and we have used the fact that \( \| \widetilde{\mathbf{E}} \| \leq 1 \). Similarly, by II.2,

\[
\| \nabla f(q; X_0) - \nabla f(q; \widetilde{X}_0) \|
\leq \frac{1}{p} \sum_{k=1}^{p} \left[ L_{h^*_k} \| \widetilde{\mathbf{E}} \| \| (x_0)_k \| + \| \widetilde{\mathbf{E}} \| \| (x_0)_k \| \right]
\]

\[
\leq \left( L_{h^*_k} + 1 \right) \| \widetilde{\mathbf{E}} \| \sqrt{\| X_0 \|},
\]

where \( L_{h^*_k} \) is the Lipschitz constant for the function \( \tilde{h}_k \). Since \( L_{h^*_k} \leq 2/\mu^2 \) and \( L_{h^*_k} \leq 1/\mu \), and \( \| X_0 \| \leq 4\sqrt{\log(np)} \) w.h.p. (Lemma 31). By (II.52), w.h.p. we have

\[
\| \nabla f(q; X_0) - \nabla f(q; \widetilde{X}_0) \| \leq \frac{1}{2} c_4 \theta, \quad \text{and}
\]

\[
\| \nabla^2 f(q; X_0) - \nabla^2 f(q; \widetilde{X}_0) \| \leq \frac{1}{2} c_4 \theta,
\]

provided the constant \( C \) in (II.51) for \( p \) is large enough. Thus, by (II.6) and the above estimates we have

\[
\| \nabla f(q; X_0) - \nabla f(q; \widetilde{X}_0) \|
\leq \| \nabla f(q; X_0) - \nabla f(q; \widetilde{X}_0) \|
\]

\[
+ \| \nabla^2 f(q; X_0) - \nabla^2 f(q; \widetilde{X}_0) \|
\]

\[
\leq c_4 \theta \leq \frac{1}{2} c_4 \theta,
\]

provided \( \mu \leq 1/2 \). So we conclude

\[
\text{Hess}(f(q; \widetilde{X}_0)) \leq \frac{1}{2} c_4 \theta \mathcal{P}_q \mathcal{S}_n \frac{1}{\mu} \implies \widetilde{m}_H \geq \frac{1}{2} c_4 \theta \mu. \quad (II.55)
\]

- Proposition 2: From the estimate of \( M_H \) above Proposition 2 and the last point, we have

\[
\| \text{Hess}(f(q; \widetilde{X}_0)) \| \leq \frac{36 \log(np)}{\mu}, \quad \text{and}
\]

\[
\text{Hess}(f(q; \widetilde{X}_0)) \leq \frac{1}{2} c_4 \theta \mathcal{P}_q \mathcal{S}_n \frac{1}{\mu}.
\]

Also since \( \bar{\eta}_f \leq 4\eta_f \) in Lemma 6 and Lemma 9, there are only multiplicative constant change to the various quantities. We conclude that

\[
\bar{d}_{\Pi} = c_4 d_{\Pi} \quad (II.56)
\]

provided that

\[
\Delta \leq \frac{c_4 c^2 \theta^2 \mu}{n^{3/2} \log^{3/2}(np)}. \quad (II.57)
\]

- Lemma 11 is generic and nothing changes.

- Lemma 12: \( L_{H} \leq 27 L_H / 8 \).

- Proposition 3: All the quantities involved in determining \( \Delta, m_H, M_H, \) and \( L_H, \beta_{\text{grad}} \) are modified by at most constant multiplicative factors and changed to their respective tilde version, so we conclude that the TRM algorithm always takes unconstrained \( R_x \) step after taking one, provided that

\[
\Delta \leq \frac{c_4 c^3 \theta^3 \mu}{n^{3/2} \log^{7/2}(np)}. \quad (II.58)
\]

- Lemma 13: is generic and nothing changes.

- Proposition 4: Again \( m_H, M_H, L_H \) are changed to \( \widetilde{m}_H, \widetilde{M}_H, \) and \( \tilde{L}_H \) respectively, differing by at most constant multiplicative factors. So we conclude for any integer \( k' \geq 1 \),

\[
\left\| q^{(k_0+k')} - q_* \right\| \leq \frac{c_6 c_4 \theta \mu}{n^{3/2} \log^{3/2}(np)} 2^{-2k'}, \quad (II.59)
\]

provided

\[
\Delta \leq \frac{c_6 c_4 \theta^2 \mu}{n^{3/2} \log^{3/2}(np)}. \quad (II.60)
\]

The final proof to Theorem 2 is almost identical to that of Theorem 1, except that \( d_{\Pi}, d_{\Pi \Pi}, \) and \( d_{\Pi \Pi} \) are changed.
to $\tilde{d}_2, \tilde{d}_1$, and $\tilde{d}_\infty$ as defined above, respectively. The final iteration complexity to each an $\varepsilon$-near solution is hence
\begin{equation*}
\# \text{Iter} \leq \left[ \min \left\{ \frac{c_8\varepsilon^3 q^4}{n^6 \log^3 (np)}, \frac{c_9\varepsilon^2 q^2}{n} \right\} \Delta^2 \right]^{-1} f (q^{(0)}) + \log \left( \frac{c_{10\mu} \varepsilon \mu}{c_{11} \mu} \right): \end{equation*}

Hence overall the qualitative behavior of the algorithm is not changed, as compared to that for the orthogonal case.

### III. COMPLETE ALGORITHM PIPELINE AND MAIN RESULTS

For orthogonal dictionaries, from Theorem 1 (and [3, Corollary II.2]), we know that all the minimizers $q_\star$ are $O(\mu)$ away from their respective nearest “target” $q_*$, with $q_\star^\top \tilde{Y} = a \epsilon q_\star X_0$ for a certain $a \neq 0$ and $i \in [n]$; in Theorem 3, we have shown that w.h.p. the Riemannian TRM algorithm produces a solution $\hat{q} \in S^{n-1}$ that is $\varepsilon$ away to one of the minimizers, say $q_\star$. Thus, the $\hat{q}$ returned by the TRM algorithm is $O(\varepsilon + \mu)$ away from $q_*$. For exact recovery, we use a simple linear programming rounding procedure, which guarantees to produce the target $q_*$. We then use deflation to sequentially recover other rows of $X_0$. Overall, w.h.p. both the dictionary $A_0$ and sparse coefficient $X_0$ are exactly recovered up to sign permutation, when $\theta \in \Omega(1)$, for orthogonal dictionaries. We summarize relevant technical lemmas and main results in Section III-A. The same procedure can be used to recover complete dictionaries, though the analysis is slightly more complicated; we present the results in Section III-B. Our overall algorithmic pipeline for recovering orthogonal dictionaries is sketched as follows.

1) **Estimating one row of $X_0$ by the Riemannian TRM algorithm.** By Theorem 1 (resp. Theorem 2) and Theorem 3 (resp. Theorem 4), starting from any $q \in S^{n-1}$, when the relevant parameters are set appropriately (say as $\mu_\star$ and $\Delta_\star$), w.h.p., our Riemannian TRM algorithm finds a local minimizer $\hat{q}$, with $q_\star$ the nearest target that exactly recovers a row of $X_0$ and $\|\hat{q} - q_\star\| \in O(\mu)$ (by setting the target accuracy of the TRM as, say, $\varepsilon = \mu$).

2) **Recovering one row of $X_0$ by rounding.** To obtain the target solution $q_\star$, and hence recover (up to scale) one row of $X_0$, we solve the following linear program:
\begin{equation*}
\min \|q^\top \hat{Y}\|_1, \quad \text{s.t.} \quad \langle r, q \rangle = 1, \quad \text{(III.1)}
\end{equation*}
with $r = \hat{q}$. We show in Lemma 15 (resp. Lemma 17) that when $\|\hat{q}, q_\star\|$ is sufficiently large, implied by $\mu$ being sufficiently small, w.h.p. the minimizer of (III.1) is exactly $q_\star$, and hence one row of $X_0$ is recovered by $q_\star^\top \hat{Y}$.

3) **Recovering all rows of $X_0$ by deflation.** Once $\ell$ rows of $X_0$ ($1 \leq \ell \leq n-2$) have been recovered, say, by unit vectors $q^1_\ell, \ldots, q^\ell_\ell$, one takes an orthonormal basis $U$ for $[q^1_\ell, \ldots, q^\ell_\ell]^\perp$, and minimizes the new function $h(z) = f(Uz; \hat{Y})$ on the sphere $S^{n-\ell-1}$ with the Riemannian TRM algorithm (though conservative, one can again set parameters as $\mu_\star$, $\Delta_\star$, as in Step 1) to produce a $\hat{z}$. Another row of $X_0$ is then recovered via the LP rounding (III.1) with input $r = U\hat{z}$ (to produce $q^{\ell+1}_\ell$). Finally, by repeating the procedure until depletion, one can recover all the rows of $X_0$.

4) **Reconstructing the dictionary $A_0$.** By solving the linear system $Y = AX_0$, one can obtain the dictionary $A_0 = YX_0^\perp (X_0X_0^\perp)^{-1}$.

### A. Recovering Orthogonal Dictionaries

Theorem 14 (Main Theorem - Recovering Orthogonal Dictionaries): Assume the dictionary $A_0$ is orthogonal and we take $\hat{Y} = Y$. Suppose $\theta \in (0, 1/3)$, $\mu_\star < c_i \min \{\theta n^{-1}, n^{-5/4}\}$, and $p \geq Cn^3 \log n \mu_\star^2/\mu_\star^2$. The above algorithmic pipeline with parameter setting
\begin{equation*}
\Delta_\star = \frac{c_{6\mu}^2 \mu_\star^2}{n^{7/2} \log^{1/2} (np)}, \quad \text{(III.2)}
\end{equation*}
recovers the dictionary $A_0$ and $X_0$ in polynomial time, with failure probability bounded by $c_i e^{-p}$. Here $c_i$ as defined in Theorem 1, and $c_i$ through $c_i$, and $C$ are all positive constants.

Towards a proof of the above theorem, it remains to be shown the correctness of the rounding and deflation procedures.

a) **Proof of LP rounding.** The following lemma shows w.h.p. the rounding will return the desired $q_\star$, provided the estimated $\hat{q}$ is already near to it.

**Lemma 15 (LP Rounding - Orthogonal Dictionary):** For any $\theta \in (0, 1/3)$, whenever $p \geq Cn^2 \log (n/\theta)/\theta$, with probability at least $1 - e^{-p}$, the rounding procedure (III.1) returns $q_\star$ for any input vector $r$ that satisfies
\begin{equation*}
|r, q_\star| \geq 249/250.
\end{equation*}

Here $C, c$ are both positive constants.

**Proof:** See Page 908 under Section VII.

Since $\|\hat{q}, q_\star\| = 1 - \|\hat{q} - q_\star\|^2/2$, and $\|\hat{q} - q_\star\| \in O(\mu)$, it is sufficient when $\mu$ is smaller than some small constant.

b) **Proof sketch of deflation.** We show the deflation works by induction. To understand the deflation procedure, it is important to keep in mind that the “target” solutions $\{q^i_\ell\}_{i=1}^\ell$ are orthogonal to each other. W.l.o.g., suppose we have found the first $\ell$ unit vectors $q^1_\ell, \ldots, q^\ell_\ell$ which recover the first $\ell$ rows of $X_0$. Correspondingly, we partition the target dictionary $A_0$ and $X_0$ as
\begin{equation*}
A_0 = [V, V^\perp], \quad X_0 = \left[ \begin{array}{c} X_0^{[\ell]} \\ x_0^{[n-\ell]} \end{array} \right], \quad \text{(III.3)}
\end{equation*}
where $V \in \mathbb{R}^{n\times \ell}$, and $X^{[\ell]}_0 \in \mathbb{R}^{\ell \times n}$ denotes the submatrix with the first $\ell$ rows of $X_0$. Let us define a function: $f^\perp_\ell : \mathbb{R}^{n-\ell} \mapsto \mathbb{R}$ by
\begin{equation*}
f^\perp_\ell(z; W) = \frac{1}{p} \sum_{k=1}^p h_{\rho}(z^\top w_k), \quad \text{(III.4)}
\end{equation*}
for any matrix $W \in \mathbb{R}^{(n-\ell) \times p}$. Then by (I.1), our objective function is equivalent to
\begin{equation*}
h(z) = f(Uz; A_0X_0) = f^\perp_\ell(z; U^* A_0 X_0)
\end{equation*}
\begin{equation*}
= f^\perp_\ell(z; U^* XV_0^\perp + U^* V X_0^{[n-\ell]}).
\end{equation*}
Since the columns of the orthogonal matrix $U \in \mathbb{R}^{n \times (n - \ell)}$ forms the orthogonal complement of span $\{q_1^*, \ldots, q_\ell^*\}$, it is obvious that $U^*V = 0$. Therefore, we obtain

$$h(z) = f_{n-\ell}^z(z; U^*V^\perp X_0^{[n-\ell]}).$$

Since $U^*V^\perp$ is orthogonal and $X_0^{[n-\ell]} \sim_{i.i.d.} \text{BG}(\theta)$, this is another instance of orthogonal dictionary learning problem with reduced dimension. If we keep the parameter settings $\mu_\star$ and $\Delta_\star$ as Theorem 14, the conditions of Theorem 1 and Theorem 3 for all cases with reduced dimensions are still valid. So w.h.p., the TRM algorithm returns a $\hat{z}$ such that $\|\hat{z} - z_\star\| \in O(\mu_\star)$ where $z_\star$ is a "target" solution that recovers a row of $X_0$:

$$z_\star^* U^* V^\perp X_0^{[n-\ell]} = z_\star^* U^* A_0 X_0 = a e_\star^* X_0,$$

for some $i \not\in \ell$. So pulling everything back in the original space, the effective target is $q_i^{\perp, \perp} = U z_\star$, and $U \hat{z}$ is our estimation obtained from the TRM algorithm. Moreover,

$$\|U \hat{z} - U z_\star\| = \|\hat{z} - z_\star\| \in O(\mu_\star).$$

Thus, by Lemma 15, one successfully recovers $U z_\star$ from $U \hat{z}$ w.h.p. when $\mu_\star$ is smaller than a constant. The overall failure probability can be obtained via a simple union bound and simplification of the exponential tails with inverse polynomials in $p$.

**B. Recovering Complete Dictionaries**

By working with the preconditioned data samples $\hat{Y} = Y \sim (YY^\ast)^{-1/2} Y$, we can use the same procedure as described above to recover complete dictionaries.

**Theorem 16 (Main Theorem - Recovering Complete Dictionaries):** Assume the dictionary $A_0$ is complete with a condition number $\kappa(A_0)$ and we take $\hat{Y} = Y$. Suppose $\theta \in (0, 1/3)$, $\mu_\star < c_d \min \{\theta n^{-1}, n^{-5/4}\}$, and $p \geq \frac{\mu_\star}{\epsilon c_d^2} \max \left\{\frac{\mu_\star}{\theta n}, \frac{\mu_\star}{n^5}\right\} \kappa^8(A_0) \log^4 \left(\frac{\kappa(A_0) n}{\mu_\theta}\right)$. The algorithmic pipeline with parameter setting

$$\Delta_\star = \frac{c_d c^3 \mu_\star^2}{n^{7/2} \log^{7/2}(np)}$$

recovers the dictionary $A_0$ and $X_0$ in polynomial time, with failure probability bounded by $c_p p^{-6}$. Here $c_d$ is as defined in Theorem 1, and $c_d, c_p$ are both positive constants.

Similar to the orthogonal case, we need to show the correctness of the rounding and deflation procedures so that the theorem above holds.

c) Proof of LP rounding: The result of the LP rounding is only slightly different from that of the orthogonal case in Lemma 15, so is the proof.

**Lemma 17 (LP Rounding - Complete Dictionary):** For any $\theta \in (0, 1/3)$, whenever

$$p \geq \frac{C}{c_d^2 \theta} \max \left\{\frac{n}{\mu^4}, \frac{n^5}{\mu^2}\right\} \kappa^8(A_0) \log^4 \left(\frac{\kappa(A_0) n}{\mu_\theta}\right),$$

with probability at least $1 - cp^{-6}$, the rounding procedure (III.1) returns $q_i^*$ for any input vector $r$ that satisfies

$$\|r, q_i^*\| \geq 249/250.$$

Here $C, c$ are both positive constants.

**Proof:** See Page 909 under Section VII.

d) Proof sketch of deflation: We use a similar induction argument to show the deflation works. Compared to the orthogonal case, the tricky part here is that the target vectors $\{q_i^*\}_{i=1}^\ell$ are not necessarily orthogonal to each other, but they are almost so. W.l.o.g., let us again assume that $q_1^*, \ldots, q_\ell^*$ recover the first $\ell$ rows of $X_0$, and similarly partition the matrix $X_0$ as in (III.3).

By Lemma 32 and (II.50), we can write $Y = (Q + \Xi) X_0$ for some orthogonal matrix $Q$ and small perturbation $\Xi$ with $\|\Xi\| \leq \delta < 1/10$ for some large $p$ as usual. Similar to the orthogonal case, we have

$$h(z) = f(Uz; (Q + \Xi) X_0) = f_{n-\ell}^z(z; U^*(Q + \Xi) X_0),$$

where $f_{n-\ell}^z$ is defined the same as in (III.4). Next, we show that the matrix $U^*(Q + \Xi) X_0$ can be decomposed as $U^* V X_0^{[n-\ell]} + \Delta$, where $V \in \mathbb{R}^{(n - \ell) \times \kappa}$ is orthogonal and $\Delta$ is a small perturbation matrix. More specifically, we show that

**Lemma 18:** Suppose the matrices $U \in \mathbb{R}^{n \times (n - \ell)}$, $Q \in \mathbb{R}^{p \times \kappa}$ are orthogonal as defined above, $\Xi$ is a perturbation matrix with $\|\Xi\| \leq 1/20$, then

$$U^* (Q + \Xi) X_0 = U^* V X_0^{[n-\ell]} + \Delta,$$

where $V \in \mathbb{R}^{(n - \ell) \times \kappa}$ is an orthogonal matrix spanning the same subspace as that of $U$, and the norms of $\Delta$ is bounded by

$$\|\Delta\|_{\ell \rightarrow \ell^2} \leq 16 \sqrt{n} \|\Xi\| \|X_0\|_\infty, \quad \|\Delta\| \leq 16 \|\Xi\| \|X_0\|,$$

(III.7)

where $\|W\|_{\ell \rightarrow \ell^2} = \sup_{||\xi||_1 = 1} ||W \xi|| = \max_k ||w_k||$ denotes the max column $\ell^2$-norm of a matrix $W$.

**Proof:** See Page 909 under Section VII.

Since $UV$ is orthogonal and $X_0^{[n-\ell]} \sim_{i.i.d.} \text{BG}(\theta)$, we come into another instance of perturbed dictionary learning problem with a reduced dimension

$$h(z) = f_{n-\ell}^z(z; U^* V X_0^{[n-\ell]} + \Delta).$$

Since our perturbation analysis in proving Theorem 2 and Theorem 4 solely relies on the fact that $\|\Delta\|_{\ell \rightarrow \ell^2} \leq C \|\Xi\| \|X_0\|_\infty$, it is enough to make $p$ large enough so that the theorems are still applicable for the reduced version $f_{n-\ell}^z(z; U^* V X_0^{[n-\ell]} + \Delta)$. Thus, by invoking Theorem 2 and Theorem 4, the TRM algorithm provably returns one $\hat{z}$ such that $\hat{z}$ is near to a perturbed optimal $\Xi$, with

$$\hat{z}_\star^* U^* V X_0^{[n-\ell]} = z_\star^* U^* V X_0^{[n-\ell]} + z_\star^* \Delta = a e_\star^* X_0,$$

for some $i \not\in \ell$, (III.8)

where $z_\star$ with $\|z_\star\| = 1$ is the exact solution. More specifically, [3, Corollary II.4] implies that

$$\|\hat{z} - z_\star\| \leq \sqrt{\mu_\star} / 7.$$
Next, we show that $\hat{z}$ is also very near to the exact solution $z_*$. Indeed, the identity (III.8) suggests

$$\begin{align*}
(\hat{z}_* - z_*)^* W^* X_0^{n-f} &= z_*^* \Delta \\
\implies \hat{z}_* - z_* &= \left( X_0^{n-f} V^* U \right)^\top \Delta^* z_* \\
&= U^* V \left( X_0^{n-f} \right)^\top \Delta^* z_* \quad \text{(III.9)}
\end{align*}$$

where $W^\top = (W^* W)^{-1} W^*$ denotes the pseudo inverse of a matrix $W$ with full column rank. Hence, by (III.9) we can bound the distance between $\hat{z}_*$ and $z_*$ by

$$\| \hat{z}_* - z_* \| \leq \left\| \left( X_0^{n-f} \right)^* \right\| \| \Delta \| \leq \sigma_{\min}^{-1} (X_0^{n-f}) \| \Delta \|$$

By Lemma 29, when $p \geq \Omega(n^2 \log n)$, w.h.p.,

$$\theta p/2 \leq \sigma_{\min} (X_0^{n-f})(X_0^{n-f})^* \leq \left\| X_0^{n-f} (X_0^{n-f})^* \right\| \| X_0 \| \leq 3 \theta p/2.$$

Hence, combined with Lemma 18, we obtain

$$\sigma_{\min}^{-1} (X_0^{n-f}) \leq \frac{2}{\sqrt{\theta p}}, \| \Delta \| \leq 28 \sqrt{\theta p} \| \Xi \| / \sqrt{2},$$

which implies that $\| \hat{z}_* - z_* \| \leq 28 \| \Xi \|$. Thus, combining the results above, we obtain

$$\| \hat{z}_* - z_* \| \leq \| \hat{z}_* - \hat{z}_* \| + \| \hat{z}_* - z_* \| \leq \sqrt{2} \| \mu \| / 7 + 28 \| \Xi \|.$$

Lemma 32, and in particular (II.50), for our choice of $p$ as in Theorem 2, $\| \Xi \| \leq c \mu n^{-3/2}$, where $c$ can be made smaller by making the constant in $p$ larger. For $\mu$ sufficiently small, we conclude that

$$\| U \hat{z}_* - U z_* \| = \| \hat{z}_* - z_* \| \leq 2 \mu / 7.$$

In words, the TRM algorithm returns a $\hat{z}$ such that $U \hat{z}$ is very near to one of the unit vectors $\{ q_i^\top \}_{i=1}^\mu$, such that $q_i^\top = \alpha e_i^\top X_0$ for some $\alpha \neq 0$. For $\mu$ smaller than a fixed constant, one will have

$$\left\langle U \hat{z}_*, q_i^\top \right\rangle \geq 249/250,$$

and hence by Lemma 17, the LP rounding exactly returns the optimal solution $q_i^\top$ upon the input $U \hat{z}_*$.

The proof sketch above explains why the recursive TRM plus rounding works. The overall failure probability can be obtained via a simple union bound and simplifications of the exponential tails with inverse polynomials in $p$.

### IV. SIMULATIONS

#### A. Practical TRM Implementation

Fixing a small step size and solving the trust-region subproblem exactly eases the analysis, but also renders the TRM algorithm impractical. In practice, the trust-region subproblem cannot be solved exactly, and the trust-region step size is adjusted to the local geometry, say by backtracking. It is possible to modify our algorithmic analysis to account for inexact subproblem solvers and adaptive step size; for sake of brevity, we do not pursue it here. Recent theoretical results on the practical version include [72] and [73].

Here we describe a practical implementation based on the Manopt toolbox [70]. Manopt is a user-friendly Matlab toolbox that implements several sophisticated solvers for tackling optimization problems over Riemannian manifolds. The most developed solver is based on the TRM. This solver uses the truncated conjugate gradient (tCG; see, e.g., [57, Sec. 7.5.4]) method to (approximately) solve the trust-region subproblem (vs. the exact solver in our analysis). It also dynamically adjusts the step size using backtracking. However, the original implementation (Manopt 2.0) is not adequate for many purposes. Their tCG solver uses the gradient as the initial search direction, which does not ensure that the TRM solver can escape from saddle points [68], [71]. We modify the tCG solver, such that when the current gradient is small and there is a negative curvature direction (i.e., the current point is near a saddle point or a local maximizer of $f(q)$), the tCG solver explicitly uses the negative curvature direction as the initial search direction. This modification ensures the TRM solver always escape from saddle points/local maximizers with negative directional curvature. Hence, the modified TRM algorithm based on Manopt is expected to have the same qualitative behavior as the idealized version we analyzed above, with better scalability. We will perform numerical simulations using the modified TRM algorithm whenever necessary. Algorithm 3 together with Lemmas 9 and 10 and the surrounding discussion in the very recent work [72] provides a detailed description of this practical version.

#### B. Simulated Data

To corroborate our theory, we experiment with dictionary recovery on simulated data. For simplicity, we focus on recovering orthogonal dictionaries and we declare success once a single row of the coefficient matrix is recovered.

Since the problem is invariant to rotations, w.l.o.g. we set the dictionary as $A_0 = I \in \mathbb{R}^{n \times n}$. For any fixed sparsity $k$, each column of the coefficient matrix $X_0 \in \mathbb{R}^{n \times p}$ has exactly $k$ nonzero entries, chosen uniformly random from $\binom{n}{k}$. These nonzero entries are i.i.d. standard normals. This is slightly different from the Bernoulli-Gaussian model we assumed for analysis. For $n$ reasonably large, these two models have similar behaviors. For our sparsity surrogate, we fix the smoothing parameter as $\mu = 10^{-2}$. Because the target points are the signed basis vector $\pm e_i$’s (to recover rows of $X_0$), for a solution $\hat{q}$ returned by the TRM algorithm, we define the reconstruction error (RE) to be

$$\text{RE} = \min_{i \in [n]} \left( \| \hat{q} - e_i \|, \| \hat{q} + e_i \| \right).$$

One trial is determined to be a success once $\text{RE} \leq \mu$, with the idea that this indicates $\hat{q}$ is already very near the target and the target can likely be recovered via the LP rounding we described (which we do not implement here).

We consider two settings: (1) fix $p = 5n^2 \log n$ and vary the dimension $n$ and sparsity $k$; (2) fix the sparsity level as $[0.2-n]$.

---

5Available online: http://www.manopt.org.

6Adjusted in sign to ensure positive correlation with the gradient – if it does not vanish.

7The code is available online: https://github.com/sunjf/dil_focm
and vary the dimension \( n \) and number of samples \( p \). For each pair of \((k, n)\) for (1), and each pair of \((p, n)\) for (2), we repeat the simulations independently for \( T = 5 \) times. Fig. 3 shows the phase transition for the two settings. It seems that our TRM algorithm can work well into the linear region whenever the phase transition for the two settings. It seems that our pair of \((k, n)\) for five times. White indicates success, and black indicates failure.

\[
\mathbf{Y} = \left( \mathbf{Y} \mathbf{Y}^\top \right)^{-1/2} \mathbf{Y},
\]

and the resulting \( \mathbf{Y} \) is fed to the dictionary learning pipeline described in Section III. The smoothing parameter \( \mu \) is fixed to \( 10^{-2} \). Fig. 4 contains the learned dictionaries: the dictionaries generally contain localized, directional features that resemble subset of wavelets and generalizations. These are very reasonable representing elements for natural images. Thus, the BG coefficient model may be a sensible, simple model for natural images.

Another piece of strong evidence in support of the above claim is as follows. For each image, we repeat the learning pipeline for one hundred times, with independent initializations across the runs. Let \( \mathbf{A} \) be the final learned dictionary for each run, we plot the value of \( \| \mathbf{A}^{-1} \mathbf{Y} \|_1 \) across the one hundred independent runs. Strikingly, the values are virtually the same, with a relative difference of \( 10^{-3} \)! This is predicted by our theory, under the BG model. If the model is unreasonable for natural images, the preconditioning, benign function landscape, LP rounding, and the deflation process that hinge on this model would have completely fallen down.

For this image experiment, \( n = 64 \) and \( p = 4096 \). A single run of the learning pipeline, including solving 64 instances of the optimization over the sphere (with varying dimensions) and solving 64 instances of the LP rounding (using CVX), lasts about 20 minutes on a mid-range modern laptop. So with careful implementation we discussed above, the learning pipeline is actually not far from practical.

V. DISCUSSION

For recovery of complete dictionaries, the LP program approach in [81] that works with \( \theta \leq O(1/\sqrt{n}) \) only demands \( p \geq \Omega(n^2 \log n) \). The sample complexity has recently been improved to \( p \geq \Omega(n \log n) \) [82], matching the information-theoretic lower bound \( \Omega(n \log n) \) (i.e., when \( \theta \sim 1/n \); see also [83]), albeit still in the \( \theta \in O(1/\sqrt{n}) \) regime. The sample complexity of our method working in the \( \theta \sim O(1) \) regime as stated in Theorem 16 is obviously much higher. As already discussed in [3], working with other \( \ell^1 \) proxies, directly in the \( q \) space would likely save the sample complexity both for geometric characterization and algorithm analysis. Another possibility is to analyze the complete case directly, instead of treating it as transformed version of perturbed orthogonal case.

Our experiments seem to suggest the necessary complexity level lies between \( \Omega(n) \) and \( O(n^2 \log n) \) even for the orthogonal case. While it is interesting to determine the true complexity requirement for the TRM, there could be other efficient algorithms that demand less. For example, simulations in [44] seem to suggest \( O(n \log n) \) samples suffice to guarantee efficient recovery. The simulations run an alternating direction algorithm fed with problem-specific initializations, nevertheless.

C. Image Data Again

Our algorithmic framework has been derived based on the BG model on the coefficients. Real data may not admit sparse representations w.r.t. complete dictionaries, or even so, the coefficients may not obey the BG model. In this experiment, we explore how our algorithm performs in learning complete dictionaries for image patches, emulating our motivational experiment in the companion paper [3] (Section I.B). Thanks to research on image compression, we know patches of natural images tend to admit sparse representation, even w.r.t. simple orthogonal bases, such as Fourier basis or wavelets.

We take the three images that we used in the motivational experiment. For each image, we divide it into \( 8 \times 8 \) non-overlapping patches, vectorize the patches, and then stack the vectorized patches into a data matrix \( \mathbf{Y} \). \( \mathbf{Y} \) is preconditioned as

![Figure 3](image-url)  
Fig. 3. Phase transition for recovering a single sparse vector. Top: We fix \( p = 5n^2 \log n \) and vary the dimension \( n \) and sparsity level \( k \); Bottom: We fix the sparsity level as \( k = 0.2 \cdot n \) and vary the dimension \( n \) and number of samples \( p \). For each configuration, the experiment is independently repeated for five times. White indicates success, and black indicates failure.
Our analysis is based on exact trust-region subproblem solver and fixed step size. The convergence result for the practical version from [72], based on approximate solver and adaptive step size, is general, but pessimistic. It seems not difficult to adapt their analysis according to our objective geometry, and obtain a tight, practical convergence result.

Our motivating experiment on real images in introduction of our companion paper [3] remains mysterious. If we were to believe that real image data are “nice” and our objective there does not have spurious local minima either, it is surprising ADM would escape all other critical points – this is not predicted by classic or modern theories. One reasonable place to start is to look at how gradient descent algorithms with generic initializations can escape from ridable saddle points (at least with high probability). The recent work [74] has showed that randomly perturbing each iterate can help gradient algorithm to achieve this with high probability.

VI. PROOF OF CONVERGENCE FOR THE TRUST-REGION ALGORITHM

A. Proof of Lemma 5

Proof: Using the fact \( \tanh(\cdot) \) and \( 1 - \tanh^2(\cdot) \) are bounded by one in magnitude, by (II.2) and (II.3) we have

\[
\|\nabla f(q)\| \leq \frac{1}{p} \sum_{k=1}^{p} \| (x_0)_k \| \leq \sqrt{n} \| X_0 \|_\infty, \\
\|\nabla^2 f(q)\| \leq \frac{1}{p} \sum_{k=1}^{p} \frac{1}{\mu} \| (x_0)_k \|^2 \leq \frac{n}{\mu} \| X_0 \|_\infty^2, 
\]

Fig. 4. Results of learning complete dictionaries from image patches, using the algorithmic pipeline in Section III. Top: Images we used for the experiment. These are the three images we used in our motivational experiment (Section I.B) in the companion paper [3]. The way we formed the data matrix \( Y \) is exactly the same as in that experiment. Middle: The 64 dictionary elements we learned. Bottom: Let \( A \) be the final dictionary matrix at convergence. This row shows the value \( \| A^{-1} Y \|_1 \) across one hundred independent runs. The values are almost the same, with a relative difference less than \( 10^{-3} \).
for any $q \in \mathbb{S}^{n-1}$. Moreover,

$$\sup_{q,q' \in \mathbb{S}^{n-1}, q \neq q'} \frac{\|\nabla f(q) - \nabla f(q')\|}{\|q - q'\|} \leq \frac{1}{p} \sum_{k=1}^{p} \|\langle x(0) \rangle\| \cdot \sup_{q,q' \in \mathbb{S}^{n-1}, q \neq q'} \frac{\|\tanh \left( \frac{q \cdot x(0) \mu}{\mu} \right) - \tanh \left( \frac{q' \cdot x(0) \mu}{\mu} \right)\|}{\|q - q'\|},$$

where at the last line we have used the fact the mapping $q \mapsto q^\delta \langle x(0) \rangle / \mu$ is $\|\langle x(0) \rangle\| / \mu$ Lipschitz, and $x \mapsto \tanh(x)$ is 1-Lipschitz, and the composition rule of Lipschitz functions (i.e., [3, Lemma V.5]). Similar argument yields the final bound.

**B. Proof of Lemma 6**

**Proof:** Suppose we can establish

$$|f(\exp_q(\delta)) - \tilde{f}(q, \delta)| \leq \frac{1}{6} \|\delta\|^3.$$

Applying this twice we obtain

$$f(\exp_q(\delta)) \leq \tilde{f}(q, \delta) + \frac{1}{6} \|\delta\|^3 \leq \tilde{f}(q, \delta) + \frac{1}{6} \|\delta\|^3 \leq f(\exp_q(\delta)) + \frac{1}{3} \|\delta\|^3 \leq f(q) - s + \frac{1}{3} \|\delta\|^3,$$

as claimed. Next we establish the first result. Let $\delta_0 = \frac{\delta}{\|\delta\|}$, and $t = \|\delta\|$. Consider the composite function

$$\zeta(t) = f(\exp_q(t\delta_0)) = f(q \cos(t) + \delta_0 \sin(t)),$$

and also

$$\dot{\zeta}(t) = \langle \nabla f(q \cos(t) + \delta_0 \sin(t)), -q \sin(t) + \delta_0 \cos(t) \rangle$$

$$\ddot{\zeta}(t) = \langle \nabla^2 f(q \cos(t) + \delta_0 \sin(t))(\delta \sin(t) + \delta_0 \cos(t)), -q \sin(t) + \delta_0 \cos(t) \rangle$$

$$+ \langle \nabla f(q \cos(t) + \delta_0 \sin(t)), -q \cos(t) - \delta_0 \sin(t) \rangle.$$

In particular, this gives that

$$\zeta(0) = f(q)$$

$$\dot{\zeta}(0) = \langle \delta_0, \nabla f(q) \rangle$$

$$\ddot{\zeta}(0) = \delta_0 \left( \nabla^2 f(q)(\delta \sin(t) + \delta_0 \cos(t)), -q \sin(t) + \delta_0 \cos(t) \right).$$

We next develop a bound on $|\ddot{\zeta}(t) - \ddot{\zeta}(0)|$. Using the triangle inequality, we can casually bound this difference as

$$|\ddot{\zeta}(t) - \ddot{\zeta}(0)| \leq |\nabla^2 f(q \cos(t) + \delta_0 \sin(t))(-q \sin(t) + \delta_0 \cos(t)), -q \sin(t) + \delta_0 \cos(t) - \delta_0 \nabla^2 f(q)\delta_0|$$

$$+ |\nabla f(q \cos(t) + \delta_0 \sin(t)), -q \cos(t) - \delta_0 \sin(t)|$$

$$+ |\nabla f(q), q|$$

$$\leq \|\nabla^2 f(q \cos(t) + \delta_0 \sin(t)) - \nabla^2 f(q)\|_{\infty}$$

$$\cdot \|\delta_0 \| + |\nabla f(q \cos(t) + \delta_0 \sin(t)), -q \sin(t) + \delta_0 \cos(t) - \delta_0 \sin(t)|$$

$$+ |\nabla f(q \cos(t) + \delta_0 \sin(t)), q|$$

$$+ |\nabla f(q), q|$$

$$\leq L_{V_2} \|q \cos(t) + \delta_0 \sin(t) - q\|$$

$$+ M_{V_2} \|\delta_0 \sin(t) - \delta_0 \cos(t)\|$$

$$+ M_{V_2} \|\delta_0 \cos(t) - \delta_0 \sin(t)\|$$

$$+ M_V \|q \cos(t) + \delta_0 \sin(t) + q\|$$

$$+ L_{V} \|q \cos(t) + \delta_0 \sin(t) - q\|$$

$$= (L_{V_2} + 2M_{V_2} + M_V + L_V) \|q\| \sqrt{1 - \cos(t)^2 + \sin(t)^2}$$

$$= \frac{\eta f^2}{\sqrt{2}} \leq \frac{\eta f^2}{\sqrt{2}},$$

where in the final line we have used the fact $1 - \cos x = 2 \sin^2(x/2)$ and that $\sin x \leq x$ for $x \in [0, 1]$, and $M_V$, $M_{V_2}$, $L_V$, and $L_{V_2}$ are the quantities defined in Lemma 5. By the integral form of Taylor’s theorem in Lemma 28 and the result above, we have

$$f(\exp_q(\delta)) \leq f(q) + \frac{1}{6} \|\delta\|^3.$$
there exists a tangent vector \( \delta \) such that
\[
\| \exp_q(\delta) - q \| 
\leq 2 - 2 \cos \| \delta \| = 4 \sin^2 \| \delta \| \leq \frac{9 \Delta^2}{2^8},
\]
which means that \( \sin(\| \delta \|) \leq 3 \Delta / (2\pi) \). Since \( \sin x \geq \frac{3}{2} x \) over \( x \in [0, \pi/6] \), it implies that \( \| \delta \| \leq \Delta \). Since \( g(w) = f(q(w)) \), by summarizing all the results, we conclude that there exists a \( \delta \) with \( \| \delta \| \leq \Delta \), such that
\[
f(\exp_q(\delta)) \leq f(q) - \min \left\{ \frac{\beta_g^2}{2L_g}, \frac{3\beta_g \Delta}{4\pi \sqrt{n}} \right\},
\]
as claimed.

D. Proof of Lemma 8

Proof: Let \( \sigma = \text{sign}(w^* \nabla g(w)) \). For any \( t \in \left[ 0, \frac{\Delta}{2\pi \sqrt{n}} \right] \), by integral form of Taylor’s theorem in Lemma 28, we have
\[
g\left( w - t \sigma \frac{w}{\|w\|} \right) = g(w) - t \sigma \frac{w^* \nabla g(w)}{\|w\|} + t^2 \int_0^1 \left[ \frac{w^* \nabla^2 g(w - t \sigma \frac{w}{\|w\|}) w}{\|w\|^2} \right] ds
\]
\[
\leq g(w) + \frac{t^2 \|w^* \nabla^2 g(w) w\|}{\|w\|^2}
\]
\[
+ t^2 \int_0^1 \left[ \frac{w^* \nabla^2 g(w - t \sigma \frac{w}{\|w\|}) w}{\|w\|^2} \right] ds
\]
Minimizing this function over \( t \in \left[ 0, \frac{\Delta}{2\pi \sqrt{n}} \right] \), we obtain that there exists a \( w' \in B\left( w, \frac{3\Delta}{2\pi \sqrt{n}} \right) \) such that
\[
g\left( w' \right) \leq g(w) - \min \left\{ \frac{\beta_g^2}{2L_g}, \frac{3\beta_g \Delta}{4\pi \sqrt{n}} \right\},
\]
Given such a \( w' \in B\left( w, \frac{3\Delta}{2\pi \sqrt{n}} \right) \), there must exist some \( \delta \in T_q S^{n-1} \) such that \( q(w') = \exp_q(\delta) \). It remains to show that \( \| \delta \| \leq \Delta \). It is easy to verify that \( \| q(w') - q(w) \| \leq 2/\sqrt{n} \| w' - w \| \leq 3 \Delta / \sqrt{n} \). Hence,
\[
\| \exp_q(\delta) - q \|^2 = \| q(1 - \cos \| \delta \|) + \frac{\delta}{\| \delta \|} \sin \| \delta \| \|^2
\]
\[
= 2 - 2 \cos \| \delta \| = 4 \sin^2 \| \delta \| \leq \frac{9 \Delta^2}{2^8},
\]
which means that \( \sin(\| \delta \| / 2) \leq 3 \Delta / (2\pi) \). Since \( \sin x \geq \frac{3}{2} x \) over \( x \in [0, \pi/6] \), it implies that \( \| \delta \| \leq \Delta \). Since \( g(w) = f(q(w)) \), by summarizing all the results, we conclude that there exists a \( \delta \) with \( \| \delta \| \leq \Delta \), such that
\[
f(\exp_q(\delta)) \leq f(q) - \min \left\{ \frac{\beta_g^2}{2L_g}, \frac{3\beta_g \Delta}{4\pi \sqrt{n}} \right\},
\]
as claimed.

E. Proof of Lemma 9

Proof: For any \( t \in \left[ 0, \frac{\Delta}{\| \text{grad} f(q^{(r)}) \|} \right] \), it holds that \( \| t \text{ grad } f(q^{(r)}) \| \leq \Delta \), and the quadratic approximation
\[
\tilde{f}(q^{(r)} - t \text{ grad } f(q^{(r)}))
\]
\[
\leq f(q^{(r)}) - t \| \text{grad } f(q^{(r)}) \|^2 + \frac{M_H t}{2} \| \text{grad } f(q^{(r)}) \|^2
\]
\[
= f(q^{(r)}) - t \left( 1 - \frac{M_H t}{2} \right) \| \text{grad } f(q^{(r)}) \|^2.
\]
Taking \( t_0 = \min \left\{ \frac{\Delta}{\| \text{grad } f(q^{(r)}) \|}, \frac{1}{M_H} \right\} \), we obtain
\[
\tilde{f}(q^{(r)} - t_0 \text{ grad } f(q^{(r)})) - f(q^{(r)})
\]
\[
\leq -\frac{1}{2} \min \left\{ \frac{\Delta}{\| \text{grad } f(q^{(r)}) \|}, \frac{1}{M_H} \right\} \| \text{grad } f(q^{(r)}) \|^2.
\]
Now let \( U \) be an arbitrary orthonormal basis for \( T_{q^{(r)}} S^{n-1} \). Since the norm constraint is active, by the optimality condition in (II.7), we have
\[
\Delta \leq \left\| \left[ U^* \text{ Hess } f(q^{(r)}) U \right]^{-1} U^* \text{ grad } f(q^{(r)}) \right\|
\]
\[
\leq \left\| U^* \text{ Hess } f(q^{(r)}) U \right\|^{-1} \left\| U^* \text{ grad } f(q^{(r)}) \right\|
\]
\[
\leq \frac{\| \text{grad } f(q^{(r)}) \|}{M_H}.
\]
which means that $\|\text{grad} \ f(q(r))\| \geq m_H \Delta$. Substituting this into (VI.1), we obtain

$$\hat{f}(q(r), -t_0 \ \text{grad} \ f(q(r)))$$

$$\leq f(q(r)) - \frac{1}{2} \min \left\{ m_H \Delta^2, \frac{m_H^2 \Delta^2}{2M_H} \right\}$$

$$\leq f(q(r)) - \frac{m_H \Delta^2}{2M_H} + \frac{1}{6} \eta_f \Delta^3.$$ 

By the key comparison result established in proof of Lemma 6, we have

$$f\left( \exp_{q(r)} \left( -t_0 \ \text{grad} \ f(q(r)) \right) \right)$$

$$\leq \hat{f}(q(r), -t_0 \ \text{grad} \ f(q(r))) + \frac{1}{6} \eta_f \Delta^3$$

$$\leq f(q(r)) - \frac{m_H \Delta^2}{2M_H} + \frac{1}{6} \eta_f \Delta^3.$$ 

This completes the proof.

F. Proof of Lemma 10

It takes certain delicate work to prove Lemma 10. Basically to use discretization argument, the degree of continuity of the Hessian is needed. The tricky part is that for continuity, we need to compare the Hessian operators at different points, while these Hessian operators are only defined on the respective tangent planes. This is the place where parallel translation comes into play. The next two lemmas compute spectral bounds for the forward and inverse parallel translation operators.

Lemma 19: For $\tau \in [0, 1]$ and $\|\delta\| \leq 1/2$, we have

$$\left\| P_{\gamma}^{\tau - 0} - I \right\| \leq \frac{5}{4} \tau \|\delta\|,$$  

$$(VI.2)$$

$$\left\| P_{\gamma}^{0 - \tau} - I \right\| \leq \frac{3}{2} \tau \|\delta\|.$$  

$$(VI.3)$$

Proof: By (II.8), we have

$$\left\| P_{\gamma}^{\tau - 0} - I \right\| = \left\| \left( \cos(\tau \|\delta\|) - 1 \right) \frac{\delta^*}{\|\delta\|^2} - \sin(\tau \|\delta\|) \frac{q \delta^*}{\|\delta\|} \right\|$$

$$\leq 1 - \cos(\tau \|\delta\|) + \sin(\tau \|\delta\|)$$

$$\leq 2 \sin^2 \left( \frac{\tau \|\delta\|}{2} \right) + \sin(\tau \|\delta\|)$$

$$\leq \frac{1}{4} \tau \|\delta\| + \tau \|\delta\| \leq \frac{5}{4} \tau \|\delta\|,$$ 

where we have used the fact $\sin(t) \leq t$ and $1 - \cos x = 2 \sin^2 (x/2)$. Moreover, $P_{\gamma}^{0 - \tau}$ is in the form of $(I + uv^*)^{-1}$ for some vectors $u$ and $v$. By the Sherman-Morrison matrix inverse formula, i.e., $(I + uv^*)^{-1} = I - uv^*/(1 + v^*u)$ (justified as $\left\| \cos(\tau \|\delta\|) - 1 \right\| \frac{\delta^*}{\|\delta\|^2} - \sin(\tau \|\delta\|) \frac{q \delta^*}{\|\delta\|}$ $\leq 5 \tau$

completing the proof.

The next lemma establishes the “local-Lipschitz” property of the Riemannian Hessian.

Lemma 20: Let $\gamma(t) = \exp_{q_0}(t \delta)$ denote a geodesic curve on $S^{n-1}$. Whenever $\|\delta\| \leq 1/2$ and $\tau \in [0, 1]$, we have

$$\left\| P_{\gamma}^{0 - \tau} \text{Hess} f(\gamma(\tau)) P_{\gamma}^{\tau - 0} - \text{Hess} f(q) \right\| \leq L_H \tau \|\delta\|,$$  

$$(VI.4)$$

where $L_H = 5n^{3/2} \|X_0\|_\infty/(2\mu^2) + \frac{n}{\pi} \|X_0\|_\infty + 9\sqrt{n} \|X_0\|_\infty$. 

Proof: First of all, by (II.6) and using the fact that the operator norm of a projection operator is unitary bounded, we have

$$\|\text{Hess} f(\gamma(t)) - \text{Hess} f(q)\|$$

$$\leq \left\| P_{T_{\gamma}(\delta)}^{S^{n-1}} \left[ \nabla^2 f(\gamma(t)) - \nabla^2 f(q) \right] - \left( \langle \nabla f(\gamma(t)), \gamma(t) \rangle - \langle \nabla f(q), q \rangle \right) I \right\|_{P_{T_{\gamma}(\delta)}^{S^{n-1}} - 1}$$

$$+ \left\| P_{T_{\gamma}(\delta)}^{S^{n-1}} \left( \nabla^2 f(\gamma(t)) - \nabla^2 f(q) \right) P_{T_{\gamma}(\delta)}^{S^{n-1}} - P_{T_{\gamma}(\delta)}^{S^{n-1}} \left( \nabla^2 f(q) - \nabla^2 f(q) \right) P_{T_{\gamma}(\delta)}^{S^{n-1}} \right\|_{P_{T_{\gamma}(\delta)}^{S^{n-1}} - 1}$$

$$\leq \left\| \nabla^2 f(\gamma(t)) - \nabla^2 f(q) \right\| + \left\| \nabla f(\gamma(t)) - \nabla f(q), \gamma(t) - q \right\| + \left\| \nabla f(q), \gamma(t) - q \right\|$$

$$+ \left\| P_{T_{\gamma}(\delta)}^{S^{n-1}} - P_{T_{\gamma}(\delta)}^{S^{n-1}} \right\|_{P_{T_{\gamma}(\delta)}^{S^{n-1}} - 1} + \left\| P_{T_{\gamma}(\delta)}^{S^{n-1}} - P_{T_{\gamma}(\delta)}^{S^{n-1}} \right\|_{P_{T_{\gamma}(\delta)}^{S^{n-1}} - 1}$$

$$\cdot \left\| \nabla^2 f(q) - \nabla f(q), q \right\|.$$ 

By the estimates in Lemma 5, we obtain

$$\|\text{Hess} f(\gamma(t)) - \text{Hess} f(q)\|$$

$$\leq \frac{2}{\mu^2} n^{3/2} \|X_0\|_\infty \|\gamma(t) - q\| + \frac{n}{\mu} \|X_0\|_\infty \|\gamma(t) - q\|$$

$$+ \sqrt{n} \|X_0\|_\infty \|\gamma(t) - q\| + 2 \|\gamma(t) - q\|^* \left( \frac{n}{\mu} \|X_0\|_\infty + \sqrt{n} \|X_0\|_\infty \right)$$

$$\leq \left( \frac{5}{2\mu^2} n^{3/2} \|X_0\|_\infty^3 + \frac{25n}{4\mu} \|X_0\|_\infty^2 + \frac{5}{4} \sqrt{n} \|X_0\|_\infty \right) \cdot \tau \|\delta\|,$$  

$$(VI.5)$$
where at the last line we have used the following estimates:
\[
\|\gamma (r) - q\| = \|q (\cos (r \|\delta\|)) + \delta (r \|\delta\|)\| \\
\leq \frac{5}{4} r \|\delta\|, (Proof of Lemma 19)
\]
\[
\|\gamma (r) \gamma^* (r) - q q^*\| \leq \left( \left( \frac{\delta^*}{\|\delta\|} - qq^* \right) \right) \sin^2 (r \|\delta\|) + 2 \sin (r \|\delta\|) \cos (r \|\delta\|) \leq \sin^2 (r \|\delta\|) + (2r \|\delta\|) \leq \frac{5}{2} r \|\delta\|.
\]

Therefore, by Lemma 19, we obtain
\[
\frac{\partial^0_r \text{Hess} f(\gamma (r)) \partial^r_{\gamma} \gamma - \text{Hess} f(q)}{\|\text{Hess} f(\gamma (r))\|} \leq \left( \frac{11}{4} \right) \frac{\|\delta\|}{\|q\|} \frac{\|\nabla^2 f (\gamma (r)) - \nabla^2 f (\gamma (r)) \nabla f (\gamma (r))\|}{\|\nabla^2 f (\gamma (r))\|} \frac{\|\gamma (r)\|}{\|\delta\|} \frac{\|\nabla f (\gamma (r))\|}{\|\delta\|} \frac{\|\nabla f (\gamma (r))\|}{\|\delta\|}.
\]

By Lemma 5 and substituting the estimate in (VI.5), we obtain the claimed result.

**Proof:** (of Lemma 10) For any given \(q\) with \(\|w(q)\| \leq \mu/(4\sqrt{2})\), assume \(U\) is an orthonormal basis for its tangent space \(T_q S^{n-1}\). Again we first work with the “canonical” section in the vicinity of \(e_n\) with the “canonical” reparameterization \(q(w) = [w/\sqrt{1 - \|w\|^2}]\).

1) **Expectation of the operator.** By definition of the Riemannian Hessian in (II.6), expressions of \(\nabla^2 f\) and \(\nabla f\) in (II.2) and (II.3), and exchange of differential and expectation operators, we obtain
\[
U^* \text{Hess} E \left[ f(q) \right] U = E \left[ f(q) \right] U = E \left[ 1 - \tanh^2 \left( \frac{q x}{\mu} \right) \right] xx^* U
\]

Let \(\bar{x} \in \mathbb{R}^{n-1}\) be the first \(n-1\) coordinates of \(x\) and \(w \in \mathbb{R}^{n-1}\) the similar subvector of \(q\) (as used in [3]).

We have
\[
U^* E \left[ 1 - \tanh^2 \left( \frac{q x}{\mu} \right) \right] xx^* U \geq 1 - \theta \mu \frac{1}{4} \frac{\|w - w\|^2}{\mu^2} \left[ \frac{1}{\sqrt{2\pi}} \frac{\|w - w\|^2}{\mu^2} + q_n^2 \right] \PhiCTION^* \left[ \frac{1}{\sqrt{2\pi}} \frac{\|w - w\|^2}{\mu^2} + q_n^2 \right] U.
\]

Now consider any vector \(z \in T_q S^{n-1}\) such that \(z = U v\) for some \(v \in \mathbb{R}^{n-1}\) and \(|z| = 1\). Then
\[
E \left[ \left. \tanh \left( \frac{q x}{\mu} \right) \right| q_n \right] = \frac{1}{\mu} \left( 1 - \tanh^2 \left( \frac{w^* x}{\mu} \right) \right) \left[ \frac{1}{\sqrt{2\pi}} \frac{\|w - w\|^2}{\mu^2} + q_n^2 \right] \PhiCTION^* \left[ \frac{1}{\sqrt{2\pi}} \frac{\|w - w\|^2}{\mu^2} + q_n^2 \right] U
\]

by proof of [3, Proposition II.7], where \(v \in \mathbb{R}^{n-1}\) as above is the first \(n-1\) coordinates of \(z\). Now we know that \(\langle q, z \rangle = 0\), or
\[
w^* v + q_n z_n = 0 \quad \Rightarrow \quad \frac{\|v\|}{\|z\|} = \frac{q_n}{\|w\|} = \frac{1}{\sqrt{1 - \|w\|^2}} \geq 50,
\]
where we have used \(\|w\| \leq \mu/(4\sqrt{2})\) and \(\mu \leq 1/10\) to obtain the last lower bound. Combining the above with the fact that \(\|z\| = 1\), we obtain
\[
\frac{E \left[ \left. \tanh \left( \frac{q x}{\mu} \right) \right| q_n \right]}{\|z\|} = \frac{1}{\mu} \left( 1 - \tanh^2 \left( \frac{w^* x}{\mu} \right) \right) xx^* U
\]

where we have simplified the expression using \(\theta \leq 1/2\). To bound the second term,
\[
E \left[ \tanh \left( \frac{q x}{\mu} \right) \right] q_n \|q_n\|^2
\]

Now we have the following estimate:
\[
E \left[ \tanh \left( \frac{q x}{\mu} \right) \right] \left( 1 - \tanh^2 \left( \frac{w^* x}{\mu} \right) \right) \PhiCTION^* \left[ \frac{1}{\sqrt{2\pi}} \frac{\|w - w\|^2}{\mu^2} + q_n^2 \right] U
\]

where at the last inequality we have applied Gaussian tail upper bound of Type II in Lemma 23. Since \(\|w - w\|^2 + q_n^2 \geq 1 - \|w\|^2 \geq 1 - \mu^2/2 \geq 31/32\) for
where at the first inequality we used the fact \(|\tanh(\cdot)| \leq 1\), at the second we invoked Lemma 33, and at the third we invoked Lemma 24. Taking \(R_Z = \sigma_Z^2 = 1\), by Lemma 26, we obtain
\[
P \left[ \left| \frac{1}{p} \sum_{k=1}^p Z_k - \frac{1}{p} \sum_{k=1}^p E[Z_k] \right| > t/2 \right] \leq 2 \exp \left( -\frac{pt^2}{16} \right) \]  for any \(t > 0\). Gathering (VI.9) and (VI.10), we obtain that for any \(t > 0\),
\[
P \left[ \left\| U^* \text{Hess} E \left[ f(q) \right] U - U^* \text{Hess} f(q) U \right\| > t \right] \leq P \left[ \left\| U^* \text{Hess} \left[ f(q) \right] - \text{Hess} f(q) \right\| > t/2 \right] + P \left[ \left\| \nabla f(q), q \right\| > \left\| \nabla f(q), q \right\| > t/2 \right] \leq 2n \exp \left( -\frac{pt^2}{32n^2 + 8nt} \right) + 2 \exp \left( -\frac{pt^2}{16} \right) \leq 4n \exp \left( -\frac{pt^2}{32n^2 + 8nt} \right). \]  (VI.11)

3) **Uniformizing the bound.** Now we are ready to pull above results together for a discretization argument. For any \(\varepsilon \in (0, \mu/(4\sqrt{2}))\), there is an \(\varepsilon\)-net \(N_\varepsilon\) of size at most \((3\mu/(4\sqrt{2}))^n\) that covers the region \(\{q : \|w(q)\| \leq \mu/(4\sqrt{2})\}\). By Lemma 20, the function \(\text{Hess} f(q)\) is locally Lipschitz within each normal ball of radius \(\|q - \exp_q(1/2)\| = \sqrt{2 - 2\cos(1/2)} \geq 1/\sqrt{2}\) with Lipschitz constant \(L_H\) (as defined in Lemma 20). Note that \(\varepsilon < \mu/(4\sqrt{2}) < 1/(4\sqrt{2}) < 1/\sqrt{5}\) for \(\mu < 1\), so any choice of \(\varepsilon \in (0, \mu/(4\sqrt{2}))\) makes the Lipschitz constant \(L_H\) valid within each \(\varepsilon\)-ball centered around one element of the \(\varepsilon\)-net. Let
\[
E_\infty \doteq \left\{ 1 \leq \|X_0\| \leq 4\sqrt{\log(np)} \right\}.
\]  From Lemma 31, \(P[E_\infty^c] \leq \theta (np)^{-7} + \exp(-0.3\theta np)\). By Lemma 20, with at least the same probability,
\[
L_H \leq C_2 n^{3/2} \log^{3/2}(np). \]

Set \(\varepsilon = \frac{\mu}{12\sqrt{2\pi} \mu L_H} < \mu/(4\sqrt{2})\), so
\[
#N_\varepsilon \leq \exp \left( n \log \frac{C_2 n^{3/2} \log^{3/2}(np)}{\theta} \right). \]

Let
\[
E_H \doteq \left\{ \max_{q \in N_\varepsilon} \left\| U^* \text{Hess} E \left[ f(q) \right] U - U^* \text{Hess} f(q) U \right\| \leq \frac{\theta}{12\sqrt{2\pi} \mu} \right\}. \]

On \(E_\infty \cap E_H\),
\[
\sup_{\|q\| \leq \|q^*\|} \left\| U^* \text{Hess} E \left[ f(q) \right] U - U^* \text{Hess} f(q) U \right\| \leq \frac{\theta}{6\sqrt{2\pi} \mu}. \]
So on $E_\infty \cap E_H$, we have

$$U^* \text{Hess } f(q) U \geq c_1 \frac{\theta}{\mu} \quad (\text{VI.12})$$

for any $c_1 \leq 1/(12\sqrt{2n})$. We take $c_1 = c$, for simplicity. Setting $\tau = \frac{1}{2n \pi} \sqrt{\frac{p}{2n}}$ in (VI.11), we obtain that for any fixed $q$ in this region,

$$\mathbb{P} \left[ \left\| U^* \text{Hess } E \left[ f(q) \right] U - U^* \text{Hess } f(q) U \right\| > t \right] \leq 4n \exp \left( -\frac{p \theta^2}{c_3 n^2 + c_4 n \theta / \mu} \right).$$

Taking a union bound, we obtain that

$$\mathbb{P} \left[ E^*_{\mathcal{H}} \right] \leq 4n \exp \left( -\frac{p \theta^2}{c_3 n^2 + c_4 n \theta / \mu} + C_3 n \log n + C_4 n \log \log p \right).$$

It is enough to make $p \geq C_5 n^3 \log(n/(\mu \theta) / (\mu \theta^2))$ to make the failure probability small, completing the proof. \hfill \square

**G. Proof of Lemma 11**

**Proof:** For a given $q$, consider the vector $r = q - e_n/q_n$. It is easy to verify that $(q, r) = 0$, and hence $r \in T_q S_n^{2n-1}$. Now, by (II.2) and (II.4), we have

\[
\begin{align*}
\langle \text{grad } f(q), r \rangle &= \left( (I - q q^* \right) \nabla f(q), q - e_n/q_n \rangle \\
&= \left( (I - q q^* \right) \nabla f(q), -e_n/q_n \rangle \\
&= \frac{1}{p} \sum_{k=1}^{p} \left( (I - q q^* \right) \tanh \left( \frac{q^*(x_0)_k}{\mu} \right) (x_0)_k, -e_n/q_n \right) \\
&= \frac{1}{p} \sum_{k=1}^{p} \tanh \left( \frac{q^*(x_0)_k}{\mu} \right) \left( - \frac{x_k(n)}{q_n} + q^*(x_0)_k \right) \\
&= \frac{1}{p} \sum_{k=1}^{p} \tanh \left( \frac{q^*(x_0)_k}{\mu} \right) \left( w^*(q) \nabla x_k - \frac{x_k(n)}{q_n} \| w(q) \|^2 \right) \\
&= w^*(q) \nabla g(w),
\end{align*}
\]

where an explicit expression for $g(w)$ can be found at the start of [3, Sec. IV]. Thus,

\[
\frac{\| w^* \nabla g(w) \|}{\| w \|} = \frac{\langle \text{grad } f(q), r \rangle}{\| w \|} \leq \| \text{grad } f(q) \| \frac{\| r \|}{\| w \|},
\]

where

\[
\frac{\| r \|^2}{\| w \|^2} = \frac{\| w \|^2 + \left( q_n - \frac{1}{q_n} \right)^2}{\| w \|^2} = \frac{1}{\| w \|^2} + \| w \|^2 / q_n^2 = \frac{1}{q_n^2} \leq \frac{1}{1 - 1/2000} = 2000 \quad \text{1999},
\]

where we have invoked our assumption that $\| w \| \leq \sqrt{\frac{1}{2000}}$. Therefore we obtain

\[
\| \text{grad } f(q) \| \geq \frac{\| w \| \| w^* \nabla g(w) \|}{\| r \|} \| w \| \geq \sqrt{\frac{1999}{2000} \| w^* \nabla g(w) \| \| w \|} \geq \sqrt{\frac{9}{10} \| w^* \nabla g(w) \| \| w \|},
\]

completing the proof. \hfill \square

**H. Proof of Lemma 12**

Proof of Lemma 12 combines the local Lipschitz property of $\text{Hess } f(q)$ in Lemma 20, and the Taylor’s theorem (manifold version, [68, Lemma 7.4.7]).

**Proof:** Let $\gamma (t)$ be the unique geodesic that satisfies $\gamma (0) = q^{(r)}$, $\gamma (1) = q^{(r+1)}$, and its directional derivative $\gamma (0) = \delta_\star$. Since the parallel translation defined by the Riemannian connection is an isometry, then

$$\| \text{grad } f(q^{(r+1)}) \| = \| p^{-1} \text{grad } f(q^{(r+1)}) \| .$$

Moreover, since $\| \delta_\star \| \leq \Delta$, the unconstrained optimality condition in (II.7) implies that $\| \text{grad } f(q^{(r)}) + \text{Hess } f(q^{(r)}) \delta_\star = 0 \|$. Thus, by using Taylor’s theorem in [68], we have

\[
\begin{align*}
\| \text{grad } f(q^{(r+1)}) \| &= \| p^{-1} \text{grad } f(q^{(r+1)}) - \| \text{grad } f(q^{(r)}) \| \\
&= \| \text{grad } f(q^{(r)}) - \text{grad } f(q^{(r)}) \| \\
&= \| \text{grad } f(q^{(r)}) \|.
\end{align*}
\]

From the Lipschitz bound in Lemma 20 and the optimality condition in (II.7), we obtain

$$\| \text{grad } f(q^{(r+1)}) \| \leq \frac{1}{2} \| \delta_\star \|^2 \| H = \frac{L_H}{2000} \| \text{grad } f(q^{(r)}) \|^2.$$

This completes the proof. \hfill \square

**I. Proof of Lemma 13**

**Proof:** By invoking Taylor’s theorem in [68], we have

$$p^{-1} \text{grad } f(\gamma(t)) = \int_0^t p^{-1} \text{Hess } f(\gamma(t)) [\gamma(t)] dt.$$

Hence, we have

\[
\begin{align*}
\frac{p^{-1} \text{grad } f(\gamma(t)) - \delta}{\| \delta \|} &= \int_0^t \frac{p^{-1} \text{Hess } f(\gamma(t)) [\gamma(t)] - \delta}{\| \delta \|} dt \\
&= \int_0^t \frac{p^{-1} \text{Hess } f(\gamma(t)) [\gamma(t)] - \delta}{\| \delta \|} dt + \int_0^t \text{Hess } f(\gamma(t)) [\gamma(t)] \| \gamma(t) \| dt \\
&\geq m_H \int_0^t \| \gamma(t) \|^2 dt \geq m_H t \| \delta \|^2,
\end{align*}
\]

where we have used the fact that the parallel transport $p^{-1}$ defined by the Riemannian connection is an isometry. On the other hand, we have

\[
\| p^{-1} \text{grad } f(\gamma(t)) - \delta \| \| \delta \| = \| m_H \text{grad } f(\gamma(t)) \| \| \delta \|,
\]
where again used the isometry property of the operator $D^{0_2}_g$. Combining the two bounds above, we obtain
\[
\|\text{grad } f (\gamma (r))\| \|\delta\| \geq m_H t \|\delta\|^2,
\]
which implies the claimed result.

VII. PROOFS OF TECHNICAL RESULTS FOR SECTION III

We need one technical lemma to prove Lemma 15 and the relevant lemma for complete dictionaries.

Lemma 21: For all integer $n_1 \in \mathbb{N}$, $\theta \in (0, 1/3)$, and $n_2 \in \mathbb{N}$ with $n_2 \geq Cn_1 \log (n_1/\theta)/\theta^2$, any random matrix $M \in \mathbb{R}^{n_1 \times n_2}$ i.i.d. $BG(\theta)$ obeys the following: for any fixed index set $\mathcal{I} \subset [n_2]$ with $|\mathcal{I}| \leq \frac{4}{5} n\theta n_2$, it holds that
\[
\|v^* M_{\mathcal{I}}\|_1 \leq \|v^* M_{\mathcal{I}}\|_1 \geq \frac{n_2}{6} \frac{\sqrt{2}}{\sqrt{\pi}} \|v\| \quad \text{for all } v \in \mathbb{R}^{n_1},
\]
with probability at least $1 - cp^{-6}$. Here $C, c$ are both positive constants.

Proof: By homogeneity, it is sufficient to consider all $v \in S^{n_1}$. For any $i \in [n_2]$, let $m_i \in \mathbb{R}^{n_1}$ be a column of $M$. For a fixed $v$ such that $\|v\| = 1$, we have
\[
T(v) = \|v^* M_{\mathcal{I}}\|_1 - \|v^* M_{\mathcal{I}}\|_1
\]
\[
= \sum_{i \in \mathcal{I}} \|v^* m_i\| - \sum_{i \notin \mathcal{I}} \|v^* m_i\|,
\]
namely as a sum of independent random variables. Since $|\mathcal{I}| \leq 9n_2 \theta/8$, we have
\[
\mathbb{E}[T(v)] \geq \left( n_2 - \frac{9}{8} \theta n_2 - \frac{9}{8} \theta n_2 \right) \mathbb{E}[\|v^* m_i\|]
\]
\[
= \left( 1 - \frac{9}{4} \theta \right) n_2 \mathbb{E}[\|v^* m_i\|]
\]
\[
\geq \frac{1}{4} n_2 \mathbb{E}[\|v^* m_i\|],
\]
where the expectation $\mathbb{E}[\|v^* m_i\|]$ can be lower bounded as
\[
\mathbb{E}[\|v^* m_i\|]
\]
\[
= \sum_{k=0}^{n_i} \theta^k (1 - \theta)^{n_i-k} \sum_{\mathcal{J} \in \mathcal{J}(n_1)} \mathbb{E}_{g \sim \mathcal{N}(0, I)} \mathbb{E}_{\mathcal{J}} \left[ \|v^* g_{\mathcal{J}}\|\right],
\]
\[
= \sum_{k=0}^{n_i} \theta^k (1 - \theta)^{n_i-k} \sum_{\mathcal{J} \in \mathcal{J}(n_1)} \sqrt{\frac{2}{\pi}} \|v_{\mathcal{J}}\|
\]
\[
\geq \sqrt{\frac{2}{\pi}} \|\mathbb{E}_{\mathcal{J}} [v^* g_{\mathcal{J}}]\| = \sqrt{\frac{2}{\pi}} \theta.
\]
Moreover, by Lemma 33 and Lemma 24, for any $i \in [n_2]$ and any integer $m \geq 2$,
\[
\mathbb{E}[\|v^* m_i\|^m] \leq \mathbb{E}_{Z \sim \mathcal{N}(0, I)} [\|Z\|^m] \leq (m - 1)!! \leq \frac{m!}{2}.
\]
So invoking the moment-control Bernstein’s inequality in Lemma 26, we obtain
\[
\mathbb{P}\left[ T(v) \leq \frac{n_2}{4 \sqrt{\frac{2}{\pi}} \theta - t} \right]
\]
\[
\leq \mathbb{P}[T(v) < \mathbb{E}[T(v)] - t] \leq \exp \left( -\frac{r^2}{2n_2 + 2t} \right).\]
It is obvious that the feasible set of (VII.5) contains that of (VII.4). So if \( e_n \) is the unique optimal solution (UOS) of (VII.5), it is the UOS of (VII.4). Suppose \( \mathcal{I} = \text{supp}(x_n) \) and define an event \( \mathcal{E}_0 = \{ |I| \leq \frac{9}{4} \theta p \} \).

By Hoeffding’s inequality, we know that \( P(\mathcal{E}_0) \leq \exp \left( -\theta^2 p/2 \right) \). Now conditioned on \( \mathcal{E}_0 \) and consider a fixed support \( \mathcal{I} \). (VII.5) can be further relaxed as

\[
\min_q \| x_n \|_1 | q_n | + \frac{\theta}{\theta^*} \| q_n \| \geq \frac{p}{\theta} \| x_n \|_1 \geq 1.
\]  

The objective value of (VII.6) lower bounds that of (VII.5), and are equal when \( q = e_n \). So if \( q = e_n \) is UOS of (VII.6), it is UOS of (VII.4). By Lemma 21, we know that

\[
\| q_n \| \geq \frac{p}{\theta} \| x_n \|_1 \geq 1.
\]

holds w.h.p., when \( p \geq C_1(n-1) \log ((n-1)/\theta) / \theta^2 \). Let \( \zeta = \frac{C_1}{\theta} \), thus we can further lower bound the objective value in (VII.6) by

\[
\min_q \| x_n \|_1 | q_n | + \zeta \| q_n \|, \quad \text{s.t.} \quad q_n r_n + \| q_n r_n \| \geq 1.
\]

By similar arguments, if \( e_n \) is the UOS of (VII.7), it is also the UOS of (VII.4). For the optimal solution of (VII.7), notice that it is necessary to have \( \text{sign}(q_n) = \text{sign}(r_n) \) and \( q_n r_n + \| q_n r_n \| = 1 \). Therefore, the problem (VII.7) is equivalent to

\[
\min_{q_n} \| x_n \|_1 | q_n | + \frac{1}{\theta} | r_n | | q_n | \| q_n \|, \quad \text{s.t.} \quad \| q_n \| \leq \frac{1}{| r_n |}.
\]

Notice that the problem (VII.8) is a linear program in \( | q_n | \) with a compact feasible set, which indicates that the optimal solution only occurs at the boundary points \( | q_n | = 0 \) and \( | q_n | = 1/| r_n | \). Therefore, \( q = e_n \) is the UOS of (VII.8) if and only if

\[
\frac{1}{| r_n |} | x_n | < \frac{\zeta}{\| r_n \|}.
\]

Conditioned on \( \mathcal{E}_0 \), by using the Gaussian concentration bound, we have

\[
P \left( \| x_n \|_1 \geq \frac{9}{4} \sqrt{\frac{2}{\theta^2} p + t} \right)
\]

\[
P \left( \| x_n \|_1 \geq \mathbb{E} \left[ \| x_n \|_1 + t \right] \right) \leq \exp \left( -\frac{t^2}{2p} \right),
\]

which means that

\[
P \left( \| x_n \|_1 \geq \frac{5}{4} \sqrt{\frac{2}{\theta^2} p} \right) \leq \exp \left( -\frac{\theta^2 p}{64\pi} \right). \quad (VII.10)
\]

Therefore, by (VII.9) and (VII.10), for \( q = e_n \) to be the UOS of (VII.4) w.h.p., it is sufficient to have

\[
\frac{5}{4} \sqrt{\frac{2}{\theta^2} p} < \frac{\theta p}{64\pi}
\]

which implies by

\[
| r_n | > \frac{249}{250}.
\]

The failure probability can be estimated via a simple union bound. Since the above argument holds uniformly for any fixed support set \( \mathcal{I} \), we obtain the desired result.

When our dictionary \( A_0 \) is an arbitrary orthogonal matrix, it only rotates the row subspace of \( X_0 \). Thus, w.l.o.g., suppose the TRM algorithm returns a solution \( \tilde{q} \), to which \( A_0 q_n = a \) is the nearest “target” with \( q_n \) a signed basis vector. By a change of variable \( \tilde{q} = A_0 q \), the problem (VII.4) is of the form

\[
\min_{\tilde{q}} \| q^* X_0 \|_1, \quad \text{s.t.} \quad \langle A_0^* r, \tilde{q} \rangle = 1,
\]

obviously our target solution for \( \tilde{q} \) is again the standard basis \( q_n \). By a similar argument above, we only need \( \langle A_0^* r, q_n \rangle > 249/250 \) to exactly recover the target, which is equivalent to \( | r_n, q_n | > 249/250 \). This implies that our rounding (III.1) is invariant to change of basis, completing the proof.

\section*{B. Proof of Lemma 17}

\textbf{Proof:} Define \( \tilde{q} = (U V^* + \Xi)^* q \). By Lemma 32, and in particular (II.50), when

\[
p \geq \frac{C_1}{\epsilon^2} \max \left\{ \frac{n^4}{\mu^3} \frac{n^5}{\mu^2} \right\} K^8 \left( \frac{\kappa(A_0) n}{\mu \theta} \right),
\]

\( \| \Xi \| \leq 1/2 \) so that \( U V^* + \Xi \) is invertible. Then the LP rounding can be written as

\[
\min_{\tilde{q}} \| q^* X_0 \|_1, \quad \text{s.t.} \quad \langle (U V^* + \Xi)^{-1} r, \tilde{q} \rangle = 1.
\]

By Lemma 15, to obtain \( \tilde{q} = e_n \) from this LP, it is enough to have

\[
\langle (U V^* + \Xi)^{-1} r, e_n \rangle \geq 249/250,
\]

and \( p \geq C_2 n^2 \log(n/\theta) / \theta \) for some large enough \( C_2 \). This implies that to obtain \( q_n \) for the original LP, such that \( (U V^* + \Xi)^* q_n = e_n \), it is enough that

\[
\langle (U V^* + \Xi)^{-1} r, (U V^* + \Xi)^* q_n \rangle = \langle r, q_n \rangle \geq 249/250,
\]

completing the proof.

\section*{C. Proof of Lemma 18}

\textbf{Proof:} Note that \( [q_1^1, \ldots, q_1^\ell] = (Q^* + \Xi)^{-1} [e_1, \ldots, e_\ell] \), we have

\[
U^* (Q + \Xi) X_0 = U^* (Q^* + \Xi)^{-1} (Q + \Xi) (Q^* + \Xi) X_0 = U^* [q^1_1, \ldots, q^\ell_1] (I + \Delta_\ell) X_0,
\]

Authorized licensed use limited to: New York University. Downloaded on August 26,2020 at 06:16:20 UTC from IEEE Xplore. Restrictions apply.
where $\tilde{V} = (Q^* + \Xi^*)^{-1}(e_{t+1}, \ldots, e_n)$, and the matrix $\Delta_1 = Q^* \Xi + \Xi^* Q + \Xi^* \Xi$ so that $\|\Delta_1\| \leq 3 \|\Xi\|$. Since $U^*[q_1, \ldots, q_n | \tilde{V}] = [0 | U^* \tilde{V}]$, we have

$$
U^*(Q + \Xi)X_0 = [0 | U^* \tilde{V}]X_0 + [0 | U^* \tilde{V}] \Delta_1 X_0
= U^* \tilde{V} X_0^{n-\ell} + \Delta_2 X_0,
$$

(VII.12)

where $\Delta_2 = [0 | U^* \tilde{V}] \Delta_1$. Let $\delta = \|\Xi\|$, so that

$$
\|\Delta_2\| \leq \frac{\|\Delta_1\|}{\sigma_{\min}(Q + \Xi)} \leq \frac{3 \|\Xi\|}{\sigma_{\min}(Q + \Xi)} \leq \frac{3\delta}{1 - \delta}.
$$

(VII.13)

Since the matrix $\tilde{V}$ is near orthogonal, it can be decomposed as $\tilde{V} = V + \Delta_3$, where $V$ is orthogonal, and $\Delta_3$ is a small perturbation. Obviously, $V = UR$ for some orthogonal matrix $R$, so that $V$ spans the same subspace as that of $U$. Next, we control the spectral norm of $\Delta_3$:

$$
\|\Delta_3\| = \min_{R \in O_\ell} \|UR - \tilde{V}\|
\leq \min_{R \in O_\ell} \|UR - Q_{[n-\ell]}\| + \|Q_{[n-\ell]} - \tilde{V}\|,
$$

(VII.14)

where $Q_{[n-\ell]}$ collects the last $n - \ell$ columns of $Q$, i.e., $Q = [Q_{[\ell]}, Q_{[n-\ell]}]$.

To bound the second term on the right, we have

$$
\|Q_{[n-\ell]} - \tilde{V}\| \leq \|Q^{-1} - (Q + \Xi)^{-1}\|
\leq \|Q^{-1}\| \|Q^{-1} \Xi\| \leq \frac{\delta}{1 - \delta},
$$

where we have used perturbation bound for matrix inverse (see, e.g., [84, Th. 2.5, Ch. III]).

To bound the first term, from Lemma 30, it is enough to upper bound the largest principal angle $\theta_1$ between the subspaces span($q_1, \ldots, q_n$), and that spanned by $Q(e_1, \ldots, e_\ell)$. Write $I_{[\ell]} = [e_1, \ldots, e_\ell]$ for short, we bound $\sin \theta_1$ as

$$
\sin \theta_1 \leq \left\| I_{[\ell]}^*(Q + \Xi)^{-1} - (Q^* + \Xi^*)^{-1} I_{[\ell]} \right\|
\leq \left\| (I_{[\ell]}^*(Q + \Xi)^{-1} - (Q^* + \Xi^*)^{-1}) I_{[\ell]}^*(Q + \Xi)^{-1} \right\|
\leq \left\| Q I_{[\ell]} I_{[\ell]}^*(Q^* + \Xi^*)^{-1} I_{[\ell]} \right\|
+ \left\| (Q^* + \Xi^*)^{-1} I_{[\ell]} \left[I - (I_{[\ell]}^*(I + \Delta_1)^{-1} I_{[\ell]})^{-1}\right] I_{[\ell]}^*(Q + \Xi)^{-1} \right\|
\leq \left(1 + \frac{1}{\sigma_{\min}(Q + \Xi)}\right) \|Q - (Q + \Xi)^{-1}\|
+ \frac{1}{\sigma_{\min}(Q + \Xi)} \|I - (I_{[\ell]}^*(I + \Delta_1)^{-1} I_{[\ell]})^{-1}\|
\leq \left(1 + \frac{1}{\delta}\right) \frac{\delta}{1 - \delta}
+ \frac{1}{\delta} \frac{1}{1 - \delta} \frac{1}{(1 - \delta)^2} \frac{1}{2 - 2 \|\Delta_1\|},
$$

where to obtain the first line we used that for any full column rank matrix $M, M(M^* M)^{-1} M^*$ is the orthogonal projector onto its column span, and to obtain the fifth and sixth lines we invoked the matrix inverse perturbation bound again. Using $\delta < 1/20$ and $\|\Delta_1\| \leq 3\delta < 1/2$, we have

$$
\sin \theta_1 \leq (2 - \delta)\delta + \frac{3\delta}{(1 - \delta)^2(1 - 6\delta)}
\leq 5\delta - 13\delta^2 + 6\delta^3
\leq \frac{8\delta}{(1 - \delta)^2(1 - 6\delta)} \leq 8\delta.
$$

For $\delta < 1/20$, the upper bound is nontrivial. By Lemma 30,

$$
\min_{R \in O_\ell} \|UR - Q_{[n-\ell]}\| \leq 2\sqrt{2 - 2\cos \theta_1}
\leq 2\sqrt{2 - 2\sin^2 \theta_1}
\leq 8\sqrt{2}\delta.
$$

Put the estimates above, there exists an orthogonal matrix $R \in O_\ell$ such that $V = UR$ and $\tilde{V} = V + \Delta_3$ with

$$
\|\Delta_3\| \leq \delta/(1 - \delta) + 8\sqrt{2}\delta \leq 25/(2\delta).
$$

(VII.15)

Therefore, by (VII.12), we obtain

$$
U^*(Q + \Xi)X_0 = U^* VX_0^{n-\ell} + \Delta,
$$

(VII.16)

with $\Delta = U^* \Delta_3 X_0^{n-\ell} + \Delta_2 X_0$.

By using the results in (VII.13) and (VII.15), we get the desired result.

**APPENDIX A**

**TECHNICAL TOOLS AND BASIC FACTS USED IN PROOFS**

In this section, we summarize some basic calculations that are useful throughout, and also record major technical tools we use in proofs.

**Lemma 22 (Derivatives and Lipschitz Properties of $h_\mu(z)$):**

For the sparsity surrogate

$$
h_\mu(z) = \mu \log \cosh(z/\mu),
$$

the first two derivatives are

$$
\hat{h}_\mu(z) = \tanh(z/\mu), \quad \tilde{h}_\mu(z) = \left[1 - \tanh^2(z/\mu)\right]/\mu.
$$

Also, for any $z > 0$, we have

$$
(1 - \exp(-2z/\mu))/2 \leq \tanh(z/\mu) \leq 1 - \exp(-2z/\mu),
\quad \exp(-2z/\mu) \leq 1 - \tanh^2(z/\mu) \leq 4 \exp(-2z/\mu).
$$

Moreover, for any $z, z' \in \mathbb{R}$, we have

$$
|\hat{h}_\mu(z) - \hat{h}_\mu(z')| \leq |z - z'|/\mu,
\quad |\tilde{h}_\mu(z) - \tilde{h}_\mu(z')| \leq 2|z - z'|/\mu^2.
$$
Lemma 23 (Gaussian Tail Estimates): Let \( X \sim N(0, 1) \) and \( \Phi(x) \) be CDF of \( X \). For any \( x \geq 0 \), we have the following estimates for \( \Phi'(x) \approx 1 - \Phi(x) \):

\[
\left( \frac{1}{x} - \frac{1}{x^2} \right) \exp\left(-\frac{x^2}{2}\right) \frac{1}{\sqrt{2\pi}} \leq \Phi'(x) \leq \left( \frac{1}{x} - \frac{1}{x^2} + \frac{3}{x^3} \right) \exp\left(-\frac{x^2}{2}\right) \frac{1}{\sqrt{2\pi}}, \quad (\text{Type I})
\]

\[
\left( \frac{x}{x^2 + 1} \right) \exp\left(-\frac{x^2}{2}\right) \frac{1}{2\sqrt{2\pi}} \leq \Phi'(x) \leq \left( \frac{x}{x^2 + 1} \right) \exp\left(-\frac{x^2}{2}\right) \frac{1}{\sqrt{2\pi}}, \quad (\text{Type II})
\]

\[
\left( \sqrt{2 + x^2} - x \right) \exp\left(-\frac{x^2}{2}\right) \frac{1}{\sqrt{2\pi}} \leq \Phi'(x) \leq \left( \sqrt{2 + x^2} - x \right) \exp\left(-\frac{x^2}{2}\right) \frac{1}{\sqrt{2\pi}}, \quad (\text{Type III}).
\]

**Proof:** See proof of Lemma A.5 in the technical report [2].

Lemma 24 (Moments of the Gaussian RV): If \( X \sim N(0, \sigma^2) \), then it holds for all integer \( p \geq 1 \) that

\[
E[X^p] \leq \sigma^p (p-1)!.!
\]

Lemma 25 (Moments of the \( \chi^2 \) RV): If \( X \sim \chi^2(n) \), then it holds for all integer \( p \geq 1 \) that

\[
E[X^p] = 2^n \Gamma(\frac{p+n/2}{2}) \Gamma(\frac{n/2}{2}) = \sum_{k=1}^{\infty} ((n+2k-2) \leq p! (2n)^p / 2.
\]

Lemma 26 (Moment-Control Bernstein’s Inequality for Scalar RVs, [85, Th. 2.10]): Let \( X_1, \ldots, X_p \) be i.i.d. real-valued random variables. Suppose that there exist some positive numbers \( R \) and \( \sigma \) such that

\[
E[|X_k|^m] \leq m! n^2 R^{m-2} / 2, \quad \text{for all integers } m \geq 2.
\]

Let \( S \doteq \frac{1}{p} \sum_{k=1}^{p} X_k \), then for all \( t > 0 \), it holds that

\[
P[|S - E[S]| \geq t] \leq 2 \exp\left(-\frac{pt^2}{2\sigma^2 + 2Rt}\right).
\]

Lemma 27 (Moment-Control Bernstein’s Inequality for Matrix RVs, [86, Th. 6.2]): Let \( X_1, \ldots, X_p \in \mathbb{R}^{d \times d} \) be i.i.d. random, symmetric matrices. Suppose there exist some positive number \( R \) and \( \sigma \) such that

\[
E[X_k^p] \leq m^2 \sigma^2 R^{m-2} / 2 \cdot I \quad \text{and} \quad -E[X_k^p] \leq m^2 \sigma^2 R^{m-2} / 2 \cdot I, \quad \text{for all integers } m \geq 2.
\]

Let \( S \doteq \frac{1}{p} \sum_{k=1}^{p} X_k \), then for all \( t > 0 \), it holds that

\[
P[|S - E[S]| \geq t] \leq 2d \exp\left(-\frac{pt^2}{2\sigma^2 + 2 Rt}\right).
\]

**Proof:** See proof of Lemma A.10 in the technical report [2].

Lemma 28 (Integral Form of Taylor’s Theorem): Let \( f(x) : \mathbb{R}^n \mapsto \mathbb{R} \) be a twice continuously differentiable function, then for any direction \( y \in \mathbb{R}^n \), we have

\[
f(x + ty) = f(x) + t \int_0^1 \langle \nabla f(x + s t y), y \rangle \, ds, \quad f(x + ty) = f(x) + t \langle \nabla f(x), y \rangle + \frac{1}{2} \int_0^1 (1 - s) \langle \nabla^2 f(x + st y)y, y \rangle \, ds.
\]

**APPENDIX B**

**Auxiliary Results for Proofs**

Lemma 29: There exists a positive constant \( C \) such that for any \( \theta \in (0, 1/2) \) and \( n_2 > C n_1^2 \log n_1 \), the random matrix \( X \in \mathbb{R}^{n_1 \times n_2} \) with \( X \sim_{i.i.d.} \text{BG}(\theta) \) obeys

\[
\left\| \frac{1}{n_2^2} X X^* - I \right\| \leq 10 \sqrt{\frac{\theta n_1 \log n_2}{n_2}}.
\]

**Proof:** See proof of [3, Lemma B.3].

Lemma 30: Consider two linear subspaces \( U, V \) of dimension \( k \) in \( \mathbb{R}^n \) (\( k \in [n] \)) spanned by orthonormal bases \( U \) and \( V \), respectively. Suppose \( \pi / 2 \geq \theta_0 \geq \theta_2 \geq \cdots \geq \theta_k \geq 0 \) are the principal angles between \( U \) and \( V \). Then it holds that

i) min\( \theta_{k \in [k]} \|U - V Q\| \leq \sqrt{2 - 2 \cos \theta_1};
\]

ii) \( \sin \theta_1 = \|U^* - V^*\| \); 

iii) Let \( U^\perp \) and \( V^\perp \) be the orthogonal complement of \( U \) and \( V \), respectively. Then \( \theta_1(U, V) = \theta_1(U^\perp, V^\perp) \).

**Proof:** See proof of [2, Lemma B.4].

Below are restatements of several technical results from [3] that are important for proofs in this paper.

**Proposition 5 (Hessian Lipschitz):** Fix any \( r_\perp \in (0, 1) \). Over the set \( \Gamma \cap \{w : \|w\| \geq r_\perp\} \), \( w^* \nabla^2 g(w; X_0) w / ||w||^2 $$ is \( L_\perp \)-Lipschitz with

\[
L_\perp \leq \frac{16 n^3}{\mu^2} \|X_0\|^3 + \frac{8n^{5/2}}{\mu \rho_{\perp}^2} \|X_0\|^2 + \frac{48n^{5/2}}{\mu} \|X_0\|^2 + 96n^{5/2} \|X_0\| \|X_0\|.
\]

**Proposition 6 (Gradient Lipschitz):** Fix any \( r_\perp \in (0, 1) \). Over the set \( \Gamma \cap \{w : \|w\| \geq r_\perp\} \), \( w^* \nabla g(w; X_0) / ||w|| $$ is \( L_g \)-Lipschitz with

\[
L_g \leq \frac{2\sqrt{n} \|X_0\|}{r_\perp^2} + 8n^{3/2} \|X_0\| + 4n^2 \|X_0\|^2.
\]

**Proposition 7 (Lipschitz for Hessian Around Zero):** Fix any \( r_\perp \in (0, 1/2) \). Over the set \( \Gamma \cap \{w : \|w\| \leq r_\perp\} \), \( \nabla^2 g(w; X_0) $$ is \( L_\perp \)-Lipschitz with

\[
L_\perp \leq \frac{4n^2}{\mu^2} \|X_0\|^3 + \frac{4n}{\mu} \|X_0\| + \frac{8\sqrt{2} \sqrt{n}}{\mu} \|X_0\| + 8 \|X_0\| \|X_0\|.
\]

**Lemma 31:** For any \( \theta \in (0, 1) \), consider the random matrix \( X \in \mathbb{R}^{n_1 \times n_2} \) with \( X \sim_{i.i.d.} \text{BG}(\theta) \). Define the event \( \mathcal{E}_\infty \doteq \{1 \leq \|X\| \leq 4 \sqrt{\log(n_\rho, n_\mu)} \} \). It holds that

\[
P[\mathcal{E}_\infty] \leq \theta (n_\rho n_\mu)^{-7} + \exp(-0.3 \theta n_\rho).
\]
Lemma 32: For any \( \theta \in (0, 1/2) \), suppose \( A_0 \) is complete with condition number \( \kappa (A_0) \) and \( X_\infty \). BG (\( \theta \)). Provided \( p \geq C \kappa^4 (A_0) \theta^n \log (1 / \theta (A_0)) \), one can write \( \hat{F} \) as defined in (I.49) as
\[
\hat{F} = U V^* X_0 + Z X_0,
\]
for a certain \( Z \) obeying \( \| Z \| \leq 20 C \kappa^4 (A) \sqrt{\theta n \log p / p} \), with probability at least \( 1 - p^{-8} \). Here \( U \Sigma V^* = SVD (A_0) \), and \( C \) is a positive numerical constant.

For a certain \( \Xi \) obeying \( \| \Xi \| \leq 20 C \kappa^4 (A) \sqrt{\theta n \log p / p} \), with probability at least \( 1 - p^{-8} \). Here \( U \Sigma V^* = SVD (A_0) \), and \( C \) is a positive numerical constant.

Lemma 33: Suppose \( z, z' \in \mathbb{R}^n \) are independent and obey \( z \sim_{i.i.d.} \) BG \( (\theta) \) and \( z' \sim_{i.i.d.} N (0, 1) \). Then, for any fixed vector \( v \in \mathbb{R}^n \), it holds that
\[
E [v^* z^m] \leq \Bigg[ E [v^* z^m] \Bigg] = E_{z \sim N (0, 1)} [\| z \|^m],
\]
\[
E [\| z \|^m] \leq \Bigg[ E [\| z \|^m] \Bigg],
\]
for all integers \( m \geq 1 \).

Acknowledgment

The authors thank Dr. Boaz Barak for pointing out an inaccurate comment made on overcomplete dictionary learning using SOS. They thank Cun Mu and Henry Kuo of Columbia University for discussions related to this project. They also thank the anonymous reviewers for their careful reading of the paper, and for comments which have helped us to substantially improve the presentation. JS thanks the We Family Private Foundation for their generous support. An extended abstract of the current work has been published in [1]."
Qing Qu (’S) is a Ph.D student in Electrical Engineering, Columbia University, working with Prof. John Wright. He received a B.Eng. in Electronic Engineering from Tsinghua University, Beijing, in Jul. 2011, and a M.Sc. in Electrical Engineering from the Johns Hopkins University, Baltimore, in Dec. 2012. He interned at U.S. Army Research Laboratory and Microsoft Research from Jun. 2012 to Aug. 2013, and May 2016 to Aug 2016, respectively. His research interest lies at the intersection of machine learning, numerical optimization, signal/image processing, information theory and compressed sensing, where his current research focuses on developing practical algorithms and provable guarantees for many nonconvex problems with low intrinsic structure. He is the recipient of spotlight talk award at NY Machine Learning Symposium, Best Student Paper Award at SPARS15 (with Ju Sun, John Wright), and the recipient of 2016-18 Microsoft Research Fellowship in North America.

John Wright (’S’06–M’09) is an Associate Professor in the Electrical Engineering Department at Columbia University. He received his PhD in Electrical Engineering from the University of Illinois at Urbana-Champaign in 2009, and was with Microsoft Research from 2009-2011. His research is in the area of high-dimensional data analysis. In particular, his recent research has focused on developing algorithms for robustly recovering structured signal representations from incomplete and corrupted observations, and applying them to practical problems in imaging and vision. His work has received a number of awards and honors, including the Best Paper Award from the Conference on Learning Theory (COLT) in 2012 and the 2015 PAMI TC Young Researcher Award.